

# ProBench: Judging Multimodal Foundation Models on Open-ended Multi-domain Expert Tasks

Anonymous ACL submission

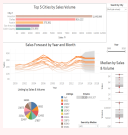
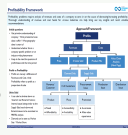
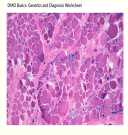
Coding; Screenshots and UI Elements	Knowledge; Document and Text-based Images	Science; Medical Images
<b>Query:</b> i want you to write a Rshiny code in rstudio to generate above visualization. Can you do that? 	<b>Query:</b> Explain this framework to me in detail and in chronological order. I am an aspiring consultant and I need to know this. Also give me potential issues and solutions that will come up through this. 	<b>Query:</b> The image above represents a H&E stain of a skeletal muscle biopsy from a young boy who came into the clinic reporting muscle weakness. You are his doctor. Does the boy have Duchenne muscular dystrophy? Explain. Your answer should include an analysis of the biopsy (you can use arrows to point to various features) and be sure to list all features of the muscle that indicate diseased or healthy conditions. 
<b>Task sub-field:</b> Code Generation <b>Image field:</b> Interactive Tools <b>Keywords:</b> Multiple complex visual elements; no domain knowledge.	<b>Task sub-field:</b> Human and Culture <b>Image sub-field:</b> Diagrams <b>Keywords:</b> Profitability framework; structured diagram; moderate reasoning.	<b>Task sub-field:</b> Life Science/Medical <b>Image sub-field:</b> Pathology Slides <b>Keywords:</b> Medical diagnosis; pathological analysis; fiber size variation; signs of necrosis and infiltration; specialized knowledge.

Figure 1: Examples of ProBench with varying lengths. We show the task and image fields in the header of each sample. Due to space limitations, more diverse and longer samples are provided in the supplementary material.

## Abstract

Solving expert-level multimodal tasks is a key milestone in general intelligence. As the capabilities of multimodal large language models (MLLMs) continue to evolve, evaluation of frontier multimodal intelligence becomes necessary yet challenging. In this work, we introduce ProBench, a benchmark of open-ended user queries encapsulating professional expertise and advanced reasoning. ProBench consists of 4,000 high-quality samples independently collected from professionals based on their productivity demands. It spans across 10 fields and 56 sub-fields, including science, arts, humanities, coding, mathematics, and creative writing. Experimentally, we evaluate and compare 24 latest models using MLLM-as-a-Judge. Our results reveal that although the best open-source models rival the proprietary ones, they all face significant challenges in visual perception, textual understanding, domain knowledge, and advanced reasoning. Our benchmark is publicly accessible at [TBC](#).

## 1 Introduction

Solving expert-level multimodal tasks with multimodal large language models (MLLMs) represents an important milestone toward achieving human-level general intelligence. However, these tasks require MLLMs to possess strong user query understanding, domain-specific knowledge, and advanced reasoning abilities. Ensuring their reliabil-

ity before deployment necessitates rigorous evaluation. To address it, we introduce ProBench, a challenging and automatic evaluation benchmark leveraging MLLM-as-a-Judge. ProBench consists of 4,000 queries from professional users, covering diverse productivity demands to assess MLLM capabilities in open-ended scenarios (Fig. 1).

One common benchmark to evaluating MLLM performance with expert knowledge is MMMU (Yue et al., 2024a). While effective for automatic evaluation using predefined answers, these closed-ended visual question answering benchmarks fail to capture MLLM capabilities in open-ended user interactions. Specifically, they do not adequately assess MLLM ability to follow user instructions or align with human preferences, both of which are fundamental for real-world applications (Lu et al., 2024; Luo et al., 2024; Chen et al., 2024b). Similar limitations apply to other benchmarks, such as MMMU-pro (Yue et al., 2024b), MMBench (Liu et al., 2025), and others (Lu et al., 2023; Masry et al., 2022; Singh et al., 2019; Wu et al., 2024).

Alternatively, MLLM-as-a-Judge is employed to automatically evaluate model performance in open-ended scenarios. However, existing benchmarks fail to rigorously assess MLLMs on expert-level professional tasks. Some (Chen et al., 2024b) are artificially constructed by a small group of experts, limiting their ability to reflect real-world user interactions. The remaining benchmarks (Luo et al.,

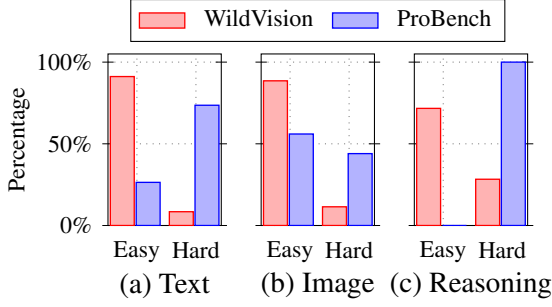


Figure 2: Comparison with WildVision (Lu et al., 2024) on challenge levels of (a) text, (b) image, and (c) reasoning for user instruction queries. To ensure a fair comparison, we follow WildVision by selecting the top 500 highest-quality queries from the single-round conversations.

2024; Lu et al., 2024), such as WildVision, are mostly set in general chat environments and require much less domain knowledge to solve.

To fill this gap, in this paper, we aim to design an *open-ended benchmark that requires expert-level knowledge* for multimodal tasks. Our ProBench is created from high-quality interactions within 100K real-world, professionally crowdsourced multimodal conversations for productivity scenarios. Specifically, samples are collected by encouraging users to ask questions related to their daily professional work, which usually require significant expert-level knowledge. This distinction sets our benchmark apart from prior works like WildVision (Lu et al., 2024) (Fig. 2). For a comprehensive evaluation, ProBench includes three tracks: single-round, multi-round, and multi-linguistic conversations. They respectively span 10 task fields and 56 sub-fields, support 17 languages, and support conversations with up to 13 conversation turns. An overview of ProBench is presented in Fig. 3.

Leveraging MLLM-as-a-Judge (e.g., gpt-4o), we assess 24 leading MLLMs on ProBench. Our evaluation reveals several key limitations in state-of-the-art MLLMs: i) current MLLMs struggle in visual perception, textual understanding, domain knowledge, and advanced reasoning, suffering from tasks like mathematics and planning; ii) multi-linguistic understanding and long-context reasoning during multi-round interaction remain challenging for most existing MLLMs. Our main contributions are summarized as follows:

- we introduce ProBench, an open-ended multimodal benchmark tailored for professional work scenarios requiring expert-level knowledge, featuring 4,000 samples across 10 task fields over 56 sub-fields. The benchmark also features multi-round conversations up to 13

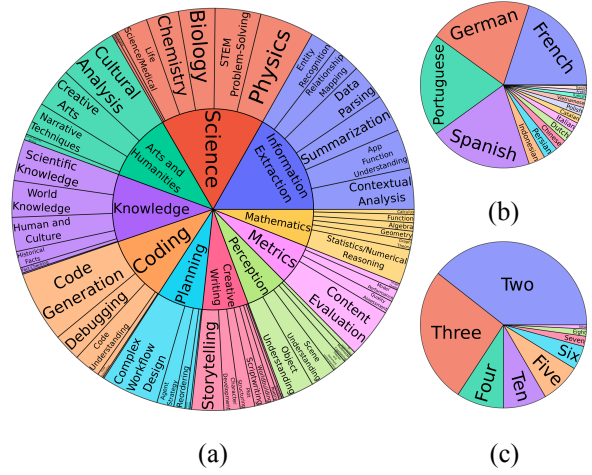


Figure 3: ProBench overview. Distributions of (a) task fields on the single-round track, (b) languages on the multi-linguistic track, and (c) conversation rounds on the multi-round tracks.

turns and multi-linguistic tracks in 17 languages;

- we design an automatic pairwise evaluation pipeline using MLLM-as-a-Judge, achieving 79.9% agreement with human experts. The evaluation is robust to different comparison baseline and judge model choices. We also provide a distilled version of Llama-vision to support cost-effective local evaluations;
- we conduct comprehensive evaluations using 24 leading MLLMs, showing that ProBench presents significant challenges for existing MLLMs, in visual perception, advanced reasoning, and domain knowledge. This signifies the need for more advanced multimodal models for high-value practical scenarios.

## 2 ProBench

**Preliminary.** The ProBench dynamically ranks MLLMs by employing the ELO rating system, implemented through statistical modeling based on direct pairwise model comparisons. In the following, we provide an overview. For further details, please refer to (Elo, 1966; Hunter, 2004). Given  $N$  MLLMs, an online ELO rating system compares model  $i$  with rating  $r_i$  and model  $j$  with rating  $r_j$  using the probability  $P(y_{i,j} = 1)$ . Here,  $y_{i,j}$  denotes the binary outcome, where  $y_{i,j} = 1$  indicates that model  $i$  wins, and  $y_{i,j} = 0$  indicates that model  $j$  wins. The probability is calculated by

$$P(y_{i,j} = 1) = \frac{1}{1 + 10^{(r_i - r_j)/\alpha}},$$

where  $\alpha$  is a hyperparameter that serves as a scaling factor, typically set to  $\alpha = 400$ . The ELO rating is

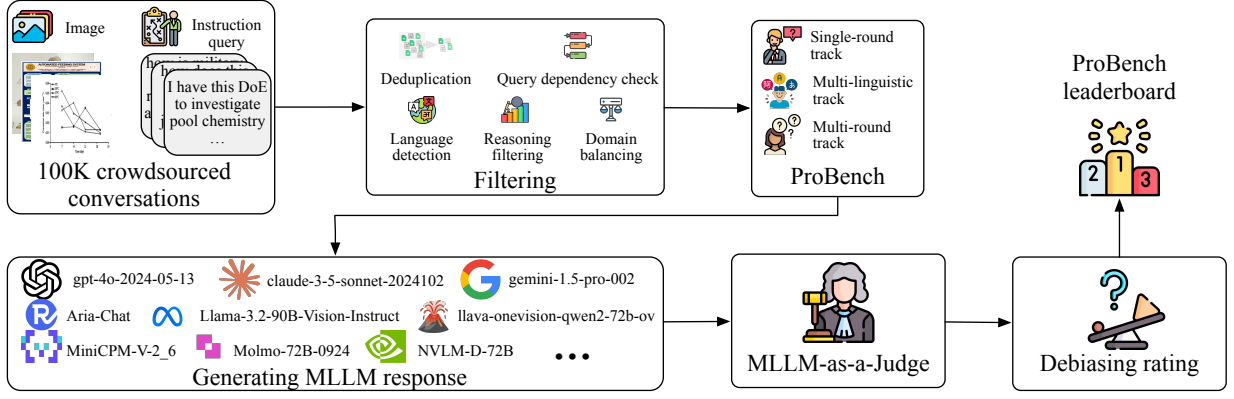


Figure 4: Framework of ProBench. Starting with 100K crowdsourced conversations, we identify high-quality user queries to curate single-round, multi-linguistic, and multi-round tracks. Using MLLM-as-a-Judge, we benchmark and rank 24 state-of-the-art MLLMs with ELO ratings. To ensure fairness, the ELO ratings are de-biased to remove confounder effects (*e.g.*, MLLM response formats), resulting in the final ProBench leaderboard. Icons in the figure are sourced from (Freepik et al., 2025).

dynamically updated after each model comparison. Taking model  $i$  as an example, the rating is updated according to the following rule:

$$r_i^{\text{upt}} = r_i + K \times (s_{i,j} - P(y_{i,j} = 1)) .$$

Similarly,  $K$  is a constant determining the magnitude of rating adjustments, commonly set to  $K = 32$ . The term  $s_{i,j}$  is a scalar representing the actual outcome: 0 for a loss, 0.5 for a tie, and 1 for a win. This updating rule encourages that a higher-rated model gains fewer points for a win, and loses more points for a defeat, while a lower-rated model experiences the opposite effect.

However, when using MLLM-as-a-Judge, the comparison results can be sensitive to model presentation order and confounded by response style variations (Li et al., 2024c). To address these challenges, the ProBench incorporates the Bradley-Terry model (Hunter, 2004) as an additional layer atop the ELO system. For  $N$  MLLMs and  $M$  pairwise comparisons, each round  $1 \leq m \leq M$  compares model  $i$  and model  $j$ . We have  $\mathbf{X}_m^{\text{win}} \in \mathbb{R}^N$  to indicate which model is presented first<sup>1</sup>, while  $\mathbf{X}_m^{\text{sty}} \in \mathbb{R}^S$  captures  $S$  stylistic differences between the outputs of models  $i$  and  $j$  (*e.g.*, word counts, and use of markdown). The Bradley-Terry model then refines the rating of model  $i$  as

$$r_i^{\text{ref}} = C + K \times \hat{\beta}_i ,$$

$$\hat{\beta}, \hat{\gamma} = \arg \min_{\beta, \gamma} \sum_{m,i,j} \ell_{\text{bce}}(\beta^\top \mathbf{X}_m^{\text{win}} + \gamma^\top \mathbf{X}_m^{\text{sty}}, s_{i,j}) ,$$

where  $\ell_{\text{bce}}(\cdot, \cdot)$  is the binary cross-entropy loss,  $C$  is a baseline rating constant,  $\beta \in \mathbb{R}^N$  and  $\gamma \in$

<sup>1</sup>This bias can be easily mitigated by evaluating twice while swapping the comparison order.

$\mathbb{R}^S$  are respectively known as the model strength and style coefficients, and  $\hat{\beta}_i$  is a scaler indicating strength of model  $i$ . This refinement known as style control in the literature (Li et al.) compensates for stylistic biases, ensuring a fair model performance evaluation.

**Overview.** Our paper aims to establish a comprehensive and challenging benchmark for evaluating MLLMs. The resulting ProBench is built on two primary components: i) curating high-quality conversations from a crowdsourced dataset, categorized into single-round, multi-linguistic, and multi-round tracks; ii) employing MLLM-as-a-Judge to compare and rank MLLMs. In total, 3000, 500, and 500 conversations are selected for the single-round, multi-linguistic, and multi-round tracks, respectively, from an initial pool of 100K crowdsourced user-MLLM conversations. An overview is presented in Fig. 4.

## 2.1 Benchmark establishment

The benchmark is curated based on three guiding principles: i) diversity, selected user instruction queries target to avoid redundancies while extensively covering MLLM-based tasks; ii) MLLM-driven, the chosen queries of conversations are tailored to evaluate the unique capabilities of MLLMs in the multimodal domain; iii) coherence, the benchmark enables targeted evaluations for specific MLLM tasks, rather than providing undifferentiated evaluations. We first describe the common steps involved in curating the three tracks, followed by a discussion of the track-specific methodologies. **Common step.** We filter out short user instruction queries that contain excessive stop words, and apply MinHash-based text deduplication (Lee et al.,

2021) to retain a pool of non-redundant queries. To address potential redundancy or irrelevance between the instructions and images within a user query, we perform image-instruction deduplication. This step removes queries that can be sufficiently answered using only the textual instructions, leveraging an MLLM-based filter.

**Single-round track.** A language detector is employed to filter out non-English user instruction queries. Starting with a pool of MLLM task and sub-task fields derived from (Chen et al., 2024b), we use an MLLM-based annotator to assign user instruction queries to existing fields or propose new ones where necessary. Additionally, the annotator assesses the challenge level of each query. To ensure diversity, domain balancing is performed, and overrepresented task fields are downsampled, resulting in 3000 user instruction queries.

**Multi-linguistic track.** User instruction queries are categorized by their languages, excluding all English-based conversations. Based on frequency, the queries are grouped into Portuguese (PT), French (FR), Spanish (ES), German (DE), and an “Other” category (*e.g.*, Chinese, Vietnamese, and more). An MLLM-based annotator is then used to assess the challenges of the queries, with the 100 most difficult queries retained for each group.

**Multi-round track.** Similar to the single-round track, we focus on user instruction queries in English for this track. Multi-round conversations are required to feature interconnected queries across rounds, demonstrating a progressive nature. To achieve this, we identify the reasoning challenges and interdependencies between queries within the conversations, applying an MLLM annotator. Ultimately, the 100 most challenging independent queries and 400 interconnected multi-round user instruction queries are preserved.

Detailed prompts used for the above steps are provided in the supplementary material. With the ProBench, we are readily to assess and rank the MLLMs.

## 2.2 MLLM-as-a-Judge and ranking

We evaluate MLLM performance in addressing user instruction queries using a 5-point Likert scale (Likert, 1932), by conducting pairwise comparisons against a baseline model (*e.g.*, GPT-4o). While evaluations by domain-specific human experts are considered as the gold standard, they are resource-intensive, time-consuming, and challenging to scale for large-scale benchmarks. As an alter-

native, we employ MLLM-as-a-Judge as an approximation of human expertise (Li et al., 2024c; Zheng et al., 2023; Chen et al., 2024a). The MLLM-as-a-Judge is guided by the following principles.

- **Correctness:** ensures the accuracy of information, absence of factual errors, and alignments with known and visual knowledge. (For the multi-linguistic track, response language consistency is emphasized).
- **Helpfulness:** provides clear, practical, and actionable guidance to address the user instruction query.
- **Relevance:** focuses on the prompt requirements, avoiding extraneous or tangential information.
- **Conciseness:** avoids unnecessary verbosity while maintaining clarity and direct language.
- **Completeness:** covers all essential aspects of the user instruction query, providing sufficient information to address it.

Details of the prompts used to guide MLLM-as-a-Judge are provided in the supplementary material. Subsequently, we apply the ELO rating system, as described in the preliminary section, to compute the de-biased ratings of each MLLM. These ratings are used for leaderboard comparisons, ensuring a fair and consistent evaluation across models.

## 3 Experiment

### 3.1 Experimental setup

**Implementation detail.** All MLLMs are benchmarked using the vllm (Kwon et al., 2023) and Hugging Face (Wolf, 2019) codebases, with greedy sampling employed for response generation. For MLLMs with limited context lengths (*e.g.*, a 4096 token context in Molmo-7B-D-0924), sliding window generation is applied to handle longer inputs. Our MLLM judge utilizes gpt-4o-2024-08-06 with greedy sampling for consistent and reproducible evaluation. For pairwise comparisons in Elo rating calculations, we set gpt-4o-2024-05-13 as the baseline, evaluate each model twice by swapping the presentation order for each user query, and de-bias the ELO ratings by following the methodology of (Li et al., 2024c).

**MLLM.** We evaluate 24 leading MLLMs: gpt-4o-mini-2024-07-18 (Hurst et al., 2024), gpt-4o-2024-08-06 (Hurst et al., 2024), gpt-4o-2024-05-13 (Hurst et al., 2024), claude-3-5-sonnet-20241022 (Anthropic, 2024), gemini-1.5-pro-002 (Team et al., 2023), gemini-1.5-flash-002 (Team et al., 2023),



Table 1: Comparisons of state-of-the-art MLLMs on the single-round track are presented using the following abbreviations: Sci. (Science), Cd. (Coding), CW. (Creative Writing), IE. (Information Extraction), Perc. (Perception), Knowl. (Knowledge), Arts (Arts), Plan. (Planning), Math (Mathematics), and Mt. (Metrics). We provide ELO ratings for each task, followed by an overview that includes the average number of output tokens (#Token), 95% confidence interval (95% CI), win rate (WR), and overall ELO rating. The MLLMs are sorted by the overall ELO rating in each group of model size.

Model		Task-Specific ELO Ratings										Overview			
		Sci.	Cd.	CW.	IE.	Perc.	Knowl.	Arts	Plan.	Math.	Mt.	#Token	95% CI	WR	Elo
Proprietary MLLMs															
🌟 claude-3-5-sonnet-20241022	🔒	1228	1252	1259	1211	1213	1272	1236	1192	1197	1251	405	(-7, 8)	65.84	1228
🌈 gemini-1.5-pro-002	🔒	1151	1145	1105	1100	1110	1067	1107	1095	1134	1147	500	(-8, 10)	50.58	1118
🌀 gpt-4o-2024-05-13	🔒	1114	1114	1114	1114	1114	1114	1114	1114	1114	1114	491	(0, 0)	50.00	1114
🌀 gpt-4o-mini-2024-07-18	🔒	1049	1074	1165	1094	1096	1101	1130	1102	1037	1159	526	(-8, 10)	47.12	1094
🌀 gpt-4o-2024-08-06	🔒	1096	1112	1050	1097	995	1080	1032	1058	1175	1015	374	(-7, 7)	44.98	1079
🌈 gemini-1.5-flash-002	🔒	1025	877	1092	1007	1022	1011	993	946	1035	1087	493	(-8, 9)	35.33	1009
70B+ Open-source MLLMs															
📦 Pixtral-Large-Instruct-2411	124B	1230	1194	1280	1242	1224	1250	1245	1221	1175	1266	715	(-8, 8)	65.97	1229
🌊 InternVL2_5-78B	78B	1083	1018	1051	1091	1031	1084	1042	1073	1065	1023	558	(-7, 10)	42.85	1064
🌟 Qwen2-VL-72B-Instruct	72B	1009	914	965	991	986	960	962	921	998	970	557	(-9, 9)	31.37	978
📦 Molmo-72B-0924	72B	828	733	953	859	903	881	862	817	871	852	301	(-12, 8)	18.46	856
🌊 NVLM-D-72B	72B	780	877	991	810	849	835	767	881	838	725	561	(-10, 10)	16.63	834
🌊 Llama-3.2-90B-Vision-Instruct	90B	830	751	624	754	806	842	626	769	940	662	448	(-11, 10)	12.89	782
🌊 llava-onevision-qwen-72b-ov	72B	696	735	762	726	767	689	663	679	853	620	360	(-11, 12)	10.09	734
10B+ Open-source MLLMs															
📦 Pixtral-12B-2409	12B	1028	965	1099	1031	1024	1057	1047	1083	996	1063	659	(-5, 8)	39.1	1037
🌊 Aria-Chat	3.9/25.3B	990	982	985	937	998	1034	1019	974	973	1016	675	(-7, 8)	32.88	990
🌊 InternVL2_5-38B	38B	1000	979	1028	987	1021	904	932	1041	1026	933	521	(-9, 9)	32.5	987
🌊 InternVL2_5-26B	26B	890	816	1008	894	944	876	864	964	880	896	490	(-10, 8)	22.59	900
🌊 Llama-3.2-11B-Vision-Instruct	11B	671	541	681	702	766	761	624	524	744	614	531	(-13, 16)	7.93	688
7B+ Open-source MLLMs															
🌊 InternVL2_5-8B	8B	824	806	983	880	914	840	915	895	835	868	644	(-11, 8)	20.45	878
🌟 Qwen2-VL-7B-Instruct	7B	803	689	827	877	861	816	736	680	858	833	787	(-9, 10)	15.40	818
🌊 MiniCPM-V-2_6	8B	644	599	767	659	812	676	673	667	656	681	646	(-12, 10)	7.97	689
🌊 llava-onevision-qwen2-7b-ov	7B	605	570	807	683	809	681	715	608	573	724	575	(-13, 10)	7.93	688
📦 Molmo-7B-D-0924	7B	536	304	720	631	638	655	681	531	613	603	310	(-14, 12)	5.41	617
📦 Molmo-7B-O-0924	7B	457	134	623	483	681	599	606	380	428	528	296	(-18, 19)	3.54	540

Aria-Chat (Li et al., 2024b), InternVL2\_5-8B (Wang et al., 2024b), InternVL2\_5-26B (Wang et al., 2024b), InternVL2\_5-38B (Wang et al., 2024b), InternVL2\_5-78B (Wang et al., 2024b), Pixtral-12B-2409 (Agrawal et al., 2024), Pixtral-Large-Instruct-2411 (Agrawal et al., 2024), Qwen2-VL-7B-Instruct (Wang et al., 2024a), Qwen2-VL-72B-Instruct (Wang et al., 2024a), MiniCPM-V-2\_6 (Yao et al., 2024), Llama-3.2-11B-Vision-Instruct (Dubey et al., 2024), Llama-3.2-90B-Vision-Instruct (Dubey et al., 2024), Molmo-7B-O-0924 (Deitke et al., 2024), Molmo-7B-D-0924 (Deitke et al., 2024), Molmo-72B-0924 (Deitke et al., 2024), NVLM-D-72B (Dai et al., 2024), llava-onevision-qwen2-7b-ov (Li et al., 2024a), and llava-onevision-qwen2-72b-ov (Li et al., 2024a).

### 3.2 Experimental result

Tab. 1 and Tab. 2 present the evaluation results. Our key observations are summarized into the following five folds: i) **best open-source models rival the best proprietary MLLMs**. claude-3-5-sonnet-20241022 and Pixtral-Large-Instruct-2411 respectively belonging to proprietary and open-source MLLMs consistently achieve leading ELO scores across all three tracks. Both models significantly outper-

form the baseline gpt-4o-2024-05-13; ii) **training recipe outweighs model size**. While scaling parameters can generally enhance performance, the performance of MLLMs can be size-agnostic, with greater emphasis placed on the training recipe (i. e., optimization stratge and training data quality). For example, Pixtral with 12B parameters and Aria-Chat with 3.9B activated parameters (out of a total of 25.3B) per token consistently demonstrate first-tier performance; iii) **reasoning tasks remain the hardest**. On the single-round track, most MLLMs generally perform well on writing-based tasks (e.g., creative writing). However, their performance on logic-intensive tasks is notably poor, similar to findings in prior LLM studies (Ahn et al., 2024; Quan et al., 2025). The two tasks separately exhibit the lowest Spearman correlation with overall ELO ratings and receive the lowest scores among task fields. Similarly, among all open-source models, performance also suffers significantly in planning tasks, which have the lowest average score (excluding coding); iv) **multi-linguistic tasks challenge MLLMs**. MLLMs face significant challenges in multi-linguistic tasks, with 11 out of 24 MLLMs showing an overall ELO decrease compared to their performance on the single-round track. Notably,

Table 2: Comparisons of state-of-the-art MLLMs on the multi-linguistic and multi-round tracks. We provide an overview that shows the average number of output tokens (#Token), 95% confidence interval (95% CI), win rate (WR), and overall ELO rating for each of the track. Refer to our supplementary material for comparison details on different languages and rounds. The MLLMs are sorted by the overall ELO rating on the multi-linguistic track in each group of model size.

Model	Overview on multi-linguistic track					Overview on multi-round track			
		#Token	95% CI	WR	Elo	#Token	95% CI	WR	Elo
Proprietary MLLMs									
🌟 claude-3-5-sonnet-20241022	🔒	485	(-21, 29)	74.58	1301	1477	(-20, 18)	70.82	1268
🌐 gpt-4o-2024-05-13	🔒	585	(0, 0)	50.00	1114	1563	(0, 0)	50.00	1114
🌐 gemini-1.5-pro-002	🔒	629	(-20, 20)	59.11	1178	1425	(-26, 19)	53.88	1141
🌐 gpt-4o-2024-08-06	🔒	480	(-17, 26)	60.35	1187	1052	(-22, 18)	45.41	1082
🌐 gpt-4o-mini-2024-07-18	🔒	657	(-21, 16)	45.84	1085	1749	(-17, 24)	55.16	1150
🌐 gemini-1.5-flash-002	🔒	567	(-25, 19)	28.47	954	1388	(-16, 19)	38.14	1030
70B+ Open-source MLLMs									
📺 Pixtral-Large-Instruct-2411	124B	966	(-23, 22)	73.81	1294	2593	(-23, 19)	69.73	1259
📺 Qwen2-VL-72B-Instruct	72B	834	(-18, 21)	47.56	1097	1608	(-21, 19)	32.24	985
📺 InternVL2_5-78B	78B	841	(-14, 20)	42.71	1063	2015	(-21, 20)	44.84	1078
📺 NVLM-D-72B	72B	907	(-17, 25)	21.99	894	1371	(-35, 33)	8.49	701
📺 Llama-3.2-90B-Vision-Instruct	90B	968	(-29, 21)	20.92	883	1350	(-36, 24)	9.88	730
📺 Molmo-72B-0924	72B	426	(-27, 19)	18.90	861	967	(-28, 25)	18.64	858
📺 llava-onevision-qwen2-72b-ov	72B	534	(-27, 24)	11.95	767	1176	(-31, 26)	10.30	738
10B+ Open-source MLLMs									
📺 InternVL2_5-38B	38B	868	(-20, 18)	43.98	1072	1734	(-18, 21)	34.68	1004
📺 Pixtral-12B-2409	12B	1199	(-14, 22)	35.73	1012	2264	(-19, 20)	40.48	1047
📺 Aria-Chat	3.9/25.3B	1014	(-23, 17)	35.33	1009	2321	(-27, 12)	23.92	913
📺 InternVL2_5-26B	26B	814	(-28, 19)	17.70	847	554	(-27, 28)	15.77	823
📺 Llama-3.2-11B-Vision-Instruct	11B	2027	(-29, 21)	8.40	699	2094	(-38, 32)	6.03	637
7B+ Open-source MLLMs									
📺 Qwen2-VL-7B-Instruct	7B	1216	(-24, 22)	12.25	772	2004	(-34, 25)	9.48	722
📺 InternVL2_5-8B	8B	1021	(-22, 20)	11.95	767	1835	(-25, 22)	11.77	764
📺 MiniCPM-V-2_6	8B	890	(-36, 35)	4.44	581	1861	(-33, 37)	5.35	615
📺 Molmo-7B-D-0924	7B	406	(-52, 33)	4.32	576	923	(-34, 26)	5.04	604
📺 llava-onevision-qwen2-7b-ov	7B	686	(-68, 37)	3.07	514	1743	(-30, 30)	6.58	653
📺 Molmo-7B-O-0924	7B	512	(-73, 51)	1.95	433	925	(-49, 37)	3.43	534

llava-onevision-qwen2-7b-ov experienced the most substantial decline; v) **multi-round evaluation enhances model performance separability**. Multi-round tasks usually demand long-context reasoning across turns, amplifying performance gaps among MLLMs. MLLMs that underperform in single-round tasks exhibit significantly lower ELO scores. This trend is particularly evident in open-source MLLMs with 7B+ and 10B+ parameters (excluding Pixtral-12B-2409).

### 3.3 Ablation and discussion

**Performance declining with difficulty.** We evaluate the ELO rating variances of MLLMs by categorizing user queries into easy and hard groups. The results are presented in Fig. 5. Existing MLLMs tend to exhibit a noticeable performance decline compared to the baseline gpt-4o-2024-05-13 as the reasoning challenge level increased from easy to hard, while MLLM with poor performance typically deteriorates further on the harder queries. This observation aligns with human intuition that more challenging tasks inherently provide better separability when evaluating the MLLM performance, highlighting the limitations of most MLLMs in effectively handling complex user queries.

**Error analysis.** We analyze scenarios in which the state-of-the-art MLLM underperforms relative to the baseline. Fig. 6 (a) illustrates the shortcomings of the MLLM compared to the baseline across five evaluation aspects, highlighting completeness and correctness as the primary issues. Fig. 6 (b) categorizes the error types in the MLLM losses relative to the baseline. Overall, the analysis underscores the need of state-of-the-art MLLM to improve their visual perception, textual understanding, domain knowledge, and reasoning capability.

**Robustness of ProBench.** We study the setting of our evaluation protocol on the 500 most challenging queries from the single-round track. Specifically, Fig. 7 considers two set of experiments: i) comparisons of using three top-performing MLLM as the judge (i.e., gpt-4o-2024-08-06, claude-3-5-sonnet-20241022, and Pixtral-Large-Instruct-2411); ii) explorations of three baseline models (i.e., gpt-4o-2024-05-13, claude-3-5-sonnet-20241022, and Pixtral-12B-2409) in comparisons, representing different model scales. The results reveal a high degree of agreement within our evaluation process, with an average Spearman correlation coefficient of 0.979 among the different MLLM judges and 0.983 among the baseline models, highlighting our robustness and consistency.

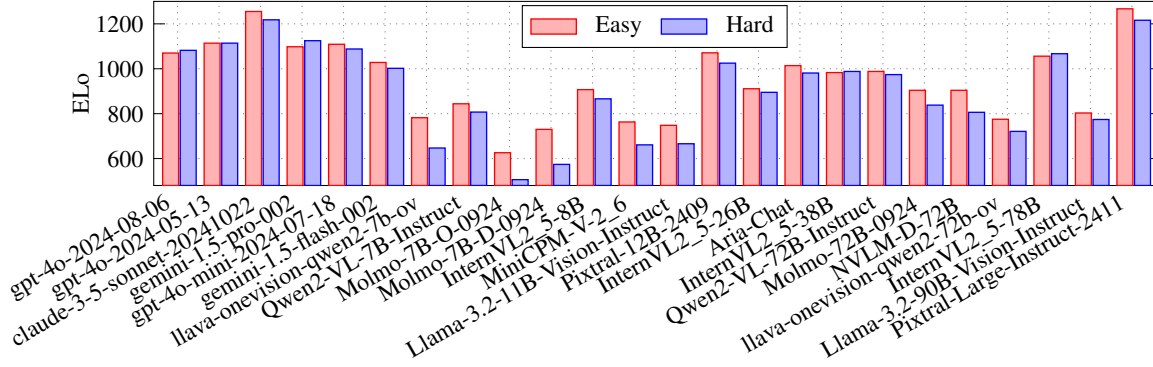


Figure 5: Ablation study of reasoning challenge. We show the ELO ratings of MLLMs on two levels: easy and hard.

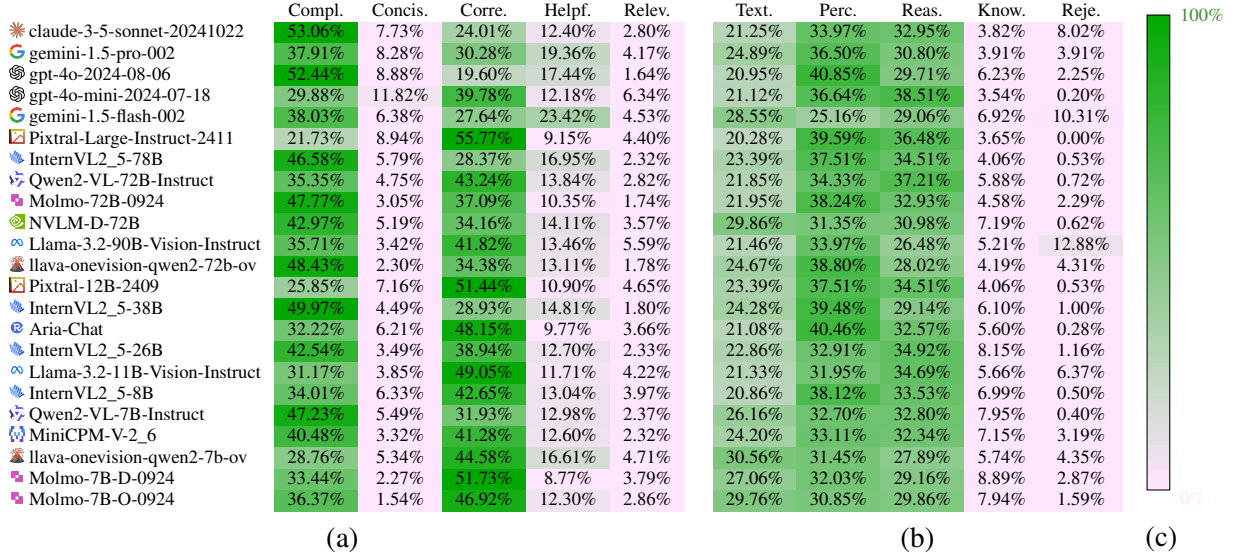


Figure 6: Error analysis. We study cases where MLLM underperforms compared to the baseline. (a) The distribution of losing cases of the MLLM across five evaluation aspects: completeness (Compl.), conciseness (Concis.), correctness (Corre.), helpfulness (Helpf.), and relevance (Relve.). (b) The distribution of error types in losses of the MLLM, categorized into five types: textual understanding error (Text.), visual perceptual error (Perc.), reasoning error (Reas.), lack of domain knowledge error (Know.), and refusal to answer (Reje.). (c) Color bar of the heatmap.

**Judge alignment with human expert.** To validate the effectiveness of MLLM-as-a-Judge, human annotators are tasked with rating the comparisons using a 5-point Likert scale. Our evaluation protocol achieves an agreement of 79.9% with human expert, indicating a strong ability of MLLM-as-a-Judge to simulate human preferences accurately. These findings demonstrate the viability of ProBench as an automatic, large-scale, and challenging benchmark for evaluating the assistance capabilities of MLLMs in professional productivity scenarios. By effectively aligning with human judgments, ProBench provides a reliable automatic framework for advancing MLLM development and assessment.

**Future work and limitation.** Although our ProBench has provided valuable insights into the performance and capabilities of MLLMs, several limitations remain that warrant further exploration.

One key limitation is potential bias in the benchmark tasks, which may not fully capture the diversity of real-world productivity scenarios for MLLMs. Future work could focus on expanding the benchmark to include a broader range of challenging tasks, potentially through the data synthesis (*e.g.*, diffusion models and MLLMs), to improve the diversity. By addressing these challenges, ProBench can continue to evolve as a robust and comprehensive tool for advancing the development and evaluation of MLLMs.

### 3.4 Distilled local evaluator

Considering the high API cost of using gpt-4o-2024-08-06 as the judge, we fine-tune a local evaluator to enable cost-effective and GPU-friendly evaluations for future MLLMs. We use the widely spread Llama-3.2-11B-Vision-Instruct as our backbone model. The Qwen and Pixtral MLLM fam-

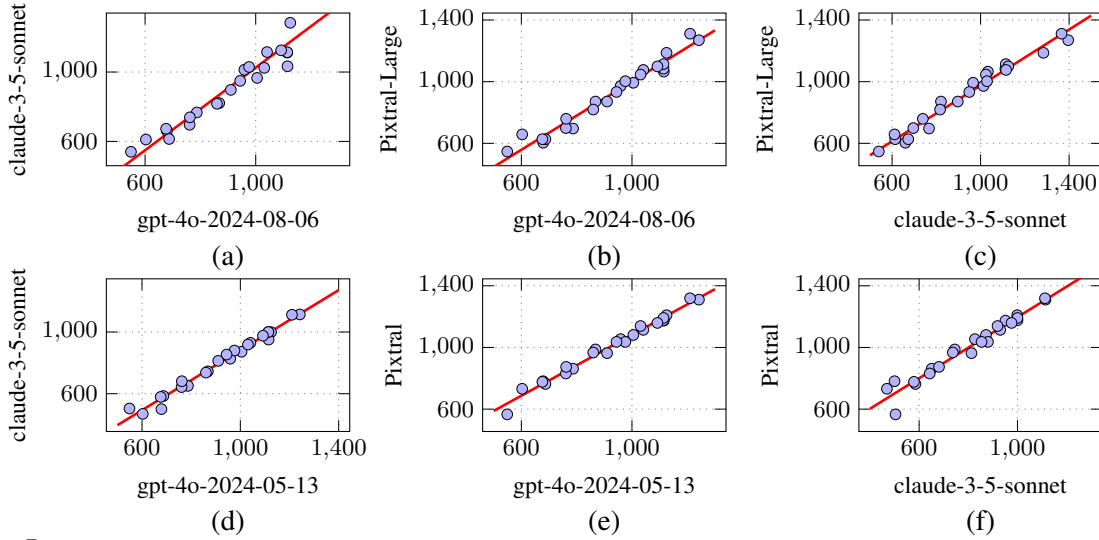


Figure 7: Ablation study of MLLM-as-the-Judge. (a-c) Pairwise comparisons of Elo scores for MLLMs evaluated using different MLLM judges. They are gpt-4o-2024-08-06, claude-3-5-sonnet-20241022 (claude-3-5-sonnet), and Pixtral-Large-Instruct-2411 (Pixtral-Large), respectively. (d-f) Comparison of using gpt-4o-2024-05-13, claude-3-5-sonnet-20241022 (claude-3-5-sonnet), and Pixtral-12B-2409 (Pixtral) as baselines. The red line in each plot indicates the best-fit curve for visualization.

ilies are reserved for testing, with the remaining data allocated for training. Our network is trained to distill both the reasoning and decisions of using gpt-4o-2024-08-06 as the judge. The network achieves an average root mean squared error of 32.58 in Elo ratings.

#### 4 Related work

The evolution of MLLM-as-a-Judge is largely inspired by the concept of LLM-as-a-Judge (Li et al., 2024c; Dubois et al., 2024; Zheng et al., 2023), which aims to automatically measure the alignment between MLLMs and human preferences. While pairwise comparison (Li et al., 2024c; Chen et al., 2024a) is considered as most preferred, it suffers from biases introduced by factors such as the presentation order of MLLM outputs, verbosity, and markdown styles. To mitigate these issues, style control has been proposed (Li et al.), using statistical modeling to de-bias these confounding effects, thereby improving the MLLM judges.

Other approaches, such as few-shot judging, have also been explored, but they face challenges such as reliance on the few-shot example selection and increased evaluation costs (Zheng et al., 2023). Existing MLLM-as-a-Judge leaderboards can be specified to (Luo et al., 2024; Lu et al., 2024; Chen et al., 2024a). However, these often focus on a narrow scope of MLLM capability dimensions (Luo et al., 2024; Lu et al., 2024), or rely on artificially posed evaluations by a limited number of human experts (Chen et al., 2024b), making them inadequate

for assessing MLLMs on professional tasks. Consequently, they fail to capture the dynamic nature of real-world human and MLLM interactions for a comprehensive assessment of MLLM capabilities. In contrast, this work introduces a challenging benchmark, ProBench, curated from large-scale crowdsourced datasets reflecting real-world professional productivity scenarios. It features three distinct evaluation tracks: single-round, multi-round, and multi-linguistic conversations, across various task fields, offering a robust framework for evaluating MLLM performance in real-world scenarios.

#### 5 Conclusion

This paper introduces the ProBench, which features single-round, multi-round, and multi-linguistic tracks to enable a comprehensive and challenging assessment of the alignment between MLLMs and human preferences across diverse professional productivity demands. By employing MLLM-as-a-Judge, the benchmark evaluates MLLM pairwise, achieving 79.9% agreement with human expert judgments, and underscoring its reliability. Through benchmarking 24 leading MLLMs, our results reveal significant shortcomings of existing MLLMs, particularly in visual perception and reasoning. Furthermore, models often struggle with multi-linguistic and multi-round tracks, highlighting the challenges of diverse language requirement and complex interactions. It reveals valuable insights for future MLLM developments. We hope it inspires successors.



## References

- Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.
- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.
- AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 3.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinyu Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*.
- Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuan-sheng Ni, Wang Zhu, Ziyang Jiang, Bohan Lyu, et al. 2024b. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*.
- Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamäki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.
- Arpad E Elo. 1966. *The USCF Rating System: Its Development, Theory, and Applications*. United States Chess Federation.
- Freepik, Eucalyp, Three Musketeers, Dewi Sari, Fantasy, Jk Icon, and Flat Icons. 2025. [Various icons](#).
- David R Hunter. 2004. Mm algorithms for generalized bradley-terry models. *The annals of statistics*, 32(1):384–406.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. 2024b. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*.
- Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. Does style matter? disentangling style and substance in chatbot arena, august 2024a. [URL https://blog.lmarena.ai/blog/2024/style-control](https://blog.lmarena.ai/blog/2024/style-control).
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024c. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*.
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mm-bench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv e-prints*, pages arXiv–2310.
- Yujie Lu, Dongfu Jiang, Wenhui Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wildvision: Evaluating vision-language models in the wild with human preferences. *arXiv preprint arXiv:2406.11069*.

Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. 2024. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. *arXiv preprint arXiv:2411.13281*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, et al. 2025. Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings. *arXiv preprint arXiv:2501.01257*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.

T Wolf. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

## A Experimental detail

We respectively present detailed comparisons of multi-linguistic and multi-round tracks in Tab. 3 and Tab. 4.

The optimization details for tuning a local evaluator based on Llama-3.2-11B-Vision-Instruct are provided below. We use a learning rate of  $1 \times 10^{-5}$  for both the projector and the LLM, while setting a lower learning rate of  $2 \times 10^{-6}$  for the vision encoder. The context length is set to 128K. A cosine annealing strategy with a 3% warm-up of the total optimization steps is employed. The AdamW optimizer is used with  $\beta_1 = 0.9$  and  $\beta_2 = 0.95$ , along with a weight decay of 0.03. We train with a batch size of 16 for 20K optimization steps. The model is trained using 16 H100 GPUs, with the training process taking approximately 2 days.

For evaluation with MLLM-as-the-Judge, the largest models require around two days for response generation on 8 GPUs, while evaluation with the local evaluator takes about one day using 2 GPUs.

All data from ProBench has been collected with explicit user consent.

## B Prompt template

We present the prompts for curating the single-round, multi-linguistic, and multi-round tracks, as well as for utilizing MLLM-as-a-Judge across the three tracks: Tab. 5, Tab. 6, Tab. 7, and Tab. 4 provide prompts for categorizing task and sub-task fields related to user instruction queries; Tab. 5 and Tab. 6 present prompts for evaluating challenges within user instruction queries; Tab. 7 and Tab. 8 are prompts for deduplications between visual and textual content in user instruction queries (i. e., image-instruction deduplication); Tab. 9 offers prompts for assessing interdependencies among multi-round user instruction queries; Tab. 10, Tab. 11, and Tab. 12 respectively give the prompts of MLLM-as-a-Judge for the three tracks.

## C Human preference evaluation

To assess the agreements and reliability of MLLM-as-a-Judge, we evaluate the alignment between human annotators and gpt-4o-2024-08-06 as a judge. All participants are volunteers who have been informed about the purpose of the study and have provided consent to share their data. In this experiment, a random sample of 300 responses is drawn from the ProBench dataset. These responses

are then evaluated by six human annotators, each tasked with comparing the outputs of two MLLMs for addressing the user instruction queries.

On average, each comparison took approximately 90.6 seconds. In contrast, the MLLM-as-a-Judge method completes the task in just a few seconds via an API call, highlighting the superior speed and efficiency of model-based evaluation. The annotation interface used for this task is shown in Fig. 8. Overall, we observe 79.9% agreement between human annotators and the MLLM-as-a-Judge. Refer to Fig. 9 that illustrates the distribution of human annotators preferences, and MLLM preferences, and human annotation time cost.

## D Analysis

In Fig. 10, we further present the distributions of image distribution, textual challenges, image challenges, and reasoning challenges across the user instruction queries. Tab. 13 provides 30 randomly sampled examples of MLLM-as-a-Judge evaluations.

Table 3: Comparisons of state-of-the-art MLLMs on the multi-linguistic track are presented using the following abbreviations: PT (Portuguese), FR (French), ES (Spanish), DE (German), and an “Other” category (e.g., Chinese, Vietnamese, and more). We provide ELO ratings for each language, followed by an overview that includes the average number of output tokens (#Token), 95% confidence interval (95% CI), win rate (WR), and overall ELO rating. The MLLMs are sorted by the overall ELo rating in each group.

Model		Language-Specific ELO Ratings					Overview			
		PT	FR	ES	DE	Other	#Token	95% CI	WR	Elo
Proprietary MLLMs										
🌟 claude-3.5-sonnet-20241022	🔒	1248	1319	1335	1389	1309	485	(-21, 29)	74.58	1301
🌀 gpt-4o-2024-05-13	🔒	1114	1114	1114	1114	1114	585	(0, 0)	50.0	1114
🌈 gemini-1.5-pro-002	🔒	1273	1168	1131	1168	1139	629	(-20, 20)	59.11	1178
🌀 gpt-4o-2024-08-06	🔒	1159	1224	1226	1259	1114	480	(-17, 26)	60.35	1187
🌀 gpt-4o-mini-2024-07-18	🔒	1038	1079	1071	1151	1099	657	(-21, 16)	45.84	1085
🌈 gemini-1.5-flash-002	🔒	1031	990	845	1015	815	567	(-25, 19)	28.47	954
70B+ Open-source MLLMs										
📺 Pixtral-Large-Instruct-2411	124B	1229	1496	1216	1324	1286	966	(-23, 22)	73.81	1294
🔗 Qwen2-VL-72B-Instruct	72B	1067	1199	944	1241	999	834	(-18, 21)	47.56	1097
🔗 InternVL2_5-78B	78B	948	1125	1035	1123	1084	841	(-14, 20)	42.71	1063
🔗 NVLM-D-72B	72B	900	863	850	898	918	907	(-17, 25)	21.99	894
🌀 Llama-3.2-90B-Vision-Instruct	90B	905	860	824	863	864	968	(-29, 21)	20.92	883
📺 Molmo-72B-0924	72B	834	835	852	853	878	426	(-27, 19)	18.9	861
📺 llava-onevision-qwen2-72b-ov	72B	782	810	609	800	729	534	(-27, 24)	11.95	767
10B+ Open-source MLLMs										
🔗 InternVL2_5-38B	38B	1038	1092	1070	1100	1044	868	(-20, 18)	43.98	1072
📺 Pixtral-12B-2409	12B	935	1096	998	1077	929	1199	(-14, 22)	35.73	1012
🔗 Aria-Chat	3.9/25.3B	964	1042	983	1041	999	1014	(-23, 17)	35.33	1009
🔗 InternVL2_5-26B	26B	779	858	782	880	839	814	(-28, 19)	17.7	847
🌀 Llama-3.2-11B-Vision-Instruct	11B	714	663	626	627	665	2027	(-29, 21)	8.4	699
7B+ Open-source MLLMs										
🔗 Qwen2-VL-7B-Instruct	7B	701	875	673	865	678	1216	(-24, 22)	12.25	772
🔗 InternVL2_5-8B	8B	760	776	765	821	602	1021	(-22, 20)	11.95	767
🔗 MiniCPM-V-2_6	8B	522	559	603	634	455	890	(-36, 35)	4.44	581
📺 Molmo-7B-D-0924	7B	445	495	577	613	505	406	(-52, 33)	4.32	576
📺 llava-onevision-qwen2-7b-ov	7B	579	386	144	403	588	686	(-68, 37)	3.07	514
📺 Molmo-7B-O-0924	7B	383	256	536	246	429	512	(-73, 51)	1.95	433

Table 4: Comparisons of state-of-the-art MLLMs on the multiround track are presented. We provide ELO ratings for rounds with lengths of 2, 3, 4, 5, and more than 6 (6+), followed by an overview that includes the average number of output tokens (#Token), 95% confidence interval (95% CI), win rate (WR), and overall ELO rating. ‘N/A’ indicates cases where the model did not apply, as it lost to gpt-4o-2024-05-13 across all samples. The MLLMs are sorted by the overall ELo rating in each group

Model		Round-Specific ELO Ratings					Overview			
		2	3	4	5	6+	#Token	95% CI	WR	Elo
Proprietary MLLMs										
🌟 claude-3.5-sonnet-20241022	🔒	1260	1249	1356	1248	1321	1477	(-20, 18)	70.82	1268
🌀 gpt-4o-2024-05-13	🔒	1114	1114	1114	1114	1114	1563	(0, 0)	50.0	1114
🌈 gemini-1.5-pro-002	🔒	1136	1140	1107	1207	1145	1425	(-26, 19)	53.88	1141
🌀 gpt-4o-2024-08-06	🔒	1146	1050	1138	1023	965	1052	(-22, 18)	45.41	1082
🌀 gpt-4o-mini-2024-07-18	🔒	1147	1143	1142	1200	1151	1749	(-17, 24)	55.16	1150
🌈 gemini-1.5-flash-002	🔒	1015	1040	1015	1119	1006	1388	(-16, 19)	38.14	1030
70B+ Open-source MLLMs										
📺 Pixtral-Large-Instruct-2411	124B	1233	1273	1304	1376	1253	2593	(-23, 19)	69.73	1259
🔗 Qwen2-VL-72B-Instruct	72B	1023	972	1033	936	875	1608	(-21, 19)	32.24	985
🔗 InternVL2_5-78B	78B	1135	1040	1148	1015	992	2015	(-21, 20)	44.84	1078
🔗 NVLM-D-72B	72B	770	557	602	641	682	1371	(-35, 33)	8.49	701
🌀 Llama-3.2-90B-Vision-Instruct	90B	754	757	784	426	605	1350	(-36, 24)	9.88	730
📺 Molmo-72B-0924	72B	886	817	787	920	808	967	(-28, 25)	18.64	858
📺 llava-onevision-qwen2-72b-ov	72B	753	721	673	525	692	1176	(-31, 26)	10.3	738
10B+ Open-source MLLMs										
🔗 InternVL2_5-38B	38B	1003	1037	1036	913	902	1734	(-18, 21)	34.68	1004
📺 Pixtral-12B-2409	12B	1054	1008	1160	1013	1035	2264	(-19, 20)	40.48	1047
🔗 Aria-Chat	3.9/25.3B	937	913	946	887	812	2321	(-27, 12)	23.92	913
🔗 InternVL2_5-26B	26B	881	811	805	753	638	1554	(-27, 28)	15.77	823
🌀 Llama-3.2-11B-Vision-Instruct	11B	741	380	487	275	490	2094	(-38, 32)	6.03	637
7B+ Open-source MLLMs										
🔗 Qwen2-VL-7B-Instruct	7B	808	622	637	557	495	2004	(-34, 25)	9.48	722
🔗 InternVL2_5-8B	8B	814	724	775	686	559	1835	(-25, 22)	11.77	764
🔗 MiniCPM-V-2_6	8B	664	575	628	530	389	1861	(-33, 37)	5.35	615
📺 Molmo-7B-D-0924	7B	672	470	523	409	618	923	(-34, 26)	5.04	604
📺 llava-onevision-qwen2-7b-ov	7B	737	591	649	N/A	512	1743	(-30, 30)	6.58	653
📺 Molmo-7B-O-0924	7B	589	413	490	N/A	402	925	(-49, 37)	3.43	534



Table 5: The prompt for identifying user instruction query task fields.

**[System]**

You are an AI assistant tasked with classifying a user-provided question and image into predefined categories. The question should be classified based on both the text of the question and the image provided, while the image classification should be based solely on the visual content of the image. Your responsibilities are:

1. Analyze the question and classify it under one category from the following list:
  - Coding: Focuses on code-related tasks such as debugging, generating, translating, and understanding programming logic.
  - Information Extraction: Involves tasks like extracting and analyzing details from data, structured parsing, summarization, and multimodal Q&A.
  - Knowledge: Covers arts, culture, fact-checking, and understanding diverse global and historical knowledge.
  - Mathematics: Includes problem-solving in algebra, calculus, geometry, number theory, graph theory, and numeric reasoning.
  - Metrics: Evaluates quality and performance in images, videos, papers, and other models or generated content.
  - Perception: Encompasses tasks like 3D understanding, image segmentation, multimodal captioning, and object or scene understanding.
  - Planning: Deals with creating strategies for agents, solving puzzles, reordering tasks, and planning complex processes.
  - Science: Applies to specialized domains like chemistry, physics, life sciences, and STEM-related problem-solving.
  - Creative Writing: Covers character development, storytelling, poetry, dialogue, scriptwriting, and worldbuilding across genres.
  - Arts and Humanities: Involves creative and cultural exploration, metaphorical thinking, narrative techniques, and genre-specific expression.
2. Classify the image into one of the main categories:
  - Document and Text-based Images: Includes scanned documents, forms, tables, and charts, used for record-keeping, data presentation, or analysis.
  - Medical Images: Diagnostic visuals like MRIs, X-rays, and pathology slides, used in healthcare and medical research.
  - Photographs: Everyday pictures, portraits, and landscapes captured with cameras, often for personal or professional use.
  - Scientific and Analytical Images: Specialized visuals like

microscopic, astronomical, or spectrogram images for research and technical analysis.

- Graphics and Artistic Images: Includes infographics, logos, cartoons, and illustrations for creative, branding, or informative purposes.
- Screenshots and UI Elements: Captures of websites, apps, or software interfaces for documentation or demonstration.
- Remote Sensing and Satellite Images: Aerial and satellite photos for mapping, monitoring, or geographic analysis.
- Security and Surveillance: CCTV footage and thermal imaging for safety, monitoring, or investigative purposes.
- Engineering and Technical Drawings: CAD designs, blueprints, and 3D models for architectural or engineering applications.
- Specialized Formats: Includes barcodes, QR codes, fingerprints, and AR/VR visuals for unique or advanced use cases.

3. If the question or image does not fit existing categories, propose a new category with justification.

4. Do not generate the answer for the user question.

Your response should be in JSON format:

```
{
  "thinking_image": "Reasoning for your classification of
image.",
  "image_category": "The category of the image."
  "thinking_question": "Reasoning for your classification of
question.",
  "question_category": "The category of the user question.",
}
```

750

Table 6: The prompt for identifying user instruction query sub-task fields.

#### **[System]**

You are an AI assistant tasked with further classifying a user-provided question and image into sub-categories. The question should be classified based on both the text of the question and the image provided, while the image classification should be based solely on the visual content of the image. Your responsibilities are:

1. **\*\*Question Classification\*\***:

- Analyze the question and assign it to the most relevant sub-category based on its content.

751

- The question belongs to the main category "{question\_category}" and should be classified into one of the following sub-categories:  
    {question\_subcats\_formatted}

2. **\*\*Image Classification\*\***:

- Analyze the image and assign it to the most relevant sub-category based solely on its visual content.  
- The image belongs to the main category "{image\_category}" and should be classified into one of the following sub-categories:  
    {image\_subcats\_formatted}

3. If the question or image does not fit any of the above sub-categories, propose a new sub-category and provide a justification.

4. Do not generate the answer for the user question.

Your response must be structured in the following JSON format:

```
{{
  "thinking_image": "Reasoning for the image sub-category classification.",
  "image_subcategory": "The sub-category for the image.",
  "thinking_question": "Reasoning for the question sub-category classification.",
  "question_subcategory": "The sub-category for the question.",
}}
```

Table 7: The task and sub-task fields for user instruction queries (*e.g.*, questions). For consistency, the naming convention aligns with Tab. 6. `question_category` represents the task field, while `question_subcats_formatted` denotes the task sub-field.

<code>question_category</code>	<code>question_subcats_formatted</code>
Information Extraction	<ul style="list-style-type: none"> <li>* App Function Understanding: Analyzing and interpreting the purpose, features, and functionality of an application.</li> <li>* Summarization: Condensing detailed information into a concise form while preserving key points and context.</li> <li>* Entity Recognition: Identifying and categorizing specific elements such as names, dates, locations, or organizations.</li> <li>* Relationship Mapping: Identifying and visualizing the connections or associations between different entities.</li> <li>* Contextual Analysis: Understanding the meaning, intent, or relevance of data within its specific context.</li> </ul>
Creative Writing	<ul style="list-style-type: none"> <li>* Storytelling: Developing compelling and engaging narratives for readers or audiences.</li> <li>* Scriptwriting: Creating scripts for various media formats, including films, television, and plays.</li> <li>* Worldbuilding: Designing intricate and immersive fictional settings, universes, or environments.</li> <li>* Character Development: Creating, evolving, and deepening the personalities and arcs of fictional characters.</li> <li>* Plot Structuring: Organizing the sequence of events and narrative flow to build tension, conflict, and resolution.</li> </ul>
Science	<ul style="list-style-type: none"> <li>* Physics: The exploration of forces, motion, energy, and the fundamental nature of the universe.</li> <li>* Biology: The study of living organisms, their functions, and interactions within ecosystems.</li> <li>* Astronomy: The observation and study of celestial objects, space, and the physical universe as a whole.</li> <li>* Life Science/Medical: The study of biological and medical sciences, including anatomy, physiology, and healthcare-related topics.</li> <li>* STEM Problem-Solving: Using interdisciplinary approaches to tackle technical and scientific challenges.</li> </ul>

*Continued on next page...*



question_category	question_subcats_formatted
Knowledge	<ul style="list-style-type: none"> <li>* Human and Culture: Insights into human behavior, societal structures, traditions, and cultural practices.</li> <li>* Scientific Knowledge: Understanding and explaining scientific concepts, theories, and principles across disciplines.</li> <li>* World Knowledge: General information about global geography, politics, economies, and cultures.</li> <li>* Fact-Checking: Verifying the accuracy of information and identifying misinformation or inaccuracies.</li> <li>* Philosophical Inquiry: Exploring existential, ethical, and metaphysical questions to gain deeper understanding.</li> </ul>
Metrics	<ul style="list-style-type: none"> <li>* Model Performance: Assessing the accuracy, efficiency, and reliability of algorithms or machine learning models.</li> <li>* Paper Review: Critiquing and analyzing research papers for quality, relevance, and scientific rigor.</li> <li>* Content Evaluation: Judging the quality, coherence, and relevance of generated or provided content.</li> <li>* Quality Assessment: Measuring and determining the overall standard or quality of various outputs or systems.</li> <li>* Reward Models: Designing and evaluating models that provide feedback or incentives for optimizing performance in systems.</li> </ul>
Coding	<ul style="list-style-type: none"> <li>* Code Generation: Creating new code based on given requirements, templates, or problem-solving scenarios.</li> <li>* Code Translation: Converting code from one programming language or framework to another.</li> <li>* Code Optimization: Enhancing the efficiency, readability, and performance of existing code.</li> <li>* Code Understanding: Interpreting and explaining the purpose, logic, or functionality of code.</li> </ul>

*Continued on next page...*

question_category	question_subcats_formatted
Perception	<ul style="list-style-type: none"> <li>* Counting: Identifying and quantifying the number of objects or elements in an image or scene.</li> <li>* Multimodal Captioning: Generating descriptive captions by combining visual and textual data for an enriched understanding.</li> <li>* Object Understanding: Recognizing, categorizing, and interpreting the attributes and roles of objects in visual content.</li> <li>* Scene Understanding: Comprehending the arrangement, context, and interactions within a visual scene.</li> <li>* Diagram and Document Understanding: Interpreting and extracting information from diagrams, charts, or text-based documents.</li> </ul>
Arts and Humanities	<ul style="list-style-type: none"> <li>* Cultural Analysis: Examining societal norms and values.</li> <li>* Narrative Techniques: Exploring storytelling methods.</li> <li>* Genre-Specific Writing: Crafting work within specific literary or artistic genres.</li> </ul>
Mathematics	<ul style="list-style-type: none"> <li>* Calculus: Analyzing rates of change and accumulation using derivatives and integrals.</li> <li>* Function: Studying relationships between inputs and outputs, represented mathematically.</li> <li>* Geometry: Exploring shapes, sizes, dimensions, and the properties of space.</li> <li>* Graph Theory: Analyzing the relationships between nodes and edges in a network or graph.</li> <li>* Number Theory: Investigating the properties, patterns, and relationships of numbers, especially integers.</li> <li>* Statistics/Numerical Reasoning: Interpreting, analyzing, and presenting data to draw logical inferences and conclusions.</li> </ul>
Planning	<ul style="list-style-type: none"> <li>* Reordering: Resequencing tasks or events to optimize efficiency and effectiveness.</li> <li>* Puzzle Solving: Finding logical or creative solutions to abstract, conceptual, or practical challenges.</li> <li>* Game Strategy: Developing tactics, plans, and approaches to achieve success in game environments.</li> <li>* Complex Workflow Design: Designing and managing intricate, multi-step processes to accomplish complex tasks or objectives.</li> </ul>

*Continued on next page...*

question_category	question_subcats_formatted
Other	Unspecified or generic category.

Table 4: The field and sub-field for images in user instruction queries. For consistency, the naming convention aligns with Tab. 6. image\_category represents the image field, while image\_subcats\_formatted denotes the image sub-field.

image_category	image_subcats_formatted
Screenshots and UI Elements	<ul style="list-style-type: none"> <li>* Mobile App UI: User interfaces for mobile applications.</li> <li>* Desktop Applications: Screenshots of software interfaces.</li> <li>* Game Interfaces: Displays from video games.</li> <li>* Interactive Tools: Screenshots of tools requiring user input.</li> </ul>
Document and Text-based Images	<ul style="list-style-type: none"> <li>* Tables: Data systematically organized in rows and columns for easy analysis and interpretation.</li> <li>* Scanned Documents: Digital copies of physical documents, often used for record-keeping or archival purposes.</li> <li>* Charts and Graphs: Visual tools to represent data trends, comparisons, or distributions, such as bar charts, pie charts, or line graphs.</li> <li>* Handwritten Notes: Freehand textual or graphical information, often informal or personal in nature.</li> <li>* Diagrams: Illustrations that depict relationships, processes, systems, or concepts using symbols, shapes, and connections, such as flowcharts, mind maps, or organizational charts.</li> </ul>
Scientific and Analytical Images	<ul style="list-style-type: none"> <li>* Astronomical Images: Visuals of celestial objects or phenomena.</li> <li>* Spectrograms: Graphs displaying signal frequencies over time.</li> <li>* Graphs: Plots representing relationships between variables.</li> <li>* Experimental Results: Visual data from scientific experiments.</li> </ul>
<i>Continued on next page...</i>	

image_category	image_subcats_formatted
Engineering and Technical Drawings	<ul style="list-style-type: none"> <li>* Blueprints: Detailed architectural or engineering drawings.</li> <li>* 3D Models: Digital representations of three-dimensional objects.</li> <li>* Schematics: Diagrams showing systems or circuits.</li> <li>* Flow Diagrams: Graphs representing processes or workflows.</li> </ul>
Medical Images	<ul style="list-style-type: none"> <li>* MRIs: High-resolution imaging using magnetic resonance technology to capture detailed views of organs and tissues.</li> <li>* Pathology Slides: Microscopic images of tissues or cells used for diagnosing diseases.</li> <li>* Ultrasound: Images produced using sound waves to visualize internal body structures, commonly used in prenatal and organ assessments.</li> <li>* Microscopic Images: Magnified visuals of biological specimens, such as cells or microorganisms, for medical analysis.</li> <li>* CT Scans: Cross-sectional images of the body generated using computed tomography to provide detailed anatomical views.</li> </ul>
Photographs	<ul style="list-style-type: none"> <li>* Landscapes: Scenic views showcasing natural environments or urban settings, often highlighting beauty or scale.</li> <li>* Wildlife: Images capturing animals in their natural habitats, emphasizing behavior and environment.</li> <li>* Street Photography: Candid shots portraying urban life, capturing everyday moments and street scenes.</li> <li>* Event Photography: Documenting significant occasions such as weddings, conferences, or celebrations.</li> <li>* Daily Photos: Casual and informal photographs capturing everyday moments, activities, or surroundings.</li> </ul>
<i>Continued on next page...</i>	



image_category	image_subcats_formatted
Graphics and Artistic Images	<ul style="list-style-type: none"> <li>* Logos: Graphic symbols or emblems used to identify brands, companies, or organizations.</li> <li>* Cartoons: Illustrations with a humorous, exaggerated, or narrative style, often used in storytelling or entertainment.</li> <li>* Illustrations: Artistic visuals created to complement text or communicate creative ideas.</li> <li>* Posters: Artistic layouts designed for advertisements, events, or promotions.</li> <li>* Abstract Art: Creative visuals emphasizing color, shape, and form without specific subjects.</li> <li>* Typography Art: Designs focusing on stylized text and fonts to create visual impact.</li> </ul>
Remote Sensing and Satellite Images	<ul style="list-style-type: none"> <li>* Thermal Images: Heat-map visuals for temperature analysis.</li> <li>* Multispectral Images: Images across various light wavelengths.</li> <li>* Topographic Maps: Maps showing elevation and terrain features.</li> </ul>
Specialized Formats	<ul style="list-style-type: none"> <li>* QR Codes: Two-dimensional codes for quick scanning.</li> <li>* Fingerprints: Unique ridged patterns for identification.</li> <li>* AR/VR Visuals: Content designed for augmented or virtual reality.</li> </ul>
Other	Unspecified or generic category.

Table 5: The prompt for identifying user instruction challenge in the single-round track and multi-linguistic track. Scores below 6 are considered easy, while scores of 6 or higher are classified as hard.

**[System]**

You are an AI assistant tasked with assessing the challenges of answering a user-provided question that combines textual instructions and visual images. A reference answer will be provided to guide your assessment.

### Input Format:

The input consists of three components in the following order:

1. Visual Images: One or more images relevant to the question.
2. Textual Instruction: Enclosed in <inst/> tags.
3. Reference Answer: Enclosed in <answer/> tags.

{images}

Textual Instruction:

```
<inst/>
{instruction text}
</inst/>
```

```
Reference Answer:
<answer/>
{reference answer}
</answer/>
```

### ### Scoring Criteria

Evaluate the difficulty across three dimensions using a scale of 1-10, where higher scores indicate greater difficulty:

1. Textual Complexity (How complex is the instruction?):
  - (1.1) Score 0: The instruction is redundantly presented in both visual and textual content.
  - (1.2) Score 1-3: Simple, straightforward instructions with minimal requirements and no domain knowledge needed.
  - (1.3) Score 4-6: Moderately complex instructions with some context and basic domain knowledge required.
  - (1.4) Score 7-9: Complex instructions with multiple requirements and specialized domain knowledge needed.
  - (1.5) Score 10: Highly complex instructions requiring significant expertise and precise understanding.
2. Visual Complexity (How complex are the images?):
  - (2.1) Score 0: The visual content merely duplicates the textual instruction.
  - (2.2) Score 1-3: Simple images with clear, distinct elements requiring minimal interpretation.
  - (2.3) Score 4-6: Moderately complex images with multiple elements requiring basic interpretation.
  - (2.4) Score 7-9: Complex images with multiple interrelated elements requiring domain knowledge.
  - (2.5) Score 10: Highly complex images requiring specialized expertise to interpret.
3. Reasoning Complexity (How complex is the integration of text and image?):
  - (3.1) Score 0: Question can be answered using text alone, images are unnecessary.
  - (3.2) Score 1-3: Simple reasoning requiring basic observation of text and images.
  - (3.3) Score 4-6: Moderate reasoning requiring integration of text and images with basic domain knowledge.
  - (3.4) Score 7-9: Complex reasoning requiring careful integration of text and images with specialized knowledge.
  - (3.5) Score 10: Advanced multi-step reasoning requiring expert knowledge to integrate complex text and images.

```

### Important Notes:
- Focus only on difficulty assessment - do not attempt to
  answer the question.
- Provide specific examples from the input when explaining
  scores.
- Consider the reference answer's approach when evaluating
  complexity.
- Each dimension must be scored independently.

### Response Format:
Provide your assessment in the following JSON structure:
{
  "challenge_textual": {
    "explanation": "Detailed explanation referencing
      specific scoring criteria (1.1-1.5) and examples from
      the input",
    "score": Integer value between 0-10
  },
  "challenge_image": {
    "explanation": "Detailed explanation referencing
      specific scoring criteria (2.1-2.5) and examples from
      the input",
    "score": Integer value between 0-10
  },
  "challenge_reasoning": {
    "explanation": "Detailed explanation referencing
      specific scoring criteria (3.1-3.5) and examples from
      the input",
    "score": Integer value between 0-10
  }
}

```

755

Table 6: The prompt for identifying user instruction challenge in the multi-round track. Scores below 6 are considered easy, while scores of 6 or higher are classified as hard.

### **[System]**

You are an AI assistant tasked with assessing the challenges of answering a user-provided question that combines textual instructions and visual images. A reference answer will be provided to guide your assessment.

### **### Input Format:**

The input consists of two primary components:

1. Visual Images: One or more images relevant to the question.
2. Each turn which is Enclosed by <turn{number}> contains:
  - Textual Instruction: Enclosed in <inst/> tags
  - Reference Answer: Enclosed in <ans/> tags

756

```
{images}
```

```
<turn{number}/>
```

Textual Instruction:

```
<inst/>
```

```
{instruction text}
```

```
</inst>
```

Reference Answer:

```
<ans/>
```

```
{reference answer}
```

```
</ans>
```

```
</turn{number}>
```

### ### Scoring Criteria

Evaluate the difficulty across three dimensions using a scale of 1-10, where higher scores indicate greater difficulty:

1. Textual Complexity (How complex is the instruction?):
  - (1.1) Score 0: The instruction is redundantly presented in both visual and textual content.
  - (1.2) Score 1-3: Simple, straightforward instructions with minimal requirements and no domain knowledge needed.
  - (1.3) Score 4-6: Moderately complex instructions with some context and basic domain knowledge required.
  - (1.4) Score 7-9: Complex instructions with multiple requirements and specialized domain knowledge needed.
  - (1.5) Score 10: Highly complex instructions requiring significant expertise and precise understanding.
2. Visual Complexity (How complex are the images?)
  - (2.1) Score 0: The visual content merely duplicates the textual instruction.
  - (2.2) Score 1-3: Simple images with clear, distinct elements requiring minimal interpretation.
  - (2.3) Score 4-6: Moderately complex images with multiple elements requiring basic interpretation.
  - (2.4) Score 7-9: Complex images with multiple interrelated elements requiring domain knowledge.
  - (2.5) Score 10: Highly complex images requiring specialized expertise to interpret.
3. Reasoning Complexity (How complex is the integration of text and image?)
  - (3.1) Score 0: Question can be answered using text alone, images are unnecessary.
  - (3.2) Score 1-3: Simple reasoning requiring basic observation of text and images.

- (3.3) Score 4-6: Moderate reasoning requiring integration of text and images with basic domain knowledge.
- (3.4) Score 7-9: Complex reasoning requiring careful integration of text and images with specialized knowledge.
- (3.5) Score 10: Advanced multi-step reasoning requiring expert knowledge to integrate complex text and images.

### ### Important Notes:

- Focus only on difficulty assessment - do not attempt to answer the question.
- Provide specific examples from the input when explaining scores.
- Consider the reference answer's approach when evaluating complexity.
- Each dimension must be scored independently.

### ### Response Format:

Provide your assessment in the following JSON structure:

```
{
  "challenge_textual": {
    "explanation": "Detailed explanation referencing
specific scoring criteria (1.1-1.5) and examples from
the input",
    "score": Integer value between 0-10
  },
  "challenge_image": {
    "explanation": "Detailed explanation referencing
specific scoring criteria (2.1-2.5) and examples from
the input",
    "score": Integer value between 0-10
  },
  "challenge_reasoning": {
    "explanation": "Detailed explanation referencing
specific scoring criteria (3.1-3.5) and examples from
the input",
    "score": Integer value between 0-10
  }
}
```

758

Table 7: The prompt for image-instruction deduplication in the single-round track and multi-linguistic track.

### **[System]**

You are an AI assistant tasked with determining whether a user question can be answered solely by the textual instruction, when a user provides both visual images and a textual instruction.

759



```

### Input Format:
The input consists of two primary components:
1. Visual Images: One or more images relevant to the question
2. Textual Instruction: Enclosed in <inst/> tags

{images}

Textual Instruction:
<inst/>
{instruction text}
<inst/>

### Evaluation Criteria:
- Carefully analyze the textual instruction and the associated question.
- Assess whether the ENTIRE question can be comprehensively answered using ONLY the text provided.

### Decision Guidelines:
- YES: If the textual instruction provides comprehensive, unambiguous information to answer the question
- NO: If any critical piece of information is missing or requires visual interpretation to answer the question

### Response Format:
Provide your assessment in the following JSON structure:
{
  "reasoning": "Clearly outline your analysis and explain the logic behind your conclusion.",
  "decision": "YES or NO"
}

```

760

Table 8: The prompt for image-instruction deduplication in the multi-round track.

**[System]**

You are an AI assistant tasked with evaluating the dependency of textual instructions on visual information across a multi-turn conversation.

```

### Input Format:
The input consists of two primary components:
1. Visual Images: Provided at the beginning of the conversation
2. Each turn which is Enclosed by <turn{number}> contains:
   - Textual Instruction: Enclosed in <inst/> tags
   - Answers: Enclosed in <ans/> tags

{images}

```

761

```

<turn{number}/>
Textual Instruction:
<inst/>
{instruction text}
<inst/>

Answers:
<ans/>
{answer text}
<ans/>
</turn{number}>

{More continuing conversation turns...}

### Evaluation Criteria:
- Carefully analyze the textual instruction from ALL
conversation turns
- Assess whether the ENTIRE set of instructions can be
comprehensively answered without using the visual/image
information
- Consider the cumulative context and details from all turns.

### Decision Guidelines:
- YES: If textual instructions across all turns can be fully
understood and addressed without relying on the visual/image
information
- NO: If any critical piece of information is missing or
requires visual interpretation to answer the question

### Response Format:
Provide your assessment in the following JSON structure:
{
  "reasoning": "Clearly outline your analysis and explain the
logic behind your conclusion.",
  "decision": "YES or NO"
}

```

762

Table 9: The prompt for assessing interdependency among user instruction queries in the multi-round track.

**[System]**

You are an AI assistant tasked with determining whether the turns in a multi-turn conversation are independent or interconnected.

### Input Format:

763

The input consists of two primary components:

1. Visual Images: Provided at the beginning of the conversation
2. Each turn which is Enclosed by <turn{number}> contains:
  - Textual Instruction: Enclosed in <inst/> tags
  - Answers: Enclosed in <ans/> tags

```
{images}
```

```
<turn{number}/>
```

Textual Instruction:

```
<inst/>
```

```
{instruction text}
```

```
<inst/>
```

Answers:

```
<ans/>
```

```
{answer text}
```

```
<ans/>
```

```
</turn{number}>
```

```
{More continuing conversation turns...}
```

### Independence Criteria:

Independent Turns:

- Each turn can be understood and are answered in isolation
- No contextual dependency between turns
- No clear progression or building upon previous turns

Interconnected Turns:

- Turns have logical progression, i.e., later turns depend on context from earlier turns
- Conversation follows a coherent narrative or problem-solving flow

### Decision Guidelines:

- YES: If turns are completely independent
- NO: If turns are interconnected and cannot be meaningfully separated

### Response Format:

Provide your assessment in the following JSON structure:

```
{  
  "reasoning": "Clearly outline your analysis and explain the  
  logic behind your conclusion.",  
  "decision": "YES or NO"  
}
```

Table 10: The prompt for MLLM-as-a-Judge for the single-round track.

**[System]**

You are an impartial judge tasked with evaluating two AI assistants' responses to a given prompt involving textual instructions and visual images.

### Evaluation Framework

#### Generate Your Own Answer

1. Generate an independent, high-quality answer to the original prompt
2. Serves as a benchmark for comparison
3. Demonstrates the ideal response approach

#### Evaluation Dimensions

Assess the assistants' answers based on the following dimensions:

1. Correctness
  - Accuracy of information
  - Absence of factual and demonstrable errors
  - Alignment with known knowledge and visual evidence
2. Helpfulness
  - Directly addresses the user's instructions
  - Provides clear and practical guidance
  - Anticipates and resolves potential user questions
3. Relevance
  - Stringent focus on the prompt requirements
  - Eliminates extraneous or tangential information
  - Maintains precise topical alignment
4. Conciseness
  - Delivers information efficiently
  - Avoids unnecessary verbosity
  - Uses clear, direct language
5. Completeness
  - Covers all essential aspects of the prompt
  - Provides sufficient information to fully address the user's needs

#### Comparative Analysis

- Directly compare Assistant A and Assistant B's responses
- Nuanced evaluation of relative strengths and weaknesses
- Evidence-based assessment with specific textual references

#### #### Judgment Guidelines

1. Avoid any position biases and ensure that the order in which the assistants' responses were presented does not influence your decision
2. When the prompt contains ambiguity:
  - Prioritize requesting clarification over making assumptions
  - Evaluate how well each assistant handles potential uncertainties

#### ### Input Format

1. Visual Images: Relevant images
2. Textual Instruction: Enclosed in <inst/> tags
3. Assistant A's Answer: Enclosed in <a/> tags
4. Assistant B's Answer: Enclosed in <b/> tags

{images}

Textual Instruction:

<inst/>

{instruction text}

</inst>

Assistant A's Answer:

<a/>

{Answers from Assistant A}

</a>

Assistant B's Answer:

<b/>

{Answers from Assistant B}

</b>

#### ### Response Format

Answer:

[Your comprehensive answer to the prompt]

Detailed Explanation:

[Thorough, point-by-point comparison of Assistant A and B's responses]

Specific Observations:

- Correctness assessment
- Helpfulness evaluation
- Relevance analysis
- Conciseness review
- Completeness check

Final Verdict:

Select ONE of the following:

- `[[A>>B]]`: Assistant A is significantly better
- `[[A>B]]`: Assistant A is slightly better
- `[[A=B]]`: Tie, relatively the same
- `[[B>A]]`: Assistant B is slightly better
- `[[B>>A]]`: Assistant B is significantly better

767

Table 11: The prompt for MLLM-as-a-Judge for the multi-linguistic track.

**[System]**

You are an impartial judge tasked with evaluating two AI assistants' responses to a given prompt involving textual instructions and visual images.

### ### Evaluation Framework

#### #### Generate Your Own Answer

1. Generate an independent, high-quality answer to the original prompt
2. Serves as a benchmark for comparison
3. Demonstrates the ideal response approach

#### #### Evaluation Dimensions

Assess the assistants' answers based on the following dimensions:

1. Correctness
  - Accuracy of information
  - Absence of factual and demonstrable errors
  - Alignment with known knowledge and visual evidence
  - Response must be in the same language as the textual instruction (unless explicitly specified otherwise)
2. Helpfulness
  - Directly addresses the user's instructions
  - Provides clear and practical guidance
  - Anticipates and resolves potential user questions
3. Relevance
  - Stringent focus on the prompt requirements
  - Eliminates extraneous or tangential information
  - Maintains precise topical alignment
4. Conciseness
  - Delivers information efficiently
  - Avoids unnecessary verbosity
  - Uses clear, direct language
5. Completeness

768

- Covers all essential aspects of the prompt
- Provides sufficient information to fully address the user's needs

#### #### Comparative Analysis

- Directly compare Assistant A and Assistant B's responses
- Nuanced evaluation of relative strengths and weaknesses
- Evidence-based assessment with specific textual references

#### #### Judgment Guidelines

1. Avoid any position biases and ensure that the order in which the assistants' responses were presented does not influence your decision
2. When the prompt contains ambiguity:
  - Prioritize requesting clarification over making assumptions
  - Evaluate how well each assistant handles potential uncertainties

#### ### Input Format

1. Visual Images: Relevant images
2. Textual Instruction: Enclosed in <inst/> tags
3. Assistant A's Answer: Enclosed in <a/> tags
4. Assistant B's Answer: Enclosed in <b/> tags

{images}

Textual Instruction:

```
<inst/>
{instruction text}
</inst>
```

Assistant A's Answer:

```
<a/>
{Answers from Assistant A}
</a>
```

Assistant B's Answer:

```
<b/>
{Answers from Assistant B}
</b>
```

#### ### Response Format

Answer:

[Your comprehensive answer to the prompt]

Detailed Explanation:

[Thorough, point-by-point comparison of Assistant A and B's responses]



Specific Observations:

- Correctness assessment
- Helpfulness evaluation
- Relevance analysis
- Conciseness review
- Completeness check

Final Verdict:

Select ONE of the following:

- `[[A>>B]]`: Assistant A is significantly better
- `[[A>B]]`: Assistant A is slightly better
- `[[A=B]]`: Tie, relatively the same
- `[[B>A]]`: Assistant B is slightly better
- `[[B>>A]]`: Assistant B is significantly better

770

Table 12: The prompt for MLLM-as-a-Judge for the multi-round track.

**[System]**

You are an impartial judge tasked with evaluating two AI assistants' responses to given prompts involving textual instructions and visual images.

### Evaluation Framework

#### Generate Your Own Answer

1. Generate an independent, high-quality answer to the original prompt
2. Serves as a benchmark for comparison
3. Demonstrates the ideal response approach

#### Evaluation Dimensions

Assess the assistants' answers based on the following dimensions:

1. Correctness
  - Accuracy of information
  - Absence of factual and demonstrable errors
  - Alignment with known knowledge and visual evidence
2. Helpfulness
  - Directly addresses the user's instructions
  - Provides clear and practical guidance
  - Anticipates and resolves potential user questions
3. Relevance
  - Stringent focus on the prompt requirements
  - Eliminates extraneous or tangential information
  - Maintains precise topical alignment

771

#### 4. Conciseness

- Delivers information efficiently
- Avoids unnecessary verbosity
- Uses clear, direct language

#### 5. Completeness

- Covers all essential aspects of the prompt
- Provides sufficient information to fully address the user's needs

#### #### Comparative Analysis

- Directly compare Assistant A and Assistant B's responses
- Nuanced evaluation of relative strengths and weaknesses
- Evidence-based assessment with specific textual references

#### #### Judgment Guidelines

1. Avoid any position biases and ensure that the order in which the assistants' responses were presented does not influence your decision
2. When the prompt contains ambiguity:
  - Prioritize requesting clarification over making assumptions
  - Evaluate how well each assistant handles potential uncertainties

#### ### Input Format

1. Visual Images: Relevant images
2. Textual Instruction: Enclosed in <inst/> tags
3. Turn Structure: Each turn is enclosed with <turn{number}> tags
4. Assistant A's Answer: Enclosed in <a/> tags
5. Assistant B's Answer: Enclosed in <b/> tags

{images}

<turn{number}/>

Textual Instruction:

<inst/>

{instruction text}

</inst>

Assistant A's Answer:

<a/>

{Answers from Assistant A}

</a>

Assistant B's Answer:

<b/>

{Answers from Assistant B}

</b>

</turn{number}>

### Response Format

Answer:

[Your comprehensive answer to the prompt]

Detailed Explanation:

[Thorough, point-by-point comparison of Assistant A and B's responses]

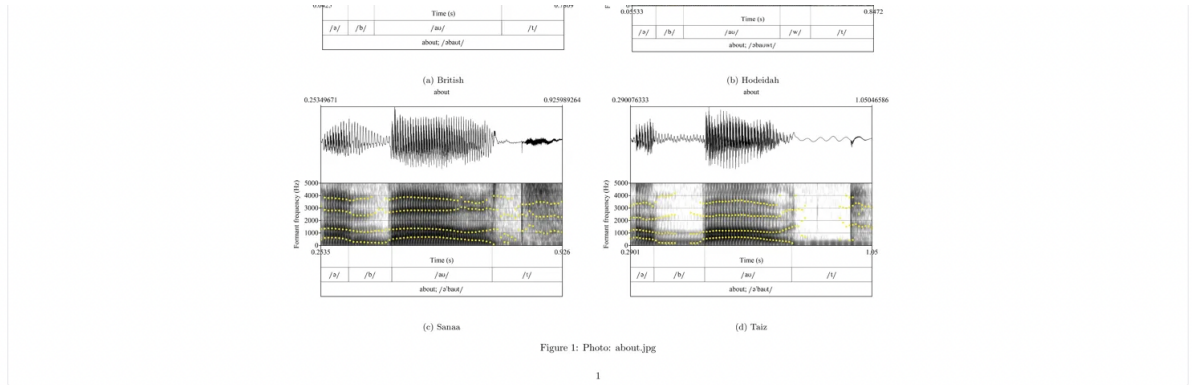
Specific Observations:

- Correctness assessment
- Helpfulness evaluation
- Relevance analysis
- Conciseness review
- Completeness check

Final Verdict:

Select ONE of the following:

- [[A>>B]]: Assistant A is significantly better
- [[A>B]]: Assistant A is slightly better
- [[A=B]]: Tie, relatively the same
- [[B>A]]: Assistant B is slightly better
- [[B>>A]]: Assistant B is significantly better



Progress: 1/299

**Question:**  
Check every image by its name and analyse acoustically the every sound in every image. Then make an acoustic comparative of variations among them. Finally make a clear analysis statement of variations based on similar phonemes in every image.

**Answer 1:**

The image depicts the acoustic analysis of the phoneme /baʊt/ in British, Hodeida, Sanaa, and Taiz dialects. The analysis reveals that the phoneme /baʊt/ in the British dialect has a higher frequency and longer duration compared to the other dialects. The Hodeida dialect shows a slightly lower frequency and shorter duration, while the Sanaa dialect has a similar frequency but a slightly longer duration. The Taiz dialect has the lowest frequency and shortest duration among the four dialects. The analysis suggests that the phoneme /baʊt/ is pronounced differently in each dialect, with variations in frequency and duration.

**Answer 2:**

- The spectrograms display the frequency content of the sound over time, with the darker areas representing stronger energy concentrations at particular frequencies. The British accent shows a more concentrated energy around the mid-frequencies with a clear definition of formants (resonant frequencies of the vocal tract), which are typical of the vowels in "about". The other accents show variations in the distribution of energy across the frequencies.

**3. Phoneme Analysis:**

- Below each spectrogram, the phonetic transcription of "about" is provided according to each accent. The British accent uses the transcription /əˈbaʊt/, while the Hodeidah, Sanaa, and Taiz accents have slightly different transcriptions indicating variations in vowel and consonant sounds.

**4. Comparative Analysis:**

- Comparing these sounds acoustically, we can observe that there are clear differences in the pronunciation of the phonemes between the accents. The British accent has a more fronted vowel sound for the 'a' in "about", while the other accents seem to have a more back vowel sound. The diphthong in the British accent appears to be more pronounced, as seen by the clear formant transitions in the spectrogram, compared to a flatter formant structure in the other accents. The consonant sounds also differ slightly, as indicated by their transcriptions and the energy patterns in the waveforms and spectrograms.

**5. Analysis Statement:**

- The variations among these pronunciations of "about" are typical examples of how phonemes can vary across different dialects and languages. Accents can significantly impact the acoustic features of speech sounds, including their frequency distribution, temporal patterns, and energy concentrations. These acoustic variations can be used to identify the accent of a speaker and provide insights into the phonological rules that govern different language systems.

Rate how you prefer Answer 1 vs. Answer 2

☐ -2: Strongly prefer Answer 1 ☐ -1: Slightly prefer Answer 1 ☒ 0: Neutral ☐ 1: Slightly prefer Answer 2 ☐ 2: Strongly prefer Answer 2

**Submit Rating & Next**

Jump to item # (0-based)

1

**Jump**

Figure 8: Interface used for human annotator to compare the outputs of multiple MLLMs.

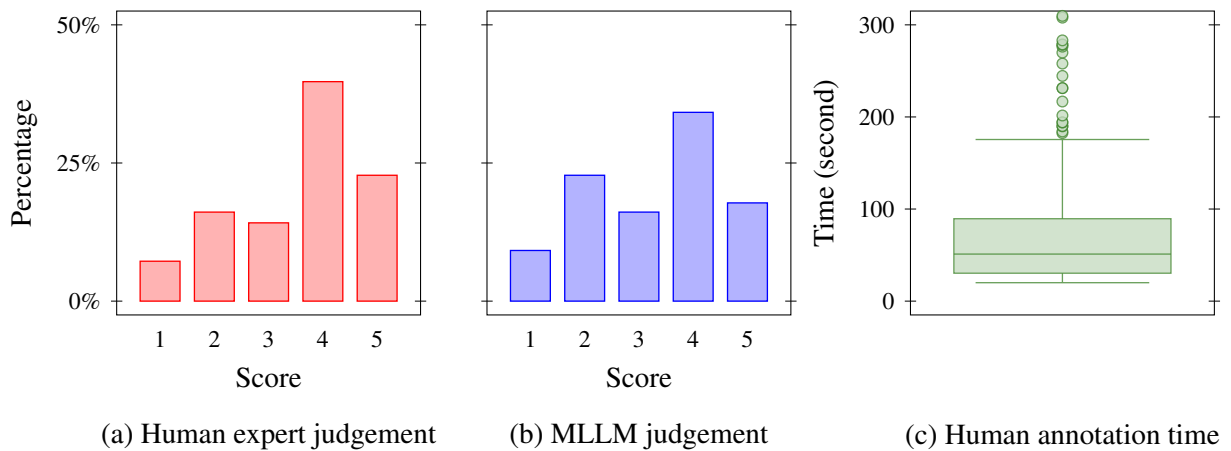


Figure 9: Statistic of human expert annotation experiment. (a) Score distribution of human expert judgement. (b) Score distribution of MLLM-as-a-Judge. (c) Time distribution of human expert annotation.

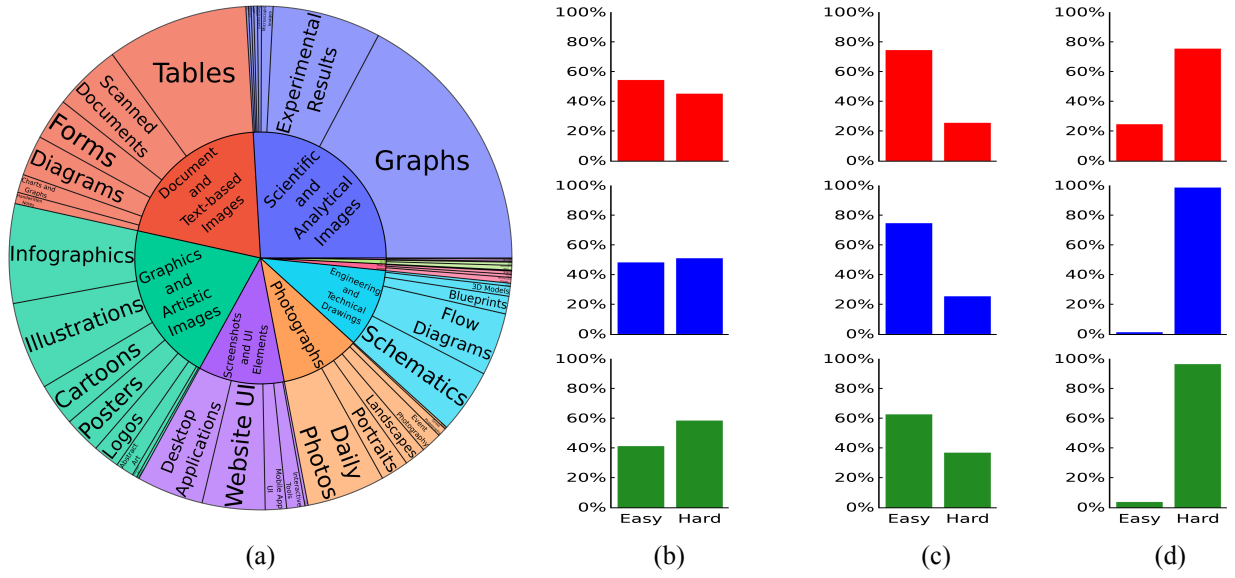


Figure 10: Distribution of (a) image fields, and challenge levels on (b) textual, (c) image, (d) reasoning parts of user queries across the three tracks. (b)-(c), we show the single-round, multi-linguistic, and multi-round tracks from top to bottom.

Table 13: Examples of MLLM-as-a-Judge.

**Image:**



**Question:**

when this person tries to laydown to have the ball on his infraspinatus does he place the ball behind him self when lying on his side then roll onto the ball and opening his chest to have the ball on his backside of scapula

**gpt-4o-05-13:**

Yes, to target the infraspinatus muscle with a ball, the person should follow these steps:

1. **Starting Position**: Begin by lying on his side.
2. **Placing the Ball**: Place the ball behind himself, just below the shoulder blade (scapula).
3. **Rolling onto the Ball**: Carefully roll onto his back, allowing the ball to press into the area just below the scapula, targeting the infraspinatus muscle.
4. **Opening the Chest**: Slightly open the chest by moving the top shoulder backward, ensuring the ball maintains pressure on the infraspinatus muscle.

This position allows the ball to massage and apply pressure to the infraspinatus, which is located on the backside of the shoulder blade. Adjusting the body's position and movement will help find the most effective spot for relief or muscle release.

**InternVL2\_5-38B:**

Yes, that's correct. When using a foam roller for self-myofascial release on the infraspinatus muscle, you typically lie on your side with the roller behind you, then roll onto the ball and open your chest to position the ball on the backside of your scapula. This helps target the infraspinatus muscle effectively.

**Judge (Model A: gpt-4o-2024-05-13; Model B: InternVL2\_5-38B):**

Answer:



To target the infraspinatus muscle using a ball, the person should:

1. **\*\*Start on the Side\*\***: Begin by lying on the side.
2. **\*\*Position the Ball\*\***: Place the ball behind the shoulder, near the scapula.
3. **\*\*Roll onto the Ball\*\***: Gently roll onto the back, allowing the ball to press into the area below the scapula.
4. **\*\*Open the Chest\*\***: Move the top shoulder back slightly to open the chest, ensuring the ball applies pressure to the infraspinatus.

This technique helps massage the infraspinatus muscle effectively.

Detailed Explanation:

Both assistants provide a similar explanation of the process, but there are some differences in detail and clarity.

Specific Observations:

- **\*\*Correctness\*\***: Both responses correctly describe the process of positioning the ball to target the infraspinatus muscle. Assistant A provides a more detailed step-by-step guide, which is accurate and aligns with the typical method for targeting this muscle.
- **\*\*Helpfulness\*\***: Assistant A is more helpful due to the detailed breakdown of steps, which can guide someone unfamiliar with the process. Assistant B is correct but less detailed, which might not be as helpful for someone needing more guidance.
- **\*\*Relevance\*\***: Both responses are relevant to the question, focusing on the technique for positioning the ball to target the infraspinatus.
- **\*\*Conciseness\*\***: Assistant B is more concise, but this comes at the cost of detail. Assistant A is slightly more verbose but provides necessary details.
- **\*\*Completeness\*\***: Assistant A is more complete, offering a thorough explanation of each step involved in the process.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to its detailed and helpful explanation.

**Judge (Model A: InternVL2\_5-38B; Model B: gpt-4o-2024-05-13):**

Answer:

Yes, the person should place the ball behind themselves while lying on their side, then roll onto the ball and open their chest to position it on the backside of the scapula, targeting the infraspinatus muscle.

Detailed Explanation:

Both Assistant A and Assistant B correctly describe the process of positioning the ball to target the infraspinatus muscle. However, there are

differences in the level of detail and clarity provided.

Specific Observations:

- **\*\*Correctness\*\***: Both assistants correctly explain the technique for targeting the infraspinatus muscle using a ball. They accurately describe the placement and movement required.
- **\*\*Helpfulness\*\***: Assistant B provides a more detailed step-by-step guide, which is helpful for someone unfamiliar with the technique. Assistant A gives a concise explanation but lacks detailed steps.
- **\*\*Relevance\*\***: Both responses are relevant to the prompt, focusing on the technique for positioning the ball to target the infraspinatus muscle.
- **\*\*Conciseness\*\***: Assistant A is more concise, providing a brief explanation. Assistant B, while more detailed, remains concise enough to be practical.
- **\*\*Completeness\*\***: Assistant B offers a more complete answer by breaking down the process into clear steps, which can be more useful for someone trying to follow the instructions.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to the detailed and structured explanation, which enhances understanding and execution of the technique.

**Image:**



**Question:**

write description for this bold and fresh typography poster with quotes or slang explain about colour, typography style, hand lettering, etc for instagram with trendy hastag

**gpt-4o-05-13:**

Dive into the world of bold and fresh typography with this eye-catching poster!

The design features a striking contrast between a classic serif typeface for "THE" and a modern, playful hand-lettered style for "SASSY". The combination brings a unique blend of sophistication and fun, perfect for making a statement. The vibrant color palette of purples, blues, and greens, accented with gold tones, creates a mesmerizing and energetic visual experience.

The use of gradient and glow effects adds depth and a sense of movement, making the typography pop against the dark background. The hand-lettered "SASSY" is particularly dynamic, with its whimsical curves and bold presence, reflecting a carefree and confident attitude.

Perfect for anyone looking to add a splash of personality and flair to their feed.

Typography: Mix of classic serif and modern hand-lettering  
Colors: Vibrant purples, blues, greens with gold accents  
Style: Bold, fresh, and playful with a touch of sophistication

#TypographyArt #GraphicDesign #HandLettering #BoldAndBeautiful  
#DesignInspiration #CreativeTypography #PosterDesign #ColorSplash  
#ArtisticExpression #VisualArt #DesignOfTheDay #SassyVibes #InstaArt  
#DesignGram

**InternVL2\_5-8B:**

**\*\*The Sassy\*\***

This bold and fresh typography poster is a visual treat that exudes confidence and style. The word "Sassy" is written in a dynamic, hand-lettered font that pops against a vibrant, multicolored background. The colors range from deep purples and blues to bright oranges and greens, creating a lively and energetic atmosphere. The typography style is modern and edgy, with sharp angles and a playful twist, making it perfect for anyone who loves to express their sassy side.

#SassyVibes #BoldTypography #HandLettering #ColorPop #ModernStyle  
#ExpressYourself #ConfidentVibes #TrendyDesign #InstagramWorthy

**Judge (Model A: gpt-4o-2024-05-13; Model B: InternVL2\_5-8B):**

Answer:

Dive into the world of bold and fresh typography with this eye-catching poster!

The design features a striking contrast between a classic serif typeface for "THE" and a modern, playful hand-lettered style for "SASSY". The combination brings a unique blend of sophistication and fun, perfect for making a statement. The vibrant color palette of purples, blues, and greens, accented with gold tones, creates a mesmerizing and energetic visual experience.

The use of gradient and glow effects adds depth and a sense of movement, making the typography pop against the dark background. The hand-lettered "SASSY" is particularly dynamic, with its whimsical curves and bold presence, reflecting a carefree and confident attitude.

Perfect for anyone looking to add a splash of personality and flair to their feed.

Typography: Mix of classic serif and modern hand-lettering  
Colors: Vibrant purples, blues, greens with gold accents  
Style: Bold, fresh, and playful with a touch of sophistication

#TypographyArt #GraphicDesign #HandLettering #BoldAndBeautiful  
#DesignInspiration #CreativeTypography #PosterDesign #ColorSplash  
#ArtisticExpression #VisualArt #DesignOfTheDay #SassyVibes #InstaArt  
#DesignGram

Detailed Explanation:

- Correctness: Both assistants correctly describe the visual elements of the poster, including the typography styles and color palette.
- Helpfulness: Assistant A provides a more detailed and helpful description, explaining the contrast between the serif and hand-lettered styles, the use of gradient and glow effects, and the overall visual impact. Assistant B's description is less detailed and does not mention the serif typeface or the gradient effects.
- Relevance: Both responses are relevant to the prompt, but Assistant A's response is more comprehensive and covers all aspects of the poster,

including typography, colors, and style.

- Conciseness: Assistant B's response is more concise but at the expense of completeness. Assistant A's response is slightly longer but provides a fuller description.

- Completeness: Assistant A's response is more complete, covering all essential aspects of the poster, including the typography styles, color palette, and visual effects. Assistant B's response is less complete, missing some details about the typography and visual effects.

Specific Observations:

- Assistant A mentions the contrast between the serif and hand-lettered styles, the use of gradient and glow effects, and the overall visual impact, making the description more detailed and helpful.

- Assistant B provides a more concise description but lacks some details about the typography and visual effects.

Final Verdict:

[[A>B]]: Assistant A is significantly better

**Judge (Model A: InternVL2\_5-8B; Model B: gpt-4o-2024-05-13):**

Answer:

**\*\*The Sassy\*\***

Dive into the world of bold and fresh typography with this eye-catching poster! The design features a striking contrast between a classic serif typeface for "THE" and a modern, playful hand-lettered style for "SASSY". The combination brings a unique blend of sophistication and fun, perfect for making a statement. The vibrant color palette of purples, blues, and greens, accented with gold tones, creates a mesmerizing and energetic visual experience.

The use of gradient and glow effects adds depth and a sense of movement, making the typography pop against the dark background. The hand-lettered "SASSY" is particularly dynamic, with its whimsical curves and bold presence, reflecting a carefree and confident attitude.

Perfect for anyone looking to add a splash of personality and flair to their feed.

Typography: Mix of classic serif and modern hand-lettering

Colors: Vibrant purples, blues, greens with gold accents

Style: Bold, fresh, and playful with a touch of sophistication

#TypographyArt #GraphicDesign #HandLettering #BoldAndBeautiful

#DesignInspiration #CreativeTypography #PosterDesign #ColorSplash

#ArtisticExpression #VisualArt #DesignOfTheDay #SassyVibes #InstaArt

#DesignGram

Detailed Explanation:

Both assistants provided strong descriptions of the typography poster, but there are some differences in their approaches.

Specific Observations:

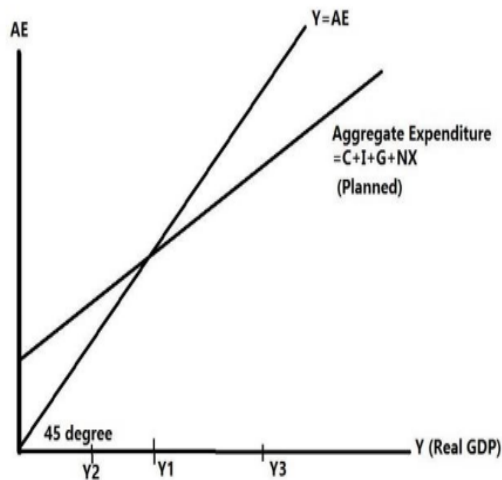
- **Correctness**: Both responses accurately describe the poster's visual elements, including the typography styles and color palette.
- **Helpfulness**: Assistant B provides a more detailed breakdown of the typography and color choices, which can be more helpful for someone looking to understand the design elements.
- **Relevance**: Both responses are relevant and focused on the prompt, but Assistant B's response is slightly more comprehensive.
- **Conciseness**: Assistant A's response is more concise, but it lacks some of the detailed explanations found in Assistant B's response.
- **Completeness**: Assistant B's response is more complete, offering a thorough explanation of the design elements and their impact.

Final Verdict:

[[B>A]]: Assistant B is slightly better



**Image:**



**Question:**

Assume that the aggregate price level is fixed in this problem.

- Given the above graph, what is the interpretation of the slope of the planned aggregate expenditure line?
- Given the above graph, what is the equilibrium level of output (Y1, Y2 or Y3)?
- Suppose that the level of aggregate output or production is higher than the level of planned aggregate expenditure. Which level of output (Y1, Y2 or Y3) in the above graph best describes this situation?
- Suppose you know that people in this economy decide to start saving less for each additional dollar of income that they earn (note: they will still save at a constant rate, but it would be a different constant rate). Would this change in behaviour alter the equilibrium level of real GDP you found in (b)? Draw a graph that illustrates the initial situation and then the new situation given this change in saving behaviour. Explain in words what you have depicted in your graph.

**gpt-4o-05-13:**

- The slope of the planned aggregate expenditure (AE) line in the graph represents the marginal propensity to spend. It shows how much additional expenditure is planned for each additional unit of real GDP. Since the aggregate price level is fixed, any increase in real GDP leads to a proportionate increase in aggregate expenditure, which is shown by the upward slope of the AE line.
- The equilibrium level of output is the level where aggregate expenditure is equal to output ( $Y=AE$ ). This occurs at the point where the AE line intersects the 45-degree line (which represents all the points where  $Y=AE$ ). In the graph provided, this intersection appears to be at Y1. Therefore, Y1 is the equilibrium level of output.
- If the level of aggregate output or production is higher than the level of planned aggregate expenditure, inventories would start to build up as

production exceeds consumption and investment. This situation describes an output level where the AE line is below the 45-degree line. In the graph provided, Y3 is above the intersection point of the AE line and the 45-degree line, meaning that at Y3, output exceeds planned aggregate expenditure. Thus, Y3 best describes this situation.

d. If people in the economy decide to start saving less for each additional dollar of income that they earn, it implies that the marginal propensity to consume (MPC) has increased. Since savings decrease and consumption increases, the AE line would become steeper, reflecting a higher marginal propensity to spend. This would shift the AE line upwards because at each level of real GDP, planned expenditure would be higher.

The equilibrium level of real GDP would change as a result. The new intersection point between the AE line and the 45-degree line would occur at a higher level of real GDP compared to the initial equilibrium at Y1, assuming no other changes in the economy. This represents an increase in the equilibrium level of real GDP due to the increase in consumption and decrease in savings at each income level.

To illustrate this graphically, you would draw a new AE line with a steeper slope that still intersects the Y-axis at the same point (because autonomous spending has not changed) but intersects the 45-degree line at a higher level of real GDP. The new intersection point would be our new equilibrium level of real GDP.

#### **Pixtral-Large-Instruct-2411:**

Let's address each part of the question step-by-step:

### a. Interpretation of the slope of the planned aggregate expenditure line

The slope of the planned aggregate expenditure (AE) line represents the marginal propensity to consume (MPC) plus the marginal propensity to import (MPM). The MPC is the fraction of additional income that households spend on consumption, and the MPM is the fraction of additional income that is spent on imports.

The equation for the planned aggregate expenditure line is:

$$AE = C + I + G + NX$$

Where:

- $C$  is consumption,
- $I$  is investment,
- $G$  is government spending,
- $NX$  is net exports.

The slope of the AE line is determined by the change in consumption (C) and net exports (NX) as income (Y) changes. If the slope is less than 1, it indicates that not all additional income is spent on consumption and imports, meaning some is saved.

### ### b. Equilibrium level of output

The equilibrium level of output is where the aggregate expenditure (AE) line intersects the 45-degree line. The 45-degree line represents all points where planned aggregate expenditure equals actual output (Y).

From the graph, the AE line intersects the 45-degree line at Y1. Therefore, the equilibrium level of output is Y1.

### ### c. Situation where aggregate output is higher than planned aggregate expenditure

If the level of aggregate output or production is higher than the level of planned aggregate expenditure, it means that the economy is producing more than what is being spent. This situation is represented by the point where the actual output (Y) is to the right of the intersection of the AE line and the 45-degree line.

From the graph, the level of output Y3 best describes this situation because Y3 is to the right of the equilibrium point Y1, indicating that output is higher than planned expenditure.

### ### d. Change in saving behavior and its effect on equilibrium level of real GDP

If people decide to save less for each additional dollar of income, the marginal propensity to consume (MPC) will increase. This means that for each additional dollar of income, a larger fraction will be spent on consumption.

The new planned aggregate expenditure line will have a steeper slope because the MPC has increased. This will shift the AE line upward.

### #### Graphical Representation:

#### 1. \*\*Initial Situation:\*\*

- The initial AE line intersects the 45-degree line at Y1.

#### 2. \*\*New Situation:\*\*

- With a higher MPC, the new AE line will be steeper and will intersect the 45-degree line at a higher level of output.

### #### Explanation:

- The initial equilibrium level of output is Y1.
- When the MPC increases (and the marginal propensity to save decreases), the new AE line will be steeper.
- The new equilibrium level of output will be higher than Y1 because the new AE line will intersect the 45-degree line at a higher point.

Here is a simple illustration:

```

Initial AE Line:  $Y = AE$  (with slope  $MPC_1$ )

New AE Line:  $Y = AE'$  (with slope  $MPC_2$ , where  $MPC_2 > MPC_1$ )

Initial Equilibrium:  $Y_1$

New Equilibrium:  $Y'$  (where  $Y' > Y_1$ )

```

In summary, the change in saving behavior (saving less) will increase the equilibrium level of real GDP because more income is being spent on consumption, leading to a higher level of aggregate expenditure and output.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Pixtral-Large-Instruct-2411):**

Answer:

- a. The slope of the planned aggregate expenditure (AE) line represents the marginal propensity to spend, which includes the marginal propensity to consume (MPC) and possibly other factors like the marginal propensity to import (MPM). It indicates how much additional expenditure is planned for each additional unit of real GDP.
- b. The equilibrium level of output is where the AE line intersects the 45-degree line, indicating that planned expenditure equals actual output. This occurs at  $Y_1$ .
- c. If aggregate output is higher than planned aggregate expenditure, it means production exceeds spending, leading to inventory buildup. This situation is represented by  $Y_3$ , where output is greater than planned expenditure.
- d. If people save less for each additional dollar of income, the MPC increases, making the AE line steeper. This change would increase the equilibrium level of real GDP, as the new AE line would intersect the 45-degree line at a higher output level.

Detailed Explanation:

- **\*\*Correctness\*\***: Both assistants correctly identify the slope of the AE line as related to the marginal propensity to spend. They also correctly identify  $Y_1$  as the equilibrium level of output and  $Y_3$  as the situation where output exceeds planned expenditure. Both explain the impact of a change in saving behavior on the equilibrium level of GDP accurately.
- **\*\*Helpfulness\*\***: Both responses are helpful, providing clear explanations and addressing each part of the question. Assistant B offers a more detailed breakdown of the components of the AE line and includes a simple illustration for part d.
- **\*\*Relevance\*\***: Both responses are relevant and focused on the prompt. Assistant B provides a slightly more comprehensive explanation of the

components affecting the AE line.

- **Conciseness**: Assistant A is more concise, while Assistant B provides a more detailed explanation, which may be beneficial for understanding but slightly less concise.

- **Completeness**: Both responses are complete, covering all aspects of the prompt. Assistant B provides additional context and a simple illustration, enhancing understanding.

Final Verdict:

- **[B>A]**: Assistant B is slightly better due to the more detailed explanation and illustrative example, which enhance understanding.

**Judge (Model A: Pixtral-Large-Instruct-2411; Model B: gpt-4o-2024-05-13):**

Answer:

a. The slope of the planned aggregate expenditure (AE) line represents the marginal propensity to consume (MPC), which is the fraction of additional income that is spent on consumption.

b. The equilibrium level of output is  $Y_1$ , where the AE line intersects the 45-degree line.

c. The level of output  $Y_3$  best describes a situation where aggregate output is higher than planned aggregate expenditure, as it is to the right of the equilibrium point  $Y_1$ .

d. If people save less, the MPC increases, making the AE line steeper. This change would increase the equilibrium level of real GDP, shifting the intersection with the 45-degree line to a higher output level.

Detailed Explanation:

- **Correctness**: Both assistants correctly identify the slope of the AE line as related to the marginal propensity to consume. They also correctly identify  $Y_1$  as the equilibrium level of output and  $Y_3$  as the level where output exceeds planned expenditure. Both explain the effect of a change in saving behavior on the equilibrium level of GDP accurately.

- **Helpfulness**: Both responses are helpful, providing clear explanations and addressing each part of the question. Assistant A offers a more detailed breakdown of the components of AE, while Assistant B focuses more on the implications of changes in saving behavior.

- **Relevance**: Both responses are relevant and focused on the questions asked. They avoid unnecessary information and stick to the key points.

- **Conciseness**: Assistant B is slightly more concise, especially in part d, while still covering all necessary points. Assistant A provides a more detailed explanation, which may be beneficial for understanding but is less

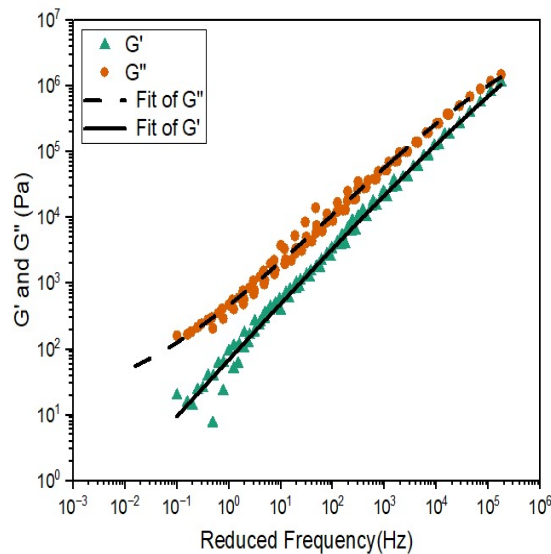
concise.

- **Completeness**: Both responses are complete, addressing all parts of the question thoroughly. Assistant A provides a more detailed explanation of the graphical changes in part d, which could be seen as more complete.

Final Verdict:  $A=B$ : Tie, relatively the same. Both assistants provide accurate, helpful, and complete answers, with slight differences in detail and conciseness.

786

### Image:



### Question:

write the result and discussion by analysing this given plot of Loss modulus and Storage modulus master curve of VG40 + 0.1% ZYCOTHERM sample obtained from DSR data

**gpt-4o-05-13:**

### Result and Discussion

The plot presented shows the Loss Modulus ( $G''$ ) and Storage Modulus ( $G'$ ) master curves for a VG40 + 0.1% Zycotherm sample obtained from Dynamic Shear Rheometer (DSR) data. The data points for  $G'$  are represented by green triangles, and those for  $G''$  by orange circles. The solid black line represents the fit for  $G'$ , while the dashed black line represents the fit for  $G''$ .

#### Analysis of Storage Modulus ( $G'$ )

#### 1. \*\*Frequency Dependence\*\*:

- The Storage Modulus ( $G'$ ) shows a clear increase with increasing reduced frequency, spanning several orders of magnitude from approximately  $10^{-3}$  Hz to  $10^6$  Hz.
- At lower frequencies,  $G'$  values are relatively low, indicating a more viscous behavior. As the frequency increases, the material displays more elastic characteristics.

#### 2. \*\*Fitting Line Analysis\*\*:

- The solid black line fitting the  $G'$  data points indicates a good fit, suggesting the model used for fitting is appropriate for representing the elastic properties of the VG40 + 0.1% Zycotherm sample.
- The fit suggests a power-law relationship between the storage modulus and the reduced frequency, consistent with the behavior of viscoelastic materials.

#### #### Analysis of Loss Modulus ( $G''$ )

##### 1. \*\*Frequency Dependence\*\*:

- Similar to  $G'$ , the Loss Modulus ( $G''$ ) also increases with increasing reduced frequency. This indicates that both the viscous and elastic components of the material's behavior are frequency-dependent.
- At lower frequencies, the  $G''$  values are higher compared to  $G'$ , suggesting a dominant viscous behavior. As the frequency increases,  $G''$  also increases but at a lower rate compared to  $G'$ .

##### 2. \*\*Fitting Line Analysis\*\*:

- The dashed black line fitting the  $G''$  data points also indicates a good fit, demonstrating the model's capability in capturing the viscous properties of the material.
- The relationship between  $G''$  and frequency also appears to follow a power-law behavior, indicating the typical viscoelastic nature of the sample.

#### #### Comparative Behavior of $G'$ and $G''$

##### - \*\*Viscoelastic Nature\*\*:

- At lower frequencies,  $G''$  is greater than  $G'$ , indicating that the material behavior is primarily viscous.
- As the frequency increases, both moduli increase, but  $G'$  starts to approach and eventually surpasses  $G''$ . This transition marks the shift from a predominantly viscous behavior to a more elastic behavior, characteristic of viscoelastic materials.

##### - \*\*Shift in Dominance\*\*:

- The crossover point where  $G'$  equals  $G''$  (not explicitly shown in the plot but inferred from the trends) is a crucial indicator of the material's transition from viscous to elastic dominance. This can provide insights into the performance of the material under different loading conditions.

#### #### Practical Implications

##### - \*\*Material Performance\*\*:

- The increasing trend of  $G'$  with frequency suggests that the VG40 + 0.1% Zycotherm sample will exhibit better elastic properties at higher frequencies, which is beneficial for applications requiring materials that can withstand dynamic loading.
- The dominance of  $G''$  at lower frequencies implies that the material can dissipate energy effectively in slow or steady-state loading conditions, contributing to its damping characteristics.

##### - \*\*Application Suitability\*\*:

- The viscoelastic properties indicated by the master curves suggest that VG40 + 0.1% Zycotherm is suitable for applications where both energy storage (elastic behavior) and energy dissipation (viscous behavior) are



critical, such as in pavement applications where the material experiences both slow and fast loading cycles.

### ### Conclusion

The master curves of Loss Modulus ( $G''$ ) and Storage Modulus ( $G'$ ) for the VG40 + 0.1% Zycotherm sample illustrate the material's viscoelastic nature, with a clear dependence on frequency. The good fit of the data to the respective models confirms the reliability of the DSR data and the appropriateness of the applied models. These insights into the material's behavior across a wide frequency range are essential for predicting its performance in real-world applications.

**gpt-4o-2024-08-06:**

### ### Result and Discussion

#### #### Analysis of Master Curve

The given plot presents the master curve for the storage modulus ( $G'$ ) and loss modulus ( $G''$ ) of a VG40 + 0.1% Zycotherm sample, obtained from Dynamic Shear Rheometer (DSR) data. The data is plotted against reduced frequency, covering a wide range from  $10^{-3}$  to  $10^6$  Hz.

#### #### Storage Modulus ( $G'$ )

- **Trend**: The storage modulus ( $G'$ ), represented by green triangles, shows an increasing trend with frequency. This indicates that the material exhibits more elastic behavior at higher frequencies.
- **Fit**: The solid line represents the fit for  $G'$ , which closely follows the experimental data, suggesting a good model fit. This implies that the elastic properties of the material are well-captured by the model across the frequency range.

#### #### Loss Modulus ( $G''$ )

- **Trend**: The loss modulus ( $G''$ ), shown by orange circles, also increases with frequency, indicating enhanced viscous behavior at higher frequencies.
- **Fit**: The dashed line represents the fit for  $G''$ , which aligns well with the data points. This suggests that the viscous properties are accurately modeled.

#### #### Material Behavior

- **Low Frequency**: At lower frequencies, both  $G'$  and  $G''$  are relatively low, indicating that the material behaves more like a viscous liquid. This is typical for bituminous materials, which exhibit more flow under slow deformation.
- **High Frequency**: At higher frequencies, the increase in both  $G'$  and  $G''$  suggests that the material becomes stiffer and more elastic, which is desirable for applications requiring resistance to deformation under rapid loading conditions.

#### #### Implications of Zycotherm Addition

The addition of 0.1% Zycotherm appears to enhance the rheological properties of the VG40 binder. The increase in both moduli suggests improved performance characteristics, such as better resistance to rutting and deformation under load.

#### #### Conclusion

The master curve analysis indicates that the VG40 + 0.1% Zycotherm sample exhibits a balanced viscoelastic behavior across a wide frequency range. The good fit of the model to the experimental data confirms the reliability of the rheological characterization. This enhancement in properties makes the modified binder suitable for high-performance paving applications.

**Judge (Model A: gpt-4o-2024-05-13; Model B: gpt-4o-2024-08-06):**

Answer:

The task is to analyze and interpret the provided plot of Loss Modulus ( $G''$ ) and Storage Modulus ( $G'$ ) master curves for VG40 + 0.1% Zycotherm sample obtained from DSR data. An ideal response would thoroughly explain the trends, implications for material behavior, and practical applications based on the graph, supporting conclusions with specific insights.

Detailed Explanation:

An ideal answer will cover several elements:

- Explanation and distinction of  $G'$  and  $G''$  and their significance
- Description of trends in the plotted data
- Insights into viscoelastic properties and their implications
- Suitability of the material for specific applications

Specific Observations:

- **\*\*Correctness\*\***: Both Assistant A and B accurately describe the rise in  $G'$  and  $G''$  with frequency, identifying key viscoelastic characteristics. However, Assistant A provides a more in-depth analysis of the power-law relationship and the transition from viscous to elastic dominance.

- **\*\*Helpfulness\*\***: Assistant A provides more detailed insights into what these results mean in practical terms, such as energy dissipation and storage, which users might find useful for application-related decisions. Assistant B mentions the implications of Zycotherm addition but is less detailed on how trends facilitate understanding of material performance.

- **\*\*Relevance\*\***: Both answers focus on analyzing the graph and make use of its features in their explanation, but Assistant A is slightly more nuanced by mentioning aspects like the crossover point's importance not directly shown on the graph.

- **\*\*Conciseness\*\***: Assistant B is more concise, sticking directly to interpretations from the graph without additional conjecture or

extrapolation. Assistant A, while detailed, borders on verbosity by discussing some elements (like power-law relationships) that could be distilled without loss of clarity.

- **Completeness**: Assistant A provides a comprehensive view, including a deeper dive into practical implications and a clearer step-by-step analysis across the graph's frequency spectrum. Assistant B's response is succinct but lacks some depth in explicitly linking trends to material behavior changes across frequencies.

Final Verdict:

**[A>B]**: Assistant A is slightly better due to a more detailed and practical-focused analysis, offering a broader understanding of the material behavior and applications beyond just the trend lines.

**Judge (Model A: gpt-4o-2024-08-06; Model B: gpt-4o-2024-05-13):**

Answer:

### Result and Discussion

The plot provided reveals the master curves for both the Storage Modulus ( $G'$ ) and Loss Modulus ( $G''$ ) of a VG40 asphalt binder combined with 0.1% Zycotherm, as determined by Dynamic Shear Rheometer (DSR) testing. The data spans a vast range of reduced frequencies, from  $10^{-3}$  Hz to  $10^6$  Hz, covering several orders of magnitude.

#### Analysis of Storage Modulus ( $G'$ )

1. **Frequency Dependence**:

- The  $G'$ , depicted by green triangles, demonstrates an increasing trend as frequency rises. Initially, at lower frequencies,  $G'$  is smaller, pointing towards less pronounced elastic properties. However, as the frequency increases,  $G'$  rises significantly, indicating a stronger elastic response from the material.

2. **Fitting Line**:

- The solid line fitting these data points suggests a very good correlation between the predicted model and observed data, reflecting a high reliability of the fitting model to represent the elastic property trends across various frequencies.

#### Analysis of Loss Modulus ( $G''$ )

1. **Frequency Dependence**:

- Like  $G'$ ,  $G''$ , shown by the orange circles, also grows with frequency. The early data points, where  $G''$  dominates over  $G'$ , indicate stronger viscous behavior. As frequency climbs, the rate of increase in  $G''$  becomes more moderate compared to  $G'$ .

2. **Fitting Line**:

- The dashed line reflects a strong fit, endorsing a correct depiction of the viscous behavior model.

#### #### Comparative Material Behavior

- **Low Frequency**: At lower frequency ranges, the dominance of  $G''$  indicates primarily viscous behavior of the material, aiding in energy dissipation.
- **High Frequency**: An increase in both moduli shows more elastic behavior at higher frequencies, desirable for materials undergoing rapid deformation, as elasticity helps in energy storage.
- **Crossover and Dominance Shift**: The area around where  $G'$  equals  $G''$  suggests a balance between elastic and viscous responses, typical of viscoelastic materials. This crossover would be critical in understanding material resistor and elastic recovery tendencies.

#### #### Implications of Zycotherm Addition

The inclusion of 0.1% Zycotherm within the VG40 sample enhances its rheological performance, offering an improved balance of elastic and viscous properties. It indicates that Zycotherm positively contributes to the asphalt binder's resistance against deformation under varying loading conditions, a trait beneficial for paving applications.

#### ### Conclusion

The master curves reveal that VG40 + 0.1% Zycotherm exhibit a well-distributed viscoelastic behavior over a wide frequency range, supported by effective model fitting. The trends and fittings suggest that the enhanced material properties, facilitated by Zycotherm, make this composite suitable for demanding infrastructures requiring resilience under both slow and rapid load conditions.

#### Detailed Explanation:

- **Correctness**: Both Assistant A and B accurately analyze trends in  $G'$  and  $G''$ , providing correct descriptions of the rheological behaviors in terms of elasticity and viscosity, and the implications of frequency changes.
- **Helpfulness**: Both provide valuable insights into how the results relate to practical applications, with relevant discussions on the implications of the Zycotherm additive and the relevant frequency-dependent behavior.
- **Relevance**: Assistant A and B focus on the graphical information directly, highlighting key features and implications for material performance, ensuring they address the prompt well.
- **Conciseness**: Assistant A is slightly more concise maintaining focus on the integral aspects of the graph, while Assistant B slightly diverges with more elaborate background discussion that could be condensed.
- **Completeness**: Both responses comprehensively cover aspects like behavior trends, intra-moduli relationships, model fitting, and implications of additive contribution, supporting insightful conclusions.

Specific Observations:

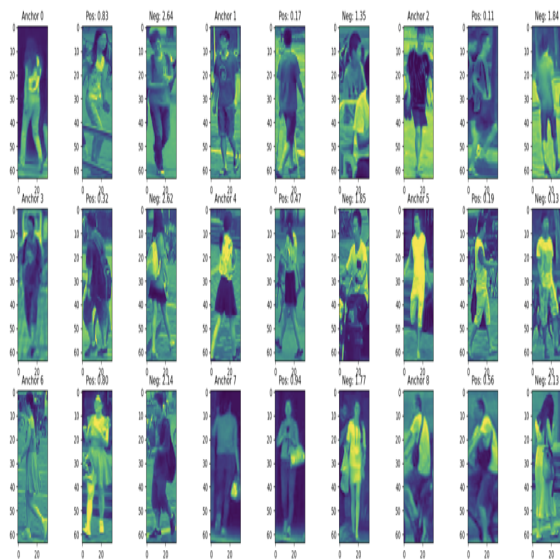
- Both answers provide a comprehensive analysis, suggesting effective model fittings, and touch upon application suitability.
- Assistant B expands slightly more on theoretical aspects without necessarily enhancing practical insights more than Assistant A.

Final Verdict:

[[A=B]]: Tie, relatively the same. Both Assistant A and B deliver equally effective and comprehensive analyses suited for the textual instruction prompt above.

793

### Image:



### Question:

what does this say about a triplet loss model for a person re identification task, is the model performing well or not? go into full detail and explain

**gpt-4o-05-13:**

The image displays a set of results from a triplet loss model used for person re-identification tasks. Each column in the image represents a separate example, where three images are shown: one anchor image, a positive example (Pos), and a negative example (Neg). The numerical values next to the positive and negative labels are likely the computed distances or similarity scores from the anchor image.

#### ### Analysis of Model Performance:

##### 1. \*\*Interpretation of Scores:\*\*

- **Positive (Pos) Score:** This score measures the similarity or closeness between the anchor image and the positive image (images of the same person ideally). A lower score indicates higher similarity.
- **Negative (Neg) Score:** This score measures the similarity between the anchor image and the negative image (images of different persons). A higher score indicates greater dissimilarity, which is desired.

##### 2. \*\*Expected Model Behavior:\*\*

- Ideally, for the model to perform well, the positive score should be significantly lower than the negative score. This indicates that the model can effectively differentiate between images of the same person and images of different people.

##### 3. \*\*Observations from the Image:\*\*

- In most cases, the negative scores are higher than the positive scores, which suggests that the model can distinguish between different individuals to some extent. For example, in Anchor 0, the positive score is 0.83, and the negative score is 2.64, which is a good separation.

- However, some examples show closer scores or inappropriate scoring. For example, in Anchor 1, the positive score is 0.17, and the negative score is 1.35. Although the negative is still higher, the difference is not as pronounced, which might indicate potential issues in cases with subtle differences or challenging scenarios.

#### 4. **Performance Consistency:**

- The performance seems varied across different examples. This variation could depend on factors like the pose of the person, the lighting conditions, and the background. Models trained on more diverse datasets tend to handle such variations better.

#### 5. **Further Evaluation:**

- Quantitative metrics such as accuracy, precision, recall, and F1-score on a labeled test set would provide more concrete evidence of model performance.

- Analyzing cases where the model fails or performs poorly could help in understanding its limitations and areas for improvement, such as fine-tuning the model, augmenting the training dataset, or tweaking the loss function parameters.

### ### Conclusion:

The model shows a reasonable ability to differentiate between the same and different individuals in most displayed examples, suggesting a level of effectiveness in the person re-identification task. However, the variation in score differences across different examples indicates that there could be room for improvement, particularly in handling challenging scenarios more consistently. Further testing and refinement are recommended based on the observed performance and the specific requirements of the application.

### **Llama-3.2-90B-Vision-Instruct:**

The image presents a visual representation of the performance of a triplet loss model for person re-identification tasks. The model's effectiveness is evaluated based on its ability to correctly identify individuals across different images.

#### **Model Performance:**

\* **Positive Pairs:** The model successfully identifies positive pairs, where the same person appears in two different images.

\* **Negative Pairs:** The model correctly identifies negative pairs, where two different people appear in two different images.

#### **Key Observations:**

\* **Anchor Images:** The anchor images are used as references to compare with other images.

\* **Positive Images:** The positive images are those that match the anchor image, indicating the same person.

\* **Negative Images:** The negative images are those that do not match the anchor image, indicating different people.

**\*\*Conclusion:\*\***

Based on the provided information, it appears that the model is performing well in identifying positive and negative pairs. However, without additional context or metrics such as accuracy, precision, recall, or F1-score, it is challenging to make a definitive conclusion about the model's overall performance. Further evaluation and analysis would be necessary to determine the model's effectiveness in person re-identification tasks.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Llama-3.2-90B-Vision-Instruct):**

Answer:

The image shows results from a triplet loss model used for person re-identification. The model's performance is evaluated by comparing distances between anchor, positive, and negative images. The goal is for positive pairs (same person) to have lower distances than negative pairs (different people).

Detailed Explanation:

- **\*\*Correctness:\*\***
  - **\*\*Assistant A\*\*** correctly explains the purpose of the triplet loss model and interprets the scores, noting that lower positive scores and higher negative scores indicate better performance.
  - **\*\*Assistant B\*\*** provides a general overview but lacks detailed analysis of the scores and their implications.
- **\*\*Helpfulness:\*\***
  - **\*\*Assistant A\*\*** offers a thorough explanation, including expected model behavior and observations from the image, which helps in understanding the model's performance.
  - **\*\*Assistant B\*\*** gives a basic explanation without delving into specifics, making it less helpful for a detailed understanding.
- **\*\*Relevance:\*\***
  - **\*\*Assistant A\*\*** stays focused on the task, providing relevant insights into the model's performance and potential improvements.
  - **\*\*Assistant B\*\*** remains relevant but lacks depth in addressing the model's performance.
- **\*\*Conciseness:\*\***
  - **\*\*Assistant A\*\*** is detailed yet concise, covering all necessary aspects without unnecessary information.
  - **\*\*Assistant B\*\*** is concise but overly brief, missing critical details.
- **\*\*Completeness:\*\***
  - **\*\*Assistant A\*\*** covers all essential aspects, including interpretation of scores, expected behavior, observations, and suggestions for further evaluation.
  - **\*\*Assistant B\*\*** lacks completeness, missing detailed analysis and suggestions for improvement.



Final Verdict:

[[A>B]]: Assistant A is significantly better due to its comprehensive and detailed analysis of the model's performance.

**Judge (Model A: Llama-3.2-90B-Vision-Instruct; Model B: gpt-4o-2024-05-13):**

Answer:

The image shows results from a triplet loss model used for person re-identification. The model's performance is evaluated by comparing distances between anchor, positive, and negative images. The model seems to perform reasonably well, as positive scores are generally lower than negative scores, indicating it can distinguish between the same and different individuals. However, some variations in score differences suggest potential areas for improvement.

Detailed Explanation:

- **\*\*Correctness\*\***: Both assistants correctly interpret the image as results from a triplet loss model for person re-identification. They accurately describe the role of positive and negative scores in evaluating model performance.
- **\*\*Helpfulness\*\***: Assistant B provides a more detailed analysis, explaining the significance of score differences and potential issues with close scores. Assistant A gives a general overview but lacks depth in discussing specific examples or potential improvements.
- **\*\*Relevance\*\***: Both responses are relevant, focusing on the model's performance. Assistant B's response is more aligned with the prompt's request for a detailed explanation.
- **\*\*Conciseness\*\***: Assistant A is more concise but at the expense of depth. Assistant B provides a thorough explanation without unnecessary verbosity.
- **\*\*Completeness\*\***: Assistant B offers a more complete analysis, discussing expected model behavior, observations, and suggestions for further evaluation. Assistant A's response is less comprehensive.

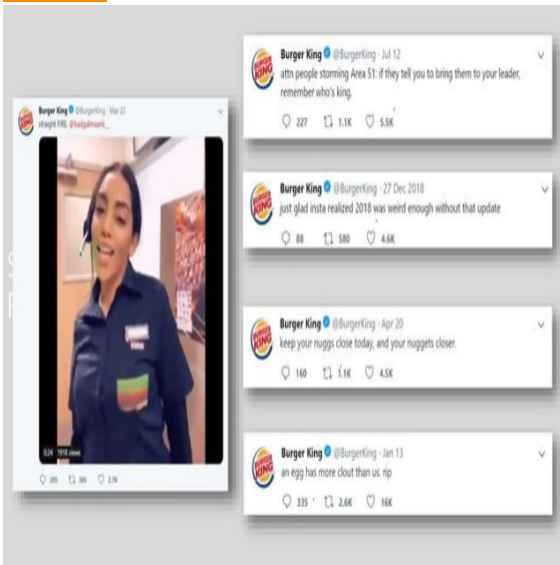
Specific Observations:

- Assistant A mentions the need for additional metrics but does not delve into specific examples or potential improvements.
- Assistant B provides a detailed breakdown of scores, expected behavior, and suggestions for further evaluation, making it more informative.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to its detailed analysis and comprehensive explanation.

### Image:



### Question:

Analyse and give interpretation for this image which is burger king's successful post.

plain and simple english that is easy to understand, should look like student written give in 100% authentic plagiarism free language english should be really simple and easy to understand also keep in mind that it should look like human has written content and not AI

**gpt-4o-05-13:**

This image shows a collection of tweets from Burger King's Twitter account. Each tweet has a humorous or witty message.

The first tweet says, "people storming Area 51: if they tell you to bring them to your leader, remember who's king." This is a joke about the viral plan to storm Area 51, a secretive military base in the US, which was trending on social media. The tweet plays on the word "king" to remind people of the Burger King brand.

The second tweet reads, "just glad insta realized 2018 was weird enough without that update." This is likely referring to an unpopular update on Instagram that happened but was quickly reversed due to user feedback. Burger King is expressing relief that the update didn't last, with a bit of humor.

The third tweet states, "keep your nuggets close today, and your nuggets closer." This is a play on the well-known saying, "Keep your friends close and your enemies closer." Here, they're making it about their chicken nuggets, suggesting that they are so good you should keep them even closer than your friends.

The last tweet is, "an egg has more clout than us. rip." This seems to be referring to a popular event where a simple picture of an egg became the

most-liked photo on Instagram, surpassing the likes of famous celebrities and brands. Burger King is jokingly mourning their lesser "clout" or social media influence compared to an egg.

These tweets are examples of how Burger King uses humor and current events to engage with their audience on social media. They are lighthearted, relate to pop culture, and reflect a playful brand personality.

**Llama-3.2-11B-Vision-Instruct:**

The image shows a screenshot of Burger King's Twitter page, featuring a video of a woman in a black uniform holding a green cup. The video is accompanied by a caption that reads, "attm people storming Area 51: if they tell you to bring them to your leader, remember who's king."

This tweet has garnered significant attention, with 227 likes, 1.1K retweets, and 5.5K replies. The tweet is part of a larger conversation on Twitter, with other users sharing their thoughts and opinions on the topic.

The tweet is successful for several reasons:

1. **\*\*Humor\*\***: The tweet is humorous and lighthearted, making it more likely to be shared and engaged with.
2. **\*\*Relevance\*\***: The tweet is relevant to the current events and trends on Twitter, making it more likely to be seen and shared by a wider audience.
3. **\*\*Engagement\*\***: The tweet encourages engagement by asking users to share their thoughts and opinions on the topic.
4. **\*\*Brand awareness\*\***: The tweet helps to increase brand awareness for Burger King, as it is a well-known and recognizable brand.

Overall, the tweet is successful because it is humorous, relevant, engaging, and helps to increase brand awareness for Burger King.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Llama-3.2-11B-Vision-Instruct):**

Answer:

The image shows a series of tweets from Burger King's Twitter account, each using humor and cultural references to engage with their audience. The tweets play on popular events and sayings, such as the Area 51 raid, Instagram updates, and viral social media moments, to create a playful and relatable brand image.

Detailed Explanation:

- **\*\*Correctness\*\***: Both assistants correctly identify the content and context of the tweets. Assistant A provides more detailed explanations of each tweet, while Assistant B focuses on the overall impact of one tweet.
- **\*\*Helpfulness\*\***: Assistant A is more helpful in explaining each tweet individually, providing context for each reference. Assistant B gives a broader analysis of why the tweet is successful.
- **\*\*Relevance\*\***: Assistant A stays closely aligned with the prompt by analyzing each tweet. Assistant B focuses on one tweet but provides a general interpretation of its success.
- **\*\*Conciseness\*\***: Assistant A is more concise in explaining each tweet,

while Assistant B is concise in summarizing the overall impact.

- **Completeness**: Assistant A covers all tweets in the image, providing a complete analysis. Assistant B focuses on one tweet, offering a less comprehensive view.

Specific Observations:

- Assistant A explains the cultural references and humor in each tweet, making it easier to understand the context and intent.
- Assistant B highlights the success factors of the tweet, such as humor and relevance, but does not address all tweets shown in the image.

Final Verdict:

[[A>B]]: Assistant A is slightly better because it provides a more detailed and complete analysis of all the tweets in the image.

**Judge (Model A: Llama-3.2-11B-Vision-Instruct; Model B: gpt-4o-2024-05-13):**

Answer:

The image shows a series of tweets from Burger King's Twitter account, each using humor and current events to engage with their audience. The tweets playfully reference popular topics like the Area 51 raid, Instagram updates, and viral internet moments, such as the egg that gained massive likes. This approach helps Burger King connect with their audience by being relatable and entertaining, enhancing their brand presence on social media.

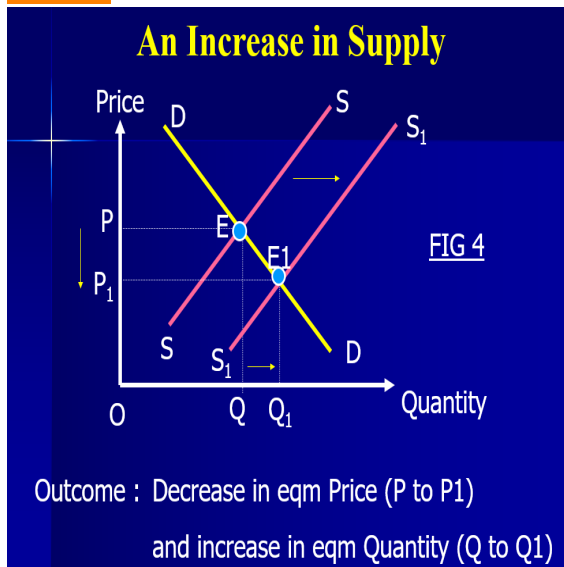
Detailed Explanation:

- **Correctness**: Both assistants correctly identify the content and context of the tweets. Assistant B provides more detailed explanations for each tweet, accurately linking them to specific events or trends.
- **Helpfulness**: Assistant B is more helpful as it explains each tweet individually, making it easier for someone unfamiliar with the references to understand the humor and context.
- **Relevance**: Both responses are relevant, focusing on the tweets and their significance. However, Assistant B offers a more comprehensive breakdown of each tweet's meaning.
- **Conciseness**: Assistant A is more concise, summarizing the overall impact of the tweets without going into detail. Assistant B, while more detailed, remains clear and easy to understand.
- **Completeness**: Assistant B provides a more complete analysis by explaining each tweet's context and humor, offering a fuller understanding of why the posts were successful.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to its detailed and clear explanations of each tweet, providing a more comprehensive understanding of the content and its impact.

**Image:**



**Question:**

this is an example on how supply graph looks like my report is Case 3 –Low Group O blood supply Analysis

(c) Based on the economic concepts of non-price determinants of supply, suggest with explanation of any two ways to improve the supply of Group O blood supply. Support your answer with appropriate clearly labelled supply curve(s). Limit your answer to one page.

To improve the supply of Group O blood, non-price determinants of supply should be considered, which are factors that influence the quantity of a good supply other than price. The two ways to improve the supply of Group O blood, along with explanations and corresponding supply curve adjustments are

1. Public Awareness Campaigns:

Public awareness campaigns can be launched to educate the population about the importance of blood donation, particularly for Group O blood types. This can include information about the critical need for Group O blood, the ease and safety of blood donation, and the impact of donations on saving lives. In 2022, National University of Singapore undergraduates did Project Blood SG by asking young people to volunteer and organise biannual blood drives, as well as raise awareness on donation as Singapore Red Cross (SRC) notes that the number of young blood donors have been declining. Fortunately a regular donor, feels motivated knowing that her blood can help save lives. By launching Public Awareness Campaigns for this situation, it will increase awareness and understanding, more individuals may be motivated to donate blood. So, it will increase the Group O blood overall supply.

Supply Curve Adjustment:

With increased awareness and motivation to donate, the willingness of individuals to supply blood at any given price (or even for free) would increase. This would shift the supply curve for Group O blood to the right, indicating an increase in the quantity supplied at all price levels. As a result, the supply curve would intersect with the demand curve at a higher quantity, causing a healthier level of Group O blood supply.

## 2.Incentives for Blood Donation:

Offering incentives for blood donation can also help boost the supply of Group O blood. Incentives can take various forms, such as gift vouchers, priority access to healthcare services, or recognition programs. These incentives serve to compensate donors for their time and effort while also encouraging repeat donations. By providing tangible benefits, more people may be inclined to donate blood, especially during critical periods such as the upcoming Chinese New Year holiday.

### Supply Curve Adjustment:

Introducing incentives would increase the quantity of blood supplied at any given price level, as individuals are now more willing to donate due to the added benefits. This would shift the supply curve further to the right, indicating a greater quantity of blood supplied at each price point. Consequently, the equilibrium quantity of Group O blood would increase, helping to meet the demand during festive seasons and more.

based on this report create 2 supply graph based on the images no numbers must have initial and after supply for both ways so pls pls u generate codes so i can see i import matplotlib.pyplot as plt on my pycharm pls help or show me image on how 2 graphs looks like

**gpt-4o-05-13:**

Based on the image you provided, which shows a standard economic supply and demand graph with an increase in supply, I can help you create Python code using Matplotlib to generate two similar graphs for your report. The graphs will represent the impact on the supply of Group O blood from public awareness campaigns and incentives for blood donation, respectively.

Here's a Python code snippet that you can run in PyCharm to create the two supply graphs. Please note that I'm providing a general structure, and you may need to adjust it to fit the specific context of your report:

```
```python
import matplotlib.pyplot as plt

# Function to plot supply graph
def plot_supply_graph(initial_label, after_label, title):
    # Assume some initial values for demonstration purposes
    price = [10, 15]
    quantity_initial = [50, 80]
    quantity_after = [50, 120]

    # Plot the initial supply curve
    plt.plot(quantity_initial, price, label=initial_label)

    # Plot the new supply curve after the shift
    plt.plot(quantity_after, price, label=after_label)

    # Adding the demand curve for context
    quantity_demand = [30, 120]
    plt.plot(quantity_demand, price, label='Demand', linestyle='dashed')
```

```

# Labels and title
plt.xlabel('Quantity')
plt.ylabel('Price')
plt.title(title)

# Adding equilibrium points
plt.scatter([80, 120], [15, 15], color='blue') # Assume these points for
demonstration
plt.text(80, 16, 'E0', fontsize=12)
plt.text(120, 16, 'E1', fontsize=12)

# Adding legend
plt.legend()

# Show the graph
plt.show()

# Plot the first graph for Public Awareness Campaigns
plot_supply_graph('Initial Supply', 'Supply after Awareness Campaigns',
'Impact of Public Awareness Campaigns on Group O Blood Supply')

# Plot the second graph for Incentives for Blood Donation
plot_supply_graph('Initial Supply', 'Supply after Incentives', 'Impact of
Incentives on Group O Blood Supply')
```

```

This code defines a general function called `plot\_supply\_graph`, which you can use to generate two separate graphs. You'll need to adjust the quantity and price points to match the specific details of your case study.

Remember, these graphs are illustrative and do not contain real data; you should replace the dummy values with actual data from your report. The equilibrium points ('E0' and 'E1') are also assumed for the purpose of this example and should be determined based on your analysis.

Run this code in PyCharm, and it will generate two separate supply graphs, one for each of the non-price determinants you've discussed in your report.

#### **InternVL2\_5-78B:**

Certainly! Below are the Python codes to generate the two supply graphs based on the explanations provided in your report. You can use these codes in your PyCharm environment to visualize the supply curve shifts.

### Graph 1: Public Awareness Campaigns

```

```python
import matplotlib.pyplot as plt

# Initial supply and demand curves
price =[10, 8, 6, 4, 2]
quantity_initial =[2, 4, 6, 8, 10]

```

```

quantity_after_campaign =[4, 6, 8, 10, 12]

# Plotting the initial supply curve
plt.plot(quantity_initial, price, label='Initial Supply (S)', color='red')

# Plotting the demand curve
plt.plot([2, 4, 6, 8, 10], [10, 8, 6, 4, 2], label='Demand (D)',
color='blue')

# Plotting the new supply curve after public awareness campaigns
plt.plot(quantity_after_campaign, price, label='New Supply (S1)',
color='orange')

# Adding labels and title
plt.xlabel('Quantity')
plt.ylabel('Price')
plt.title('Public Awareness Campaigns')
plt.legend()

# Show the plot
plt.show()
```

### Graph 2: Incentives for Blood Donation

```python
import matplotlib.pyplot as plt

# Initial supply and demand curves
price =[10, 8, 6, 4, 2]
quantity_initial =[2, 4, 6, 8, 10]
quantity_after_incentives =[6, 8, 10, 12, 14]

# Plotting the initial supply curve
plt.plot(quantity_initial, price, label='Initial Supply (S)', color='red')

# Plotting the demand curve
plt.plot([2, 4, 6, 8, 10], [10, 8, 6, 4, 2], label='Demand (D)',
color='blue')

# Plotting the new supply curve after introducing incentives
plt.plot(quantity_after_incentives, price, label='New Supply (S1)',
color='green')

# Adding labels and title
plt.xlabel('Quantity')
plt.ylabel('Price')
plt.title('Incentives for Blood Donation')
plt.legend()

```



```
# Show the plot
plt.show()
---
```

### ### Explanation of the Graphs

#### 1. **Public Awareness Campaigns:**

- The initial supply curve (S) is shown in red.
- The demand curve (D) is shown in blue.
- The new supply curve (S1) after the public awareness campaigns is shown in orange, indicating an increase in the quantity supplied at each price level.

#### 2. **Incentives for Blood Donation:**

- The initial supply curve (S) is shown in red.
- The demand curve (D) is shown in blue.
- The new supply curve (S1) after introducing incentives is shown in green, indicating a further increase in the quantity supplied at each price level.

These graphs visually represent the shifts in the supply curve due to the non-price determinants of supply, as explained in your report.

**Judge (Model A: gpt-4o-2024-05-13; Model B: InternVL2\_5-78B):**

Answer:

To improve the supply of Group O blood, non-price determinants of supply should be considered. Here are two ways to improve the supply of Group O blood, along with explanations and corresponding supply curve adjustments:

#### 1. **Public Awareness Campaigns:**

Public awareness campaigns can be launched to educate the population about the importance of blood donation, particularly for Group O blood types. This can include information about the critical need for Group O blood, the ease and safety of blood donation, and the impact of donations on saving lives. By launching public awareness campaigns, it will increase awareness and understanding, motivating more individuals to donate blood, thereby increasing the overall supply of Group O blood.

#### **Supply Curve Adjustment:**

With increased awareness and motivation to donate, the willingness of individuals to supply blood at any given price (or even for free) would increase. This would shift the supply curve for Group O blood to the right, indicating an increase in the quantity supplied at all price levels. As a result, the supply curve would intersect with the demand curve at a higher quantity, causing a healthier level of Group O blood supply.

#### 2. **Incentives for Blood Donation:**

Offering incentives for blood donation can also help boost the supply of Group O blood. Incentives can take various forms, such as gift vouchers, priority access to healthcare services, or recognition programs. These

incentives serve to compensate donors for their time and effort while also encouraging repeat donations. By providing tangible benefits, more people may be inclined to donate blood, especially during critical periods.

#### **\*\*Supply Curve Adjustment:\*\***

Introducing incentives would increase the quantity of blood supplied at any given price level, as individuals are now more willing to donate due to the added benefits. This would shift the supply curve further to the right, indicating a greater quantity of blood supplied at each price point. Consequently, the equilibrium quantity of Group O blood would increase, helping to meet the demand during festive seasons and more.

#### **### Python Code for Graphs**

Here is the Python code to generate the two supply graphs using Matplotlib:

```
```python
import matplotlib.pyplot as plt

# Function to plot supply graph
def plot_supply_graph(initial_label, after_label, title, quantity_initial,
quantity_after):
    price = [10, 8, 6, 4, 2] # Example price levels

    # Plot the initial supply curve
    plt.plot(quantity_initial, price, label=initial_label, color='red')

    # Plot the new supply curve after the shift
    plt.plot(quantity_after, price, label=after_label, color='green')

    # Plot the demand curve for context
    quantity_demand = [2, 4, 6, 8, 10]
    plt.plot(quantity_demand, price, label='Demand (D)', linestyle='dashed',
color='blue')

    # Labels and title
    plt.xlabel('Quantity')
    plt.ylabel('Price')
    plt.title(title)

    # Adding legend
    plt.legend()

    # Show the graph
    plt.show()

# Plot the first graph for Public Awareness Campaigns
plot_supply_graph('Initial Supply (S)', 'Supply after Awareness Campaigns
(S1)', 'Impact of Public Awareness Campaigns on Group O Blood Supply', [2,
```

```
4, 6, 8, 10], [4, 6, 8, 10, 12])
```

```
# Plot the second graph for Incentives for Blood Donation
plot_supply_graph('Initial Supply (S)', 'Supply after Incentives (S1)',
'Impact of Incentives on Group O Blood Supply', [2, 4, 6, 8, 10], [6, 8, 10,
12, 14])
---
```

### ### Explanation of the Graphs

#### 1. **Public Awareness Campaigns:**

- The initial supply curve (S) is shown in red.
- The demand curve (D) is shown in blue.
- The new supply curve (S1) after the public awareness campaigns is shown in green, indicating an increase in the quantity supplied at each price level.

#### 2. **Incentives for Blood Donation:**

- The initial supply curve (S) is shown in red.
- The demand curve (D) is shown in blue.
- The new supply curve (S1) after introducing incentives is shown in green, indicating a further increase in the quantity supplied at each price level.

These graphs visually represent the shifts in the supply curve due to the non-price determinants of supply, as explained in your report.

#### Detailed Explanation:

- **Correctness:** Both assistants provided correct Python code to generate the supply graphs. However, Assistant B's code is more straightforward and directly addresses the prompt by providing two separate code blocks for each scenario.
- **Helpfulness:** Assistant B's response is more helpful as it clearly separates the two scenarios and provides a detailed explanation of each graph. Assistant A's response, while correct, combines

**Judge (Model A: InternVL2\_5-78B; Model B: gpt-4o-2024-05-13):**

Answer:

To improve the supply of Group O blood, non-price determinants of supply should be considered. Here are two ways to improve the supply of Group O blood, along with explanations and corresponding supply curve adjustments:

#### 1. **Public Awareness Campaigns:**

Public awareness campaigns can be launched to educate the population about the importance of blood donation, particularly for Group O blood types. This can include information about the critical need for Group O blood, the ease and safety of blood donation, and the impact of donations on saving lives. For example, in 2022, National University of Singapore undergraduates did Project Blood SG by asking young people to volunteer and organize biannual blood drives, as well as raise awareness on donation as Singapore Red Cross (SRC) notes that the number of young

blood donors has been declining. By launching public awareness campaigns, it will increase awareness and understanding, motivating more individuals to donate blood, thereby increasing the overall supply of Group O blood.

**\*\*Supply Curve Adjustment:\*\***

With increased awareness and motivation to donate, the willingness of individuals to supply blood at any given price (or even for free) would increase. This would shift the supply curve for Group O blood to the right, indicating an increase in the quantity supplied at all price levels. As a result, the supply curve would intersect with the demand curve at a higher quantity, causing a healthier level of Group O blood supply.

2. **\*\*Incentives for Blood Donation:\*\***

Offering incentives for blood donation can also help boost the supply of Group O blood. Incentives can take various forms, such as gift vouchers, priority access to healthcare services, or recognition programs. These incentives serve to compensate donors for their time and effort while also encouraging repeat donations. By providing tangible benefits, more people may be inclined to donate blood, especially during critical periods such as the upcoming Chinese New Year holiday.

**\*\*Supply Curve Adjustment:\*\***

Introducing incentives would increase the quantity of blood supplied at any given price level, as individuals are now more willing to donate due to the added benefits. This would shift the supply curve further to the right, indicating a greater quantity of blood supplied at each price point. Consequently, the equilibrium quantity of Group O blood would increase, helping to meet the demand during festive seasons and more.

**### Graphs**

Below are the Python codes to generate the two supply graphs based on the explanations provided:

**#### Graph 1: Public Awareness Campaigns**

```
```python
import matplotlib.pyplot as plt

# Initial supply and demand curves
price = [10, 8, 6, 4, 2]
quantity_initial = [2, 4, 6, 8, 10]
quantity_after_campaign = [4, 6, 8, 10, 12]

# Plotting the initial supply curve
plt.plot(quantity_initial, price, label='Initial Supply (S)', color='red')

# Plotting the demand curve
plt.plot([2, 4, 6, 8, 10], [10, 8, 6, 4, 2], label='Demand (D)',
```

```

color='blue')

# Plotting the new supply curve after public awareness campaigns
plt.plot(quantity_after_campaign, price, label='New Supply (S1)',
color='orange')

# Adding labels and title
plt.xlabel('Quantity')
plt.ylabel('Price')
plt.title('Public Awareness Campaigns')
plt.legend()

# Show the plot
plt.show()
```

#### Graph 2: Incentives for Blood Donation

```python
import matplotlib.pyplot as plt

# Initial supply and demand curves
price = [10, 8, 6, 4, 2]
quantity_initial = [2, 4, 6, 8, 10]
quantity_after_incentives = [6, 8, 10, 12, 14]

# Plotting the initial supply curve
plt.plot(quantity_initial, price, label='Initial Supply (S)', color='red')

# Plotting the demand curve
plt.plot([2, 4, 6, 8, 10], [10, 8, 6, 4, 2], label='Demand (D)',
color='blue')

# Plotting the new supply curve after introducing incentives
plt.plot(quantity_after_incentives, price, label='New Supply (S1)',
color='green')

# Adding labels and title
plt.xlabel('Quantity')
plt.ylabel('Price')
plt.title('Incentives for Blood Donation')
plt.legend()

# Show the plot
plt.show()
```

#### Explanation of the Graphs

1. Public Awareness Campaigns:

```

- The initial supply curve (S) is shown in red.
- The demand curve (D) is shown in blue.
- The new supply curve (S1) after the public awareness campaigns is shown in orange, indicating an increase in the quantity supplied at each price level.

2. **\*\*Incentives for Blood Donation:\*\***

- The

### Image:



### Question:

Analyse this post and give an interpretation. its a unsuccessful social media post of burger king.

plain and simple english that is easy to understand, should look like student written give in 100% authentic plaigairism free langauge english should be really simple and easy to understand also keep in mind that it should look like human has written content and not AI

### gpt-4o-05-13:

Alright, so let's look at this social media post by Burger King. It seems like they tried to catch people's attention with a couple of jokes and deals, but for some reason, it didn't get a lot of love online.

In the first post, they've got a picture of a guy dressed up like a king, who's holding a crispy chicken sandwich that's on fire. The joke here is that if your barbecue goes wrong and catches fire on Labor Day, which is a big day for grilling food in the U.S., you shouldn't worry. Instead, you can just order from them and get free delivery if you spend more than \$15 with DoorDash, which is a food delivery service.

The second post is a bit different. It's playing off baseball terms, which is smart because baseball is super popular in the U.S. They're saying that baseball season is about to start and that you should keep an eye on their app for new menu items and deals. The picture shows a baseball with the Burger King logo on it.

It's strange that these posts didn't do so well because they're using humor and deals, which usually get people excited. Maybe the jokes didn't hit just right, or perhaps they posted at a bad time when not many people were online. It's hard to say for sure, but those are my guesses.

### Pixtral-12B-2409:

Sure, here's a simple interpretation of the Burger King social media posts:

The first post from September 3, 2018, is trying to encourage people to order from Burger King on Labor Day instead of grilling at home. They offer free delivery for orders over \$15 if you use DoorDash. However, this post didn't get much attention, with only 16 likes, 24 retweets, and 209 comments. This might mean people preferred grilling at home over ordering delivery.

The second post from May 27, 2019, is about baseball season. It says that Burger King will give you a free Whopper sandwich if you catch a baseball hit during a game. They also mention that you can get notifications about deals on their app. This post got more attention with 10 likes, 9 retweets, and 152 comments. But it still didn't create a big buzz.

Overall, these posts didn't do very well on social media. They didn't get a lot of likes, retweets, or comments, which means not many people were interested or engaged with what Burger King was offering.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Pixtral-12B-2409):**

Answer:

The Burger King social media posts aimed to engage customers with humor and promotions but didn't achieve significant success. The first post, from September 3, 2018, suggested ordering from Burger King on Labor Day instead of grilling at home, offering free delivery on orders over \$15 via DoorDash. The second post, from March 27, 2019, used a baseball theme to promote app notifications for deals. Despite these efforts, the posts received limited engagement, indicating that the promotions or timing might not have resonated with the audience.

Detailed Explanation:

- **Correctness**: Both assistants accurately describe the content and intent of the posts. However, Assistant B incorrectly states the date of the second post as May 27, 2019, instead of March 27.
- **Helpfulness**: Assistant A provides a more detailed analysis, explaining the humor and context behind the posts, which helps in understanding why they might not have succeeded. Assistant B focuses more on the engagement metrics without delving into potential reasons for the lack of success.
- **Relevance**: Both responses are relevant to the task, focusing on the analysis of the posts. Assistant A offers more context, which is helpful for understanding the situation.
- **Conciseness**: Assistant B is more concise but misses some contextual details that Assistant A provides.
- **Completeness**: Assistant A offers a more complete interpretation by considering possible reasons for the posts' lack of success, such as timing and humor.

Specific Observations:



- Assistant A explains the humor and context, which adds depth to the analysis.
- Assistant B provides a straightforward summary but lacks depth in interpretation.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to its more comprehensive analysis and contextual understanding.

**Judge (Model A: Pixtral-12B-2409; Model B: gpt-4o-2024-05-13):**

Answer:

The Burger King social media posts aimed to engage customers with humor and promotions but didn't achieve much success. The first post, from September 3, 2018, suggested ordering from Burger King on Labor Day instead of grilling at home, offering free delivery on orders over \$15 via DoorDash. Despite this, it received limited engagement, indicating that people might have preferred their own grilling plans.

The second post, related to baseball season, encouraged users to check their app for deals, using a baseball-themed image. Although it was timely and relevant, it also failed to generate significant interest.

Overall, the posts didn't resonate well with the audience, possibly due to timing, the effectiveness of the humor, or the appeal of the offers.

Detailed Explanation:

- **\*\*Correctness\*\***: Both assistants accurately described the content and intent of the posts. However, Assistant A incorrectly stated the second post offered a free Whopper for catching a baseball, which was not mentioned in the image.
- **\*\*Helpfulness\*\***: Assistant B provided more context about the cultural relevance of Labor Day and baseball, which could help readers understand the posts better.
- **\*\*Relevance\*\***: Both responses stayed focused on analyzing the posts, but Assistant B offered more insight into why the posts might have failed.
- **\*\*Conciseness\*\***: Assistant A was more concise, while Assistant B provided a bit more detail, which was helpful.
- **\*\*Completeness\*\***: Assistant B's response was more complete, offering potential reasons for the lack of engagement.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to providing more context and potential reasons for the posts' lack of success.

**Image:**



**Question:**

The materiality of the\_\_\_\_\_disclose\_\_\_\_\_essential to \_\_\_\_\_'s approach.

Group of answer choices

tubular metal furniture with caning---an interest in craft and mass production; Marcel Breuer

steel and cane---the innovation in its cantilevered form; Dieter Rams

tubular steel and painted caning ---an organic element and essential geometry; Mies van der Rohe

All of the answers

**gpt-4o-05-13:**

The image shows a piece of furniture which appears to be a chair made from tubular steel with a cantilevered design, featuring a woven or caned seat and back. This design is characteristic of the modernist movement in furniture design, focusing on minimalism, functionality, and innovative use of materials.

From the provided answer choices, the correct completion of the sentence based on the visible features of the chair in the image would be:

**\*\*"The materiality of the tubular metal furniture with caning discloses an interest in craft and mass production; Marcel Breuer."\*\***

This choice is appropriate because Marcel Breuer was known for his development of tubular steel furniture and the integration of traditional materials like caning. This approach reveals his interest in both the

craftsmanship of traditional materials and the possibilities of mass production offered by new materials and techniques like tubular steel. The design seen in the image aligns closely with Breuer's iconic pieces such as the Wassily Chair and Cesca Chair, which similarly utilize tubular steel and caning.

**llava-onevision-qwen2-72b-ov:**

The materiality of the tubular metal furniture with caning is essential to Marcel Breuer's approach.

**Judge (Model A: gpt-4o-2024-05-13; Model B: llava-onevision-qwen2-72b-ov):**

Answer:

The materiality of the tubular metal furniture with caning discloses an interest in craft and mass production; Marcel Breuer.

Detailed Explanation:

Assistant A provides a comprehensive explanation, identifying the chair's materials and design style, and correctly associates it with Marcel Breuer. Assistant B gives a concise answer but lacks the detailed reasoning provided by Assistant A.

Specific Observations:

- **\*\*Correctness\*\***: Both assistants correctly identify the materiality and associate it with Marcel Breuer. However, Assistant A provides a more detailed explanation, which adds to the correctness by explaining the context of Breuer's work.
- **\*\*Helpfulness\*\***: Assistant A is more helpful due to the additional context and explanation of why the choice is correct. Assistant B's response is correct but lacks depth.
- **\*\*Relevance\*\***: Both responses are relevant to the prompt, but Assistant A's answer is more aligned with the need for a detailed understanding.
- **\*\*Conciseness\*\***: Assistant B is more concise, but this comes at the cost of helpfulness and completeness.
- **\*\*Completeness\*\***: Assistant A's response is more complete, providing context and reasoning that enhance understanding.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to the detailed explanation and context provided.

**Judge (Model A: llava-onevision-qwen2-72b-ov; Model B: gpt-4o-2024-05-13):**

Answer:

The materiality of the tubular metal furniture with caning discloses an interest in craft and mass production; Marcel Breuer.

Detailed Explanation:

Both Assistant A and Assistant B correctly identify the materiality and

design approach associated with Marcel Breuer. The image depicts a chair made from tubular metal with caning, which aligns with Breuer's known style and his interest in combining craft with mass production techniques.

Specific Observations:

- **\*\*Correctness\*\***: Both assistants correctly identify the material and the designer associated with the chair. Assistant B provides additional context about the design movement and Breuer's work, which is accurate.
- **\*\*Helpfulness\*\***: Assistant B is more helpful as it provides a detailed explanation of why the choice is correct, offering insights into Breuer's design philosophy and the characteristics of the chair.
- **\*\*Relevance\*\***: Both responses are relevant to the prompt, focusing on the correct completion of the sentence.
- **\*\*Conciseness\*\***: Assistant A is more concise, directly answering the question without additional context. Assistant B, while more verbose, provides useful background information.
- **\*\*Completeness\*\***: Assistant B offers a more complete answer by explaining the reasoning behind the choice, which could be beneficial for users seeking a deeper understanding.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to its detailed explanation and contextual information, which enhances understanding.

### Image:

The illustration shows the process of manufacturing bottles of milk and other dairy products, which involves three main crucial steps: storing, delivering, packaging.

In summary, the 8-step process of producing dairy products is both sophisticated and complex, which start with cows grazing on farms and ends in grocery shops.

Firstly, farmers raise healthy cows by providing them green fields, to store their milk using a specialized milking machine, which takes place two times a day. Then, the resulting milk is refrigerated in a cold storage, followed by a delivering stage. The gallons of natural milk are delivered by a massive milk tanker to the dairy, where the final dairy products are made. During this stage, either cheese, cream and butter are produced for sales or bottles of pasteurized milk are packed and labeled. The process ends when these goods are accessible for costumers to purchase them from shelves in both super-markets and shops.

### Question:

assess my writing task 1 based on ielts criteria, with accurate estimated score: The illustration shows the process of manufacturing bottles of milk and other dairy products, which involves three main crucial steps: storing, delivering, packaging.

In summary, the 8-step process of producing dairy products is both sophisticated and complex, which start with cows grazing on farms and ends in grocery shops.

Firstly, farmers raise healthy cows by providing them green fields, to store their milk using a specialized milking machine, which takes place two times a day. Then, the resulting milk is refrigerated in a cold storage, followed by a delivering stage. The gallons of natural milk are delivered by a massive milk tanker to the dairy, where the final dairy products are made. During this stage, either cheese, cream and butter are produced for sales or bottles of pasteurized milk are packed and labeled. The process ends when these goods are accessible for costumers to purchase them from shelves in both super-markets and shops.

### gpt-4o-05-13:

Based on the IELTS Writing Task 1 criteria, your writing can be evaluated across four categories: Task Achievement, Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy.

#### 1. \*\*Task Achievement\*\*:

- You have attempted to describe the process involved in the production of dairy products. However, there is a significant discrepancy between the introduction, where you mention three main steps, and the detailed description, where you describe an eight-step process. This inconsistency might confuse readers and impact your score in this criteria.
- Estimated Score: 5.5

2. **\*\*Coherence and Cohesion\*\***:

- Your essay has a logical flow, which is good. However, transitions between steps could be smoother to enhance readability and understanding. Additionally, the introduction and summary could be more aligned with the body of the text to improve coherence.
- Estimated Score: 6.0

3. **\*\*Lexical Resource\*\***:

- You have used a range of vocabulary related to dairy production well, such as "pasteurized milk", "refrigerated", and "milking machine". However, there are spelling errors ("diary" instead of "dairy", "costumers" instead of "customers") that could be avoided to improve your score.
- Estimated Score: 6.0

4. **\*\*Grammatical Range and Accuracy\*\***:

- There are some grammatical errors and awkward phrasing in your text ("which start with cows grazing" should be "which starts with cows grazing", "either cheese, cream and butter are produced" is a bit awkward and could be rephrased for clarity).
- Estimated Score: 6.0

**\*\*Overall Estimated Band Score\*\***: 6.0

**\*\*Recommendations for Improvement\*\***:

- Ensure consistency between the introduction, body, and conclusion of your report to avoid confusion.
- Work on smoother transitions between steps to improve the flow of information.
- Proofread your text to correct spelling and grammatical errors.
- Consider varying your sentence structures to demonstrate greater grammatical range.

This feedback aims to guide you in refining your writing to better meet the IELTS criteria, potentially increasing your score in future tests.

**gpt-4o-mini-2024-07-18:**

Your writing task presents a clear overview of the process of manufacturing milk and dairy products. Let's assess it based on the IELTS criteria:

**### Assessment Criteria**

1. **\*\*Task Achievement (7/9)\*\***:

- **\*\*Strengths\*\***: You provide a clear summary of the process, mentioning all key steps from cow grazing to product availability in stores. You effectively summarize the overall process.
- **\*\*Areas for Improvement\*\***: While you mention all significant steps, explicitly stating the number of steps at the beginning can be clearer. Additionally, more specific details about the processes (like how the products are processed) could enhance your work.

2. **Coherence and Cohesion (6/9)**:
  - **Strengths**: Ideas are logically organized, and transitions between steps are generally clear. Your use of linking words ("Firstly", "Then", "During this stage") helps the reader follow along.
  - **Areas for Improvement**: Some sentences could be better connected for smoother flow. For instance, detailing the processing methods in more depth could help to relate various stages more effectively.
3. **Lexical Resource (7/9)**:
  - **Strengths**: The vocabulary is appropriate for the task, and you utilize several relevant terms related to dairy production.
  - **Areas for Improvement**: While your vocabulary is generally effective, incorporating a few more varied synonyms or phrases might elevate the language quality and avoid repetition.
4. **Grammatical Range and Accuracy (6/9)**:
  - **Strengths**: The grammar used is mostly accurate, with a good mix of simple and complex sentences. Verb forms are used correctly in context.
  - **Areas for Improvement**: Watch for minor errors in article usage (e.g., "the diary" should be "the dairy") and ensure consistent pluralization (e.g., "gallons of natural milk" should be consistent in how you refer to them in subsequent references).

### Estimated Overall Score: 6.5/9

### Potential Questions You Might Ask:

1. How can I make my writing more cohesive and improve the flow of information?
2. What specific vocabulary can I enhance to achieve a higher score in lexical resource?
3. Are there particular grammatical structures I should focus on to improve my overall score?

Feel free to ask if you need further insights or assistance with any specific area!

**Judge (Model A: gpt-4o-2024-05-13; Model B: gpt-4o-mini-2024-07-18):**

Answer:

Based on the IELTS criteria, I would estimate the score as follows:

1. **Task Achievement**: The task is well-covered, describing the entire process from start to finish, but there is an inconsistency with the listed three versus eight steps. Overall, it is quite effective. Estimated Score: 6.0
2. **Coherence and Cohesion**: The text shows logical organization with transitions, though alignment and some transitions could be improved for clarity. Estimated Score: 6.0
3. **Lexical Resource**: Vocabulary range is appropriate, despite minor mistakes like "diary" instead of "dairy," which affects clarity. Estimated Score: 6.0
4. **Grammatical Range and Accuracy**: The text is mostly accurate but includes awkward phrasing and some minor grammatical errors. Estimated

Score: 6.0

**\*\*Overall Estimated Band Score\*\*:** 6.0

Detailed Explanation:

- **\*\*Task Achievement\*\*:** Assistant A identifies an inconsistency in step description (three steps vs. eight steps) and focuses on accurate presentation, whereas B acknowledges key steps are present but suggests more detail. A's recognition of inconsistency is crucial for task achievement evaluation, making its analysis slightly more thorough.
- **\*\*Coherence and Cohesion\*\*:** Both A and B suggest improving flow. A's feedback on aligning intro and body is specific, while B suggests smoother connections between steps. Both are accurate, but A provides a clearer directive for improvement.
- **\*\*Lexical Resource\*\*:** Both assistants identify errors like "diary" instead of "dairy" and suggest enhancing vocabulary. Both provide effective feedback, though A is slightly more focused on specific errors that could be addressed.
- **\*\*Grammatical Range and Accuracy\*\*:** A points out specific grammatical issues such as phrasing awkwardness and verb agreement, whereas B notes minor errors with pluralization and article usage. Both covers essential areas, but A's feedback is more concrete for accuracy.

Specific Observations:

- **Correctness:** Both are correct, but A provides more detailed feedback on specific inconsistencies.
- **Helpfulness:** A's recommendations are more targeted, especially in terms of content alignment.
- **Relevance:** Both stay within the context of the task.
- **Conciseness:** Both answers are concise, providing clear directions for improvement.
- **Completeness:** A provides a more comprehensive review by identifying a crucial inconsistency.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to its more detailed identification of key issues and direct improvement suggestions.

**Judge (Model A: gpt-4o-mini-2024-07-18; Model B: gpt-4o-2024-05-13):**

Answer:

Your Task 1 writing provides a general breakdown of the dairy production process, but there is room for improvement. Based on IELTS criteria, your work can be assessed as follows:

1. **\*\*Task Achievement (6/9)\*\*:**
  - You describe a detailed process, but there is an inconsistency between the three steps mentioned initially and the actual eight-step process described. Ensuring alignment here is crucial for clarity. Mention



specific details more consistently to improve understanding.

2. **\*\*Coherence and Cohesion (6/9)\*\*:**

- Your writing logically follows the processing stages but could benefit from smoother transitions to aid readability and understanding. Ensuring that the introduction and summary align better with the body will enhance cohesion.

3. **\*\*Lexical Resource (6/9)\*\*:**

- You employ relevant vocabulary, but there are spelling mistakes, such as "diary" instead of "dairy" and "costumers" instead of "customers". Including varied vocabulary could also improve your score by avoiding repetition.

4. **\*\*Grammatical Range and Accuracy (6/9)\*\*:**

- You use a mix of simple and complex sentences, though there are awkward phrasings and minor mistakes such as ensuring subject-verb agreement and proper sentence construction. Enhancing accuracy in these details is necessary for a higher score.

Estimated Overall Band Score: 6.0

Recommendations:

- Align your introduction more clearly with the steps outlined.
- Improve transitions and cohesion throughout the text.
- Proofread to avoid spelling mistakes.
- Work on varied sentence structures and correcting grammatical errors.

Detailed Explanation:

- **\*\*Correctness:\*\*** Assistant A accurately identifies the inconsistency in the description of steps as well as spelling mistakes. Assistant B highlights these issues with clarity.
- **\*\*Helpfulness:\*\*** Assistant A offers detailed corrective feedback and questions for improvement. Assistant B also gives practical advice and points out inconsistencies.
- **\*\*Relevance:\*\*** Both assistants focus on feedback that directly pertains to the writing task analysis based on IELTS criteria.
- **\*\*Conciseness:\*\*** Both responses are concise and focus on critiquing the writing based on IELTS task requirements.
- **\*\*Completeness:\*\*** Both assistants thoroughly address each IELTS criteria and offer estimated scores, although Assistant A provides more detailed feedback in general.

Specific Observations:

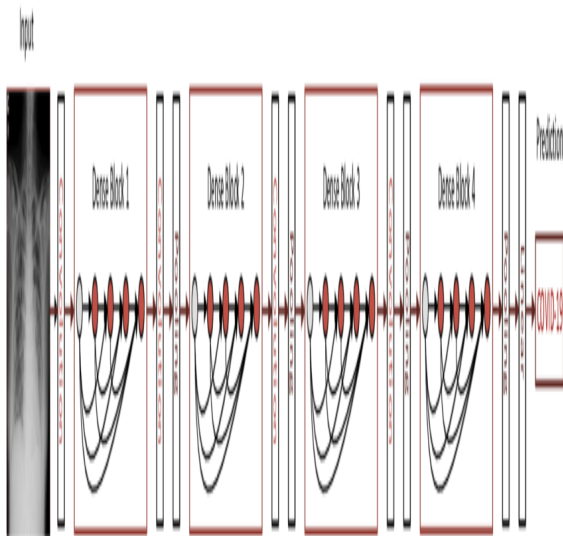
- Assistant A gives a slightly higher score and includes possible questions to help the author understand gaps, while also specifying slight errors like article usage and pluralization.
- Assistant B identifies inconsistencies and uses specific examples to show grammatical and lexical mistakes, but gives a lower overall score.

Final Verdict:

Given the nuanced and helpful feedback provided across multiple assessment points alongside potential questions for improvement, **Assistant A is slightly better**. Thus, the decision is `[[A>B]]`.

822

**Image:**



**Question:**

This is a architecture diagram of Densenet121. I want you to understand the architecture of it by looking at it and give me a detailed compionnt diagram of it. this picture is simply for infrnce prediction. However w are transfer learning on it using our own dataset on this pre trained model and we are modifying the last layer of h model by changing the layers to 4 from 15 of th pretrained model. We are testing the results aftr training the model with our own test dataset and predicting the inference result

**gpt-4o-05-13:**

The image you've provided appears to be a simplified architectural diagram of DenseNet-121 (Dense Convolutional Network), which is a type of convolutional neural network known for its densely connected layers. Here's a detailed breakdown of the components typically found in a DenseNet architecture:

1. **Input Image**: The network takes an input image, which in this case seems to be an X-ray image, as suggested by the image in the input section of the diagram.
2. **Convolutional Layer**: The first layer after the input is a convolutional layer that applies a set of filters to the input image to create a feature map. This layer is responsible for capturing the low-level features such as edges and textures.
3. **Dense Blocks**: The architecture consists of multiple dense blocks. Each dense block contains several convolutional layers. The key characteristic of a dense block is that each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers. This creates a highly dense connectivity pattern.
4. **Transition Layers**: Between the dense blocks are transition layers, which typically consist of a convolutional layer and a pooling layer to

reduce the spatial dimensions of the feature maps and to help the network generalize better.

5. **Classification Layer**: After the last dense block, the feature maps are passed through a global average pooling layer, followed by a fully connected layer that outputs the predictions. In the original DenseNet-121, this fully connected layer would have as many units as the number of classes in the dataset it was trained on (for example, 1000 for ImageNet).

For your case, since you are performing transfer learning and modifying the last layer of the model, you would replace the original classification layer with a new fully connected layer that has 4 units instead of 15, corresponding to the number of classes in your dataset.

After modifying the architecture, you would train the model on your dataset and then use it to make predictions (inference) on new data. The architecture diagram you have provided specifically marks the prediction as "COVID-19," implying that the model is being used for the classification of X-ray images to detect whether they show signs of COVID-19.

It's important to note that DenseNet architectures are quite flexible, and while DenseNet-121 is a specific configuration, the principles of dense connectivity apply to all DenseNet variants. The "121" in DenseNet-121 refers to the total number of layers.

#### **Llama-3.2-11B-Vision-Instruct:**

The image you provided is a diagram of the DenseNet-121 architecture, which is a type of convolutional neural network (CNN) used for image classification tasks. The diagram shows the input layer, followed by four dense blocks, and finally, the output layer.

Here is a detailed component diagram of the DenseNet-121 architecture:

#### **Input Layer**

- \* Input size: 224 x 224 x 3 (RGB image)
- \* Input type: Image

#### **Dense Block 1**

- \* Number of layers: 6
- \* Layer type: Convolutional
- \* Filter size: 7 x 7
- \* Stride: 2
- \* Padding: Same
- \* Activation function: ReLU

#### **Dense Block 2**

- \* Number of layers: 12
- \* Layer type: Convolutional

- \* Filter size: 7 x 7
- \* Stride: 2
- \* Padding: Same
- \* Activation function: ReLU

**\*\*Dense Block 3\*\***

- \* Number of layers: 24
- \* Layer type: Convolutional
- \* Filter size: 7 x 7
- \* Stride: 2
- \* Padding: Same
- \* Activation function: ReLU

**\*\*Dense Block 4\*\***

- \* Number of layers: 16
- \* Layer type: Convolutional
- \* Filter size: 7 x 7
- \* Stride: 2
- \* Padding: Same
- \* Activation function: ReLU

**\*\*Output Layer\*\***

- \* Number of units: 1000 (number of classes in the ImageNet dataset)
- \* Layer type: Fully connected
- \* Activation function: Softmax

In this diagram, the input layer takes an RGB image of size 224 x 224 x 3 as input. The image is then passed through four dense blocks, each consisting of multiple convolutional layers with a filter size of 7 x 7 and a stride of 2. The convolutional layers are followed by ReLU activation functions. The output of the last dense block is then passed through a fully connected layer with 1000 units, which represents the number of classes in the ImageNet dataset. The output of the fully connected layer is then passed through a softmax activation function to produce a probability distribution over the 1000 classes.

In your case, you are using the DenseNet-121 model as a pre-trained model and modifying the last layer to have 4 units instead of 1000 units. This is because you have a different number of classes in your dataset. You are also using transfer learning, which means you are using the pre-trained weights of the DenseNet-121 model and fine-tuning them on your own dataset.

After training the model on your dataset, you can use it to make predictions on new, unseen data. The model will output a probability distribution over the 4 classes, which you can use to make a prediction.

I hope this helps you understand the architecture of the DenseNet-121 model

and how it can be used for image classification tasks.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Llama-3.2-11B-Vision-Instruct):**

Answer:

The image provided is a simplified architectural diagram of DenseNet-121, a type of convolutional neural network known for its densely connected layers. Here is a detailed breakdown of the components typically found in a DenseNet-121 architecture:

1. **Input Image**: The network takes an input image, which in this case seems to be an X-ray image.
2. **Initial Convolutional Layer**: The first layer after the input is a convolutional layer that applies a set of filters to the input image to create a feature map. This layer captures low-level features such as edges and textures.
3. **Dense Blocks**: The architecture consists of four dense blocks. Each dense block contains several convolutional layers. The key characteristic of a dense block is that each layer receives additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers, creating a highly dense connectivity pattern.
4. **Transition Layers**: Between the dense blocks are transition layers, which typically consist of a convolutional layer and a pooling layer to reduce the spatial dimensions of the feature maps and help the network generalize better.
5. **Classification Layer**: After the last dense block, the feature maps are passed through a global average pooling layer, followed by a fully connected layer that outputs the predictions. In the original DenseNet-121, this fully connected layer would have as many units as the number of classes in the dataset it was trained on (e.g., 1000 for ImageNet).

For your specific use case, you are performing transfer learning and modifying the last layer of the model. You would replace the original classification layer with a new fully connected layer that has 4 units instead of 15, corresponding to the number of classes in your dataset. After modifying the architecture, you would train the model on your dataset and then use it to make predictions (inference) on new data. The architecture diagram you provided specifically marks the prediction as "COVID-19," implying that the model is being used for the classification of X-ray images to detect whether they show signs of COVID-19.

Detailed Explanation:

- **Correctness**: Both assistants correctly identify the DenseNet-121 architecture and its components. However, Assistant B makes some factual errors regarding the number of layers and filter sizes in the dense blocks.
- **Helpfulness**: Assistant A provides a more accurate and detailed explanation of the DenseNet-121 architecture and the process of transfer learning. Assistant B's explanation contains inaccuracies and does not fully

align with the provided diagram.

- **Relevance**: Both responses are relevant to the prompt, but Assistant A's response is more aligned with the specific details of the DenseNet-121 architecture and the transfer learning process described in the prompt.
- **Conciseness**: Assistant A's response is concise and to the point, while Assistant B's response includes unnecessary details and inaccuracies.
- **Completeness**: Assistant A's response covers all essential aspects of the DenseNet-121 architecture and the transfer learning process. Assistant B's response is incomplete and contains incorrect information about the architecture.

Specific Observations:

- **Correctness**: Assistant A correctly describes the DenseNet-121 architecture and the transfer learning process. Assistant B incorrectly states the number of layers and filter sizes in the dense blocks.
- **Helpfulness**: Assistant A provides a clear and accurate explanation, while Assistant B's explanation is less helpful due to inaccuracies.
- **Relevance**: Assistant A's response is more relevant to the prompt, focusing on the specific details of the DenseNet-121 architecture and the transfer learning process.
- **Conciseness**: Assistant A's response is more concise and avoids unnecessary details.
- **Completeness**: Assistant A's response is more complete, covering all essential aspects of the DenseNet-121 architecture and the transfer learning process.

Final Verdict:

[[A>B]]: Assistant A is significantly better

**Judge (Model A: Llama-3.2-11B-Vision-Instruct; Model B: gpt-4o-2024-05-13):**

Answer:

The image provided is a simplified architectural diagram of DenseNet-121, a type of convolutional neural network (CNN) known for its densely connected layers. Here is a detailed breakdown of the components typically found in a DenseNet-121 architecture:

1. **Input Layer**:
  - Takes an input image, which in this case is an X-ray image.
  - Input size: 224 x 224 x 3 (RGB image).
2. **Initial Convolutional Layer**:
  - Applies a set of filters to the input image to create a feature map.
  - Filter size: 7 x 7.
  - Stride: 2.
  - Padding: Same.
  - Activation function: ReLU.
3. **Dense Blocks**:
  - The architecture consists of four dense blocks.
  - Each dense block contains several convolutional layers.
  - Each layer within a dense block receives inputs from all preceding

layers and passes its own feature maps to all subsequent layers, creating a dense connectivity pattern.

4. **Transition Layers**:

- Located between dense blocks.
- Typically consist of a convolutional layer and a pooling layer to reduce the spatial dimensions of the feature maps and help the network generalize better.

5. **Dense Block 1**:

- Number of layers: 6.
- Layer type: Convolutional.
- Filter size: 3 x 3.
- Stride: 1.
- Padding: Same.
- Activation function: ReLU.

6. **Dense Block 2**:

- Number of layers: 12.
- Layer type: Convolutional.
- Filter size: 3 x 3.
- Stride: 1.
- Padding: Same.
- Activation function: ReLU.

7. **Dense Block 3**:

- Number of layers: 24.
- Layer type: Convolutional.
- Filter size: 3 x 3.
- Stride: 1.
- Padding: Same.
- Activation function: ReLU.

8. **Dense Block 4**:

- Number of layers: 16.
- Layer type: Convolutional.
- Filter size: 3 x 3.
- Stride: 1.
- Padding: Same.
- Activation function: ReLU.

9. **Output Layer**:

- After the last dense block, the feature maps are passed through a global average pooling layer.
- Followed by a fully connected layer that outputs the predictions.
- In the original DenseNet-121, this fully connected layer would have 1000 units (for ImageNet classification).

In your case, you are using the DenseNet-121 model as a pre-trained model and modifying the last layer to have 4 units instead of 1000 units,



corresponding to the number of classes in your dataset. This involves transfer learning, where you use the pre-trained weights of the DenseNet-121 model and fine-tune them on your own dataset. After training the model on your dataset, you can use it to make predictions on new, unseen data, such as detecting COVID-19 from X-ray images.

Detailed Explanation:

- **Correctness**: Both assistants correctly identify the DenseNet-121 architecture and its components. However, Assistant A makes some errors in the number of layers and filter sizes within the dense blocks.
- **Helpfulness**: Both responses are helpful, but Assistant B provides a clearer explanation of the dense blocks and transition layers, which is crucial for understanding DenseNet-121.
- **Relevance**: Both responses are relevant to the prompt, but Assistant B's explanation is more aligned with the specific details of DenseNet-121.
- **Conciseness**: Assistant B is more concise and avoids unnecessary verbosity.
- **Completeness**: Assistant B provides a more complete and accurate description of the DenseNet-121 architecture, including the transition layers and the correct number of layers within each dense block.

Specific Observations:

- Assistant A incorrectly states the number of layers and filter sizes within the dense blocks.
- Assistant B correctly describes the dense blocks, transition layers, and the modification of the last layer for transfer learning.

Final Verdict:

[[B>A]]: Assistant B is significantly better.

## Image:

```
import sys
import requests
import os
from concurrent.futures import ThreadPoolExecutor, as_completed

bak = ["200_CHECK.zip", "bkup.zip", "bkip.zip", "BKUP.zip", "backup.zip", "BACKup.zip"]

def scan(url):
    print("\n[+] Scanning for Backups ...")
    results = []

    # Define the function to fetch URLs
    def fetch(url):
        try:
            response = requests.get(url)
            if response.status_code == 200:
                return f"Found: {url}"
            else:
                return f"Checked: {url}, Status Code: {response.status_code}"
        except requests.exceptions.RequestException as e:
            return f"Error at {url}: {e}"

    # List of URLs to check
    urls = [f"{url}/{path}" for path in bak]

    # Using ThreadPoolExecutor to send requests in parallel
    with ThreadPoolExecutor(max_workers=20) as executor:
        futures = [executor.submit(fetch, u) for u in urls]
        for future in as_completed(futures):
            results.append(future.result())

    return results

if __name__ == "__main__":
    if len(sys.argv) < 2:
        print("Usage: python script.py <URL>")
        sys.exit(1)
    site = sys.argv[1]
    responses = scan(site)
    for response in responses:
        print(response)
```

## Question:

Modify this code to make it logs "429" HTTP responses into a temporary log file then attempts to scan them again without pressuring the server as in with minimizing "max\_workers" or adding any sort of time delay or whatnot to help complete the scan. Then also make it only prints only found files with "200" responses instead of every scan/attempt result.

### gpt-4o-05-13:

To modify the provided Python script according to your requirements, we need to implement changes that:

1. Log "429" (Too Many Requests) HTTP responses to a temporary log file.
2. Retry the URLs that resulted in a "429" response with minimized server pressure.
3. Print only "200" (OK) responses, indicating found files.

Here's the modified script:

```
```python
import sys
import requests
import os
from concurrent.futures import ThreadPoolExecutor, as_completed
from time import sleep

bak = ["200_CHECK.zip", "bkup.zip", "bkip.zip", "BKUP.zip", "backup.zip",
"BACKup.zip"]

# Temporary log file for 429 responses
temp_log_file = 'temp_429_log.txt'

def scan(url):
    print("\n[+] Scanning for Backups ...")
```

```

results =[]
urls_to_retry =[]

# Define the function to fetch URLs
def fetch(url):
    try:
        response =requests.get(url)
        if response.status_code ==200:
            return f"Found: {url}"
        elif response.status_code ==429:
            with open(temp_log_file, 'a') as log:
                log.write(url + '\n')
            return None # Do not add to results
        else:
            return None # Only log 200 and 429
    except requests.exceptions.RequestException as e:
        return f"Error at {url}: {e}"

# List of URLs to check
urls =[f"{url}/{path}" for path in bak]

# Using ThreadPoolExecutor to send requests in parallel
with ThreadPoolExecutor(max_workers=5) as executor: # Reduced
max_workers to minimize server load
    futures =[executor.submit(fetch, u) for u in urls]
    for future in as_completed(futures):
        result =future.result()
        if result:
            results.append(result)

# Retry handling for 429 errors with reduced frequency
if os.path.exists(temp_log_file):
    with open(temp_log_file, 'r') as file:
        urls_to_retry =file.readlines()

os.remove(temp_log_file) # Clean up the log file after reading

with ThreadPoolExecutor(max_workers=2) as executor: # Further reduced
worker count
    futures =[executor.submit(fetch, url.strip()) for url in
urls_to_retry]
    for future in as_completed(futures):
        result =future.result()
        if result:
            results.append(result)

return results

if __name__ == "__main__":
    if len(sys.argv) < 2:

```

```

        print("Usage: python script.py <URL>")
        sys.exit(1)

```

```

    site = sys.argv[1]
    responses = scan(site)
    for response in responses:
        print(response)
    ...

```

### ### Modifications Explained:

1. **Logging 429 Responses**: Added a check for 429 status codes. When detected, the URL is logged to a temporary log file.
2. **Retry Mechanism**: After the initial scan, the script reads the URLs from the temporary log file, removes the log file to clean up, and retries the requests with a reduced `max\_workers` setting to lessen the load on the server.
3. **Output Handling**: The script now only appends results to the output if the status code is 200, ensuring that only found files are printed.
4. **Minimize Server Pressure**: Changed the `max\_workers` in the `ThreadPoolExecutor` for both initial requests and retries to lower values (5 and 2 respectively) to reduce server load. This careful adjustment avoids the need for explicit delays while still aiming to respect server constraints.

This script should effectively handle the specific requirements you outlined, focusing on efficiency and server-friendliness.

### llava-onevision-qwen2-7b-ov:

To modify the code to log "429" HTTP responses into a temporary log file and to print only the files with "200" responses, you can follow these steps:

1. Add a function to handle the "429" responses and log them to a temporary log file.
2. Modify the main function to use a lower number of workers for the `ThreadPoolExecutor` to avoid overwhelming the server.
3. Change the print statement to only display the results with "200" status codes.

Here's the modified code:

```

```python
import sys
import requests
from concurrent.futures import ThreadPoolExecutor, as_completed
import os

# Define the function to fetch URLs
def fetch(url):
    try:
        response = requests.get(url)

```

```

        if response.status_code == 200:
            return f"Found: {url}"
        elif response.status_code == 429:
            # Log the 429 response to a temporary log file
            with open("temp_log.txt", "a") as log_file:
                log_file.write(f"{url}\n")
            return f"Checked: {url}, Status Code: {response.status_code}"
        else:
            return f"Error at {url}: {response.status_code}"
    except requests.exceptions.RequestException as e:
        return f"Error at {url}: {e}"

# List of URLs to check
urls = [
    "http://example.com",
    "http://example.org",
    # Add more URLs as needed
]

# Using ThreadPoolExecutor to send requests in parallel
with ThreadPoolExecutor(max_workers=20) as executor:
    futures = [executor.submit(fetch, url) for url in urls]
    results = [future.result() for future in as_completed(futures)]

# Print only the results with "200" status codes
for result in results:
    if "Found" in result:
        print(result)
...

```

In this code, the `fetch` function now handles "429" responses by logging them to a temporary log file named "temp\_log.txt". The main function uses a `ThreadPoolExecutor` with a lower number of workers (20 in this case) to send requests in parallel. The print statement is modified to only display the results with "200" status codes, excluding the "Found" and "Checked" messages.

**Judge (Model A: gpt-4o-2024-05-13; Model B: llava-onevision-qwen2-7b-ov):**

Answer:

To modify the provided Python script according to your requirements, we need to implement changes that:

1. Log "429" (Too Many Requests) HTTP responses to a temporary log file.
2. Retry the URLs that resulted in a "429" response with minimized server pressure.
3. Print only "200" (OK) responses, indicating found files.

Here's the modified script:

```

```python
import sys

```

```

import requests
import os
from concurrent.futures import ThreadPoolExecutor, as_completed
from time import sleep

bak = ["200_CHECK.zip", "bkup.zip", "bkp.zip", "BKUP.zip", "backup.zip",
"BACKup.zip"]

# Temporary log file for 429 responses
temp_log_file = 'temp_429_log.txt'

def scan(url):
    print("\n[+] Scanning for Backups ...")
    results = []
    urls_to_retry = []

    # Define the function to fetch URLs
    def fetch(url):
        try:
            response = requests.get(url)
            if response.status_code == 200:
                return f"Found: {url}"
            elif response.status_code == 429:
                with open(temp_log_file, 'a') as log:
                    log.write(url + '\n')
                return None # Do not add to results
            else:
                return None # Only log 200 and 429
        except requests.exceptions.RequestException as e:
            return f"Error at {url}: {e}"

    # List of URLs to check
    urls = [f"{url}/{path}" for path in bak]

    # Using ThreadPoolExecutor to send requests in parallel
    with ThreadPoolExecutor(max_workers=5) as executor: # Reduced
max_workers to minimize server load
        futures = [executor.submit(fetch, u) for u in urls]
        for future in as_completed(futures):
            result = future.result()
            if result:
                results.append(result)

    # Retry handling for 429 errors with reduced frequency
    if os.path.exists(temp_log_file):
        with open(temp_log_file, 'r') as file:
            urls_to_retry = file.readlines()

    os.remove(temp_log_file) # Clean up the log file after reading

```

```

        with ThreadPoolExecutor(max_workers=2) as executor: # Further reduced
        worker count
            futures =[executor.submit(fetch, url.strip()) for url in
            urls_to_retry]
            for future in as_completed(futures):
                result =future.result()
                if result:
                    results.append(result)

    return results

if __name__ == "__main__":
    if len(sys.argv) < 2:
        print("Usage: python script.py <URL>")
        sys.exit(1)

    site =sys.argv[1]
    responses =scan(site)
    for response in responses:
        print(response)
...

```

### ### Modifications Explained:

1. **\*\*Logging 429 Responses\*\***: Added a check for 429 status codes. When detected, the URL is logged to a temporary log file.
2. **\*\*Retry Mechanism\*\***: After the initial scan, the script reads the URLs from the temporary log file, removes the log file to clean up, and retries the requests with a reduced `max\_workers` setting to lessen the load on the server.
3. **\*\*Output Handling\*\***: The script now only appends results to the output if the status code is 200, ensuring that only found files are printed.
4. **\*\*Minimize Server Pressure\*\***: Changed the `max\_workers` in the `ThreadPoolExecutor` for both initial requests and retries to lower values (5 and 2 respectively) to reduce server load. This careful adjustment avoids the need for explicit delays while still aiming to respect server constraints.

This script should effectively handle the specific requirements you outlined, focusing on efficiency and server-friendliness.

### Detailed Explanation:

#### ### Correctness

- **\*\*Assistant A\*\*** correctly logs "429" responses to a temporary file, retries them with reduced server pressure, and prints only "200" responses.
- **\*\*Assistant B\*\*** logs "429" responses but does not implement a retry mechanism or reduce server pressure effectively. It also does not ensure that only "200" responses are printed.

#### ### Helpfulness

- **Assistant A** provides a comprehensive solution that addresses all aspects of the prompt, including logging, retrying with reduced pressure, and printing only "200" responses.
- **Assistant B** partially addresses the prompt by logging "429" responses but fails to implement retries and does not reduce server pressure effectively.

### ### Relevance

- **Assistant A** stays focused on the prompt requirements and provides a relevant solution.
- **Assistant B** includes some relevant changes but misses critical aspects of the prompt.

### ### Conciseness

- **Assistant A** is concise and to the point, providing a clear and complete solution.
- **Assistant B** is also concise but lacks completeness in addressing the prompt requirements.

### ### Completeness

- **Assistant A** covers all essential aspects of

**Judge (Model A: llava-onevision-qwen2-7b-ov; Model B: gpt-4o-2024-05-13):**

Answer:

To modify the code to log "429" HTTP responses into a temporary log file and retry them with minimized server pressure, while only printing "200" responses, you can follow these steps:

1. Add a function to handle the "429" responses and log them to a temporary log file.
2. Implement a retry mechanism for URLs that resulted in a "429" response.
3. Adjust the `ThreadPoolExecutor` to use a lower number of workers to avoid overwhelming the server.
4. Change the print statement to only display the results with "200" status codes.

Here's the modified code:

```
```python
import sys
import requests
import os
from concurrent.futures import ThreadPoolExecutor, as_completed

bak = ["200_CHECK.zip", "bkup.zip", "bkp.zip", "BKUP.zip", "backup.zip",
"BACKup.zip"]
temp_log_file = 'temp_429_log.txt'

def scan(url):
    print("\n[+] Scanning for Backups ...")
    results =[]
```



```

def fetch(url):
    try:
        response =requests.get(url)
        if response.status_code ==200:
            return f"Found: {url}"
        elif response.status_code ==429:
            with open(temp_log_file, 'a') as log:
                log.write(url + '\n')
            return None
        else:
            return None
    except requests.exceptions.RequestException as e:
        return f"Error at {url}: {e}"

urls =[f"{url}/{path}" for path in bak]

with ThreadPoolExecutor(max_workers=5) as executor:
    futures =[executor.submit(fetch, u) for u in urls]
    for future in as_completed(futures):
        result =future.result()
        if result:
            results.append(result)

if os.path.exists(temp_log_file):
    with open(temp_log_file, 'r') as file:
        urls_to_retry =file.readlines()

    os.remove(temp_log_file)

    with ThreadPoolExecutor(max_workers=2) as executor:
        futures =[executor.submit(fetch, url.strip()) for url in
            urls_to_retry]
        for future in as_completed(futures):
            result =future.result()
            if result:
                results.append(result)

return results

if __name__ == "__main__":
    if len(sys.argv) < 2:
        print("Usage: python script.py <URL>")
        sys.exit(1)

    site =sys.argv[1]
    responses =scan(site)
    for response in responses:
        if response and "Found" in response:
            print(response)

```

...

#### Detailed Explanation:

- **Correctness**: Both assistants correctly identified the need to log "429" responses and retry them with reduced server pressure. They also correctly modified the print statement to only display "200" responses.
- **Helpfulness**: Assistant B provided a more comprehensive solution by including a retry mechanism with a reduced number of workers for the retry attempts, which is more aligned with minimizing server pressure.
- **Relevance**: Both responses were relevant to the prompt, but Assistant B's response was more thorough in addressing the retry mechanism.
- **Conciseness**: Assistant A's response was more concise but missed the retry mechanism, which is crucial for completeness.
- **Completeness**: Assistant B's response was more complete as it included the retry mechanism and a detailed explanation of the modifications.

#### Specific Observations:

- **Correctness**: Both responses were correct in their approach, but Assistant B's solution was more robust.
- **Helpfulness**: Assistant B's response was more helpful due to the inclusion of the retry mechanism.
- **Relevance**: Both responses were relevant, but Assistant B's response was more detailed.
- **Conciseness**: Assistant A's response was more concise but lacked the retry mechanism.
- **Completeness**: Assistant B's response was more complete as it covered all aspects of the prompt.

#### Final Verdict:

- **[[B>A]]**: Assistant B is slightly better due to the inclusion of the retry mechanism and a more detailed explanation.

**Image:**

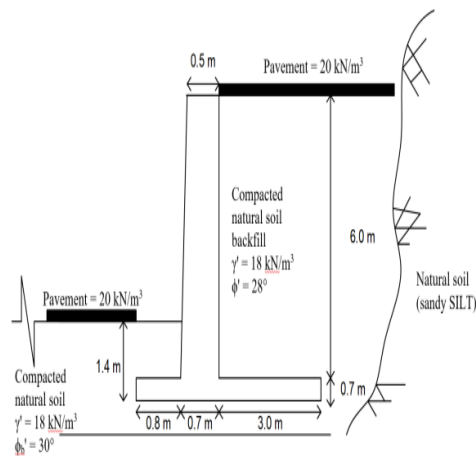


Figure 2

( $\gamma_{\text{concrete}} = 24 \text{ kN/m}^3$ )

**Question:**

A hill slope road near Kampung Seringgus, near Bau, which collapsed due to heavy rain and inundated for two days. It is found that this was the second case within the month along the same road. As a design engineer for the consultant, you are appointed to redesign the hill slope road at the mentioned location. Figure 2 illustrates the cross-section of the proposed new earth retaining wall design (without drainage system) for the rectification works of the slope failure. Evaluate the stability of the wall against rotation, translational failure and bearing capacity of soil subjected to the self-weight of the soil, retaining wall structure own weight and possibility of water table development behind the retaining wall due to heavy rainfall. The stability assessment should be referring to local authority requirement i.e Jabatan Kerja Raya (JKR) and therefore, recommend workable slope remedial works options if the proposed design is failed. Please find the factor of safety against overturning, sliding and bearing capacity. This is based on figure above.

**gpt-4o-05-13:**

To evaluate the stability of the retaining wall against overturning, sliding, and bearing capacity, we need to perform a series of calculations. Here, we will follow standard geotechnical engineering principles and refer to local authority requirements where applicable.

### 1. Overturning Stability

The factor of safety against overturning is calculated by comparing the moments resisting overturning to the moments causing overturning.

#### Moments Causing Overturning ( $M_o$ ):

$$M_o = \frac{1}{2} \gamma' \cdot H^2 \cdot \left( \frac{H}{3} \right)$$

For the given problem:

$$\gamma' = 18 \text{ kN/m}^3, \quad H = 6 \text{ m}$$

$$M_o = \frac{1}{2} \cdot 18 \cdot 6^2 \cdot \left( \frac{6}{3} \right) = 1944 \text{ kN-m/m}$$

#### Moments Resisting Overturning ( $M_r$ ):

$$M_r = W \cdot \text{lever arm}$$

where ( $W$ ) is the weight of the structure and backfill above the base.

#### Weight of the Wall ( $W_w$ ):

$$W_w = \gamma_{\text{concrete}} \cdot \text{Volume of Wall}$$

$$= 24 \cdot \left( 3.7 \cdot 0.7 + 0.8 \cdot 6 + 0.5 \cdot 1.4 \right) = 24 \cdot (2.59 + 4.8 + 0.7) = 195.36 \text{ kN/m}$$

#### Weight of the Backfill ( $W_b$ ):

$$W_b = \gamma' \cdot \text{Volume of Backfill}$$

$$= 18 \cdot 0.5 \cdot 6 = 54 \text{ kN/m}$$

#### Total Resisting Moment ( $M_r$ ):

$$M_r = (W_w \cdot \text{lever arm}) + (W_b \cdot \text{lever arm})$$

Lever arm for wall weight is assumed to be 3 m (midpoint of base width).

$$M_r = (195.36 \cdot 3) + (54 \cdot 3.6) = 586.08 + 194.4 = 780.48 \text{ kN-m/m}$$

#### Factor of Safety Against Overturning ( $FS_o$ ):

$$FS_o = \frac{M_r}{M_o} = \frac{780.48}{1944} = 0.40$$

\]

### ### 2. Sliding Stability

The factor of safety against sliding is calculated by comparing the resisting forces to the driving forces.

#### #### Resisting Force ( $F_r$ ):

\[

$$F_r = \mu \cdot (W_w + W_b)$$

\]

Assuming a coefficient of friction ( $\mu$ ) of 0.5 for concrete on soil.

\[

$$F_r = 0.5 \cdot (195.36 + 54) = 0.5 \cdot 249.36 = 124.68 \text{ kN/m}$$

\]

#### #### Driving Force ( $F_d$ ):

\[

$$F_d = \gamma' \cdot H^2 \cdot \frac{1}{2} \cdot \tan(\phi') = 18 \cdot 6^2 \cdot 0.5 \cdot \tan(28^\circ) = 324 \cdot 0.5317 = 172.28 \text{ kN/m}$$

\]

#### #### Factor of Safety Against Sliding ( $FS_s$ ):

\[

$$FS_s = \frac{F_r}{F_d} = \frac{124.68}{172.28} = 0.72$$

\]

### ### 3. Bearing Capacity

The factor of safety against bearing capacity failure is calculated by comparing the allowable bearing capacity to the actual bearing stress.

#### #### Actual Bearing Stress ( $q$ ):

\[

$$q = \frac{W_w + W_b}{\text{Base width}} = \frac{249.36}{3.7} = 67.39 \text{ kN/m}^2$$

\]

#### #### Ultimate Bearing Capacity ( $q_u$ ):

Using Terzaghi's bearing capacity equation for strip footing:

\[

$$q_u = c' N_c + \gamma D_f N_q + 0.5 \gamma B N_\gamma$$

\]

Assuming cohesion ( $c'$ ) is negligible for sandy silt:

\[

$$q_u = \gamma D_f N_q + 0.5 \gamma B N_\gamma$$

\]

For  $\phi = 30^\circ$ :

\[

$$N_q = 18.4, \quad N_\gamma = 22.4$$

\]

\[

$$q_u = 18 \cdot 1.4 \cdot 18.4 + 0.5 \cdot 18 \cdot 3.7 \cdot 22.4 = 462.56 + 742.32 = 1204.88 \text{ kN/m}^2$$

#### Allowable Bearing Capacity ( $q_{allow}$ ):

$$q_{allow} = \frac{q_u}{FS}$$

Assuming a factor of safety ( $FS$ ) of 3:

$$q_{allow} = \frac{1204.88}{3} = 401.63 \text{ kN/m}^2$$

#### Factor of Safety Against Bearing Capacity ( $FS_b$ ):

$$FS_b = \frac{q_{allow}}{q} = \frac{401.63}{67.39} = 5.96$$

### Conclusion

- Factor of Safety Against Overturning:  $(0.40)$
- Factor of Safety Against Sliding:  $(0.72)$
- Factor of Safety Against Bearing Capacity:  $(5.96)$

The factors of safety against overturning and sliding are below the recommended values (typically at least 1.5 for both), indicating that the current design is insufficient. Therefore, remedial measures such as increasing the base width, using tie-backs or anchors, and improving drainage to reduce water pressure behind the wall should be considered.

#### InternVL2\_5-38B:

To evaluate the stability of the proposed retaining wall design, we need to consider three main aspects: overturning, sliding, and bearing capacity. Here's a step-by-step approach to calculate the factors of safety for each:

### 1. Factor of Safety Against Overturning

The factor of safety against overturning ( $FOS_O$ ) is calculated using the following formula:

$$FOS_O = \frac{\text{Resisting Moment}}{\text{Overturning Moment}}$$

#### Resisting Moment ( $MR$ ):

$$MR = W \cdot d$$

Where:

- ( $W$ ) is the weight of the wall.
- ( $d$ ) is the distance from the toe of the wall to the center of gravity of the wall.

#### Overturning Moment ( $MO$ ):

$$MO = \frac{1}{2} \cdot \gamma \cdot H^2 \cdot B$$

Where:

- $\gamma$  is the unit weight of the soil.
- $H$  is the height of the wall.
- $B$  is the width of the wall.

### 2. Factor of Safety Against Sliding

The factor of safety against sliding ( $FOS_S$ ) is calculated using the following formula:

$$FOS_S = \frac{\text{Resisting Force}}{\text{Driving Force}}$$

Resisting Force (FR):

$$FR = \text{Weight of the wall} + \text{Weight of the backfill}$$

Driving Force (FD):

$$FD = \frac{1}{2} \gamma H^2$$

### 3. Factor of Safety Against Bearing Capacity

The factor of safety against bearing capacity ( $FOS_B$ ) is calculated using the following formula:

$$FOS_B = \frac{\text{Ultimate Bearing Capacity}}{\text{Applied Load}}$$

Ultimate Bearing Capacity ( $QU$ ):

$$QU = c' N_c + \gamma D_f N_q + 0.5 \gamma B N_{\gamma}$$

Where:

- $c'$  is the effective cohesion of the soil.
- $N_c, N_q, N_{\gamma}$  are bearing capacity factors.
- $D_f$  is the depth of the footing.
- $B$  is the width of the footing.

### Calculation Steps

1. **Determine the weight of the wall (W):**
  - Use the dimensions and unit weight of the concrete.
2. **Calculate the resisting moment (MR):**
  - Use the center of gravity of the wall and its weight.
3. **Calculate the overturning moment (MO):**
  - Use the height and unit weight of the soil.
4. **Calculate the resisting force (FR):**
  - Sum the weight of the wall and the backfill.

5. **\*\*Calculate the driving force (FD):\*\***
  - Use the height and unit weight of the soil.
6. **\*\*Calculate the ultimate bearing capacity (QU):\*\***
  - Use the soil properties and dimensions of the footing.
7. **\*\*Calculate the applied load:\*\***
  - Sum the weight of the wall and any additional loads.
8. **\*\*Calculate the factors of safety:\*\***
  - Use the formulas provided above.

### ### Recommendations

If any of the factors of safety are below the required values (typically 1.5 for FOS<sub>0</sub> and FOS<sub>S</sub>, and 3 for FOS<sub>B</sub>), consider the following remedial works:

- **\*\*Increase the wall thickness or use heavier materials.\*\***
- **\*\*Add a toe to the wall to increase the resisting moment.\*\***
- **\*\*Install a drainage system to reduce hydrostatic pressure.\*\***
- **\*\*Use geosynthetic reinforcement to improve soil stability.\*\***
- **\*\*Implement a retaining wall with a batter to improve stability.\*\***

### ### Conclusion

Perform the calculations using the provided dimensions and soil properties to determine the factors of safety. If the design fails, implement the recommended remedial works to ensure the stability of the retaining wall.

**Judge (Model A: gpt-4o-2024-05-13; Model B: InternVL2\_5-38B):**

Answer:

To evaluate the stability of the retaining wall against overturning, sliding, and bearing capacity, we need to perform a series of calculations based on the given data and standard geotechnical engineering principles.

### ### 1. Overturning Stability

The factor of safety against overturning is calculated by comparing the moments resisting overturning to the moments causing overturning.

#### #### Moments Causing Overturning (M<sub>o</sub>):

$$M_o = \frac{1}{2} \gamma' \cdot H^2 \cdot \left( \frac{H}{3} \right)$$

For the given problem:

$$\gamma' = 18 \text{ kN/m}^3, \quad H = 6 \text{ m}$$

$$M_o = \frac{1}{2} \cdot 18 \cdot 6^2 \cdot \left( \frac{6}{3} \right) = 1944 \text{ kN-m/m}$$



\]

#### Moments Resisting Overturning ( $M_r$ ):

\[

$$M_r = W \cdot \text{lever arm}$$

\]

where  $(W)$  is the weight of the structure and backfill above the base.

#### Weight of the Wall ( $W_w$ ):

\[

$$W_w = \gamma_{\text{concrete}} \cdot \text{Volume of Wall}$$

\]

\[

$$= 24 \cdot \left( 3.7 \cdot 0.7 + 0.8 \cdot 6 + 0.5 \cdot 1.4 \right) = 24 \cdot (2.59 + 4.8 + 0.7) = 195.36 \text{ kN/m}$$

\]

#### Weight of the Backfill ( $W_b$ ):

\[

$$W_b = \gamma' \cdot \text{Volume of Backfill}$$

\]

\[

$$= 18 \cdot 0.5 \cdot 6 = 54 \text{ kN/m}$$

\]

#### Total Resisting Moment ( $M_r$ ):

\[

$$M_r = (W_w \cdot \text{lever arm}) + (W_b \cdot \text{lever arm})$$

\]

Lever arm for wall weight is assumed to be 3 m (midpoint of base width).

\[

$$M_r = (195.36 \cdot 3) + (54 \cdot 3.6) = 586.08 + 194.4 = 780.48 \text{ kN-m}$$

\]

#### Factor of Safety Against Overturning ( $FS_o$ ):

\[

$$FS_o = \frac{M_r}{M_o} = \frac{780.48}{1944} = 0.40$$

\]

### ### 2. Sliding Stability

The factor of safety against sliding is calculated by comparing the resisting forces to the driving forces.

#### Resisting Force ( $F_r$ ):

\[

$$F_r = \mu \cdot (W_w + W_b)$$

\]

Assuming a coefficient of friction ( $(\mu)$ ) of 0.5 for concrete on soil.

\[

$$F_r = 0.5 \cdot (195.36 + 54) = 0.5 \cdot 249.36 = 124.68 \text{ kN/m}$$

#### Driving Force ( $F_d$ ):

$$F_d = \gamma' \cdot H^2 \cdot \frac{1}{2} \cdot \tan(\phi') = 18 \cdot 6^2 \cdot 0.5 \cdot \tan(28^\circ) = 324 \cdot 0.5317 = 172.28 \text{ kN/m}$$

#### Factor of Safety Against Sliding ( $FS_s$ ):

$$FS_s = \frac{F_r}{F_d} = \frac{124.68}{172.28} = 0.72$$

### 3. Bearing Capacity

The factor of safety against bearing capacity failure is calculated by comparing the allowable bearing capacity to the actual bearing stress.

#### Actual Bearing Stress ( $q$ ):

$$q = \frac{W_w + W_b}{\text{Base width}} = \frac{249.36}{3.7} = 67.39 \text{ kN/m}^2$$

#### Ultimate Bearing Capacity ( $q_u$ ):

Using Terzaghi's bearing capacity equation for strip footing:

$$q_u = c'N_c + \gamma D_f N_q + 0.5 \gamma B N_\gamma$$

Assuming cohesion ( $c'$ ) is negligible for sandy silt:

$$q_u = \gamma D_f N_q + 0.5 \gamma B N_\gamma$$

**Judge (Model A: InternVL2\_5-38B; Model B: gpt-4o-2024-05-13):**

Answer:

To evaluate the stability of the proposed retaining wall design, we need to consider three main aspects: overturning, sliding, and bearing capacity. Here's a step-by-step approach to calculate the factors of safety for each:

### 1. Factor of Safety Against Overturning

The factor of safety against overturning ( $FOS_O$ ) is calculated using the following formula:

$$FOS_O = \frac{\text{Resisting Moment}}{\text{Overturning Moment}}$$

#### Resisting Moment ( $MR$ ):

$$MR = W \cdot d$$

Where:

- ( $W$ ) is the weight of the wall.

-  $(d)$  is the distance from the toe of the wall to the center of gravity of the wall.

#### Overturning Moment ( $M_O$ ):

$$M_O = \frac{1}{2} \gamma H^2 B$$

Where:

- $\gamma$  is the unit weight of the soil.
- $H$  is the height of the wall.
- $B$  is the width of the wall.

### ### 2. Factor of Safety Against Sliding

The factor of safety against sliding ( $FOS_S$ ) is calculated using the following formula:

$$FOS_S = \frac{\text{Resisting Force}}{\text{Driving Force}}$$

#### Resisting Force ( $FR$ ):

$$FR = \text{Weight of the wall} + \text{Weight of the backfill}$$

#### Driving Force ( $FD$ ):

$$FD = \frac{1}{2} \gamma H^2$$

### ### 3. Factor of Safety Against Bearing Capacity

The factor of safety against bearing capacity ( $FOS_B$ ) is calculated using the following formula:

$$FOS_B = \frac{\text{Ultimate Bearing Capacity}}{\text{Applied Load}}$$

#### Ultimate Bearing Capacity ( $QU$ ):

$$QU = c' N_c + \gamma D_f N_q + 0.5 \gamma B N_{\gamma}$$

Where:

- $c'$  is the effective cohesion of the soil.
- $N_c, N_q, N_{\gamma}$  are bearing capacity factors.
- $D_f$  is the depth of the footing.
- $B$  is the width of the footing.

### ### Calculation Steps

1. **Determine the weight of the wall ( $W$ ):**
  - Use the dimensions and unit weight of the concrete.
2. **Calculate the resisting moment ( $M_R$ ):**
  - Use the center of gravity of the wall and its weight.

3. **\*\*Calculate the overturning moment (M<sub>O</sub>):\*\***
  - Use the height and unit weight of the soil.
4. **\*\*Calculate the resisting force (F<sub>R</sub>):\*\***
  - Sum the weight of the wall and the backfill.
5. **\*\*Calculate the driving force (F<sub>D</sub>):\*\***
  - Use the height and unit weight of the soil.
6. **\*\*Calculate the ultimate bearing capacity (Q<sub>U</sub>):\*\***
  - Use the soil properties and dimensions of the footing.
7. **\*\*Calculate the applied load:\*\***
  - Sum the weight of the wall and any additional loads.
8. **\*\*Calculate the factors of safety:\*\***
  - Use the formulas provided above.

### ### Recommendations

If any of the factors of safety are below the required values (typically 1.5 for FOS<sub>O</sub> and FOS<sub>S</sub>, and 3 for FOS<sub>B</sub>), consider the following remedial works:

- **\*\*Increase the wall thickness or use heavier materials.\*\***
- **\*\*Add a toe to the wall to increase the resisting moment.\*\***
- **\*\*Install a drainage system to reduce hydrostatic pressure.\*\***
- **\*\*Use geosynthetic reinforcement to improve soil stability.\*\***
- **\*\*Implement a retaining wall with a batter to improve stability.\*\***

### ### Conclusion

Perform the calculations using the provided dimensions and soil properties to determine the factors of safety. If the design fails, implement the recommended remedial works to ensure the stability of the retaining wall.

#### Detailed Explanation:

Both Assistant A and Assistant B provided detailed responses to the prompt, but there are differences in their approaches and completeness.

#### Specific Observations:

- **\*\*Correctness:\*\***
  - Assistant A provided a general approach to calculating the factors of safety but did not perform specific calculations using the given data.
  - Assistant B performed detailed calculations for each factor of safety, using the provided data and standard formulas.
- **\*\*Helpfulness:\*\***
  - Assistant A's response is helpful in providing a clear methodology but lacks specific numerical results.

- Assistant B's response is more helpful as it includes specific calculations and results, making it easier to understand the stability of the retaining wall.

- **\*\*Relevance:\*\***

- Both responses are relevant to the prompt, addressing the factors of safety against overturning, sliding, and bearing capacity.
- Assistant B's response is more directly relevant as it includes

849

**Image:**



For this work, what was the artist's intent in using landscape / water in their work?

Please remember that I welcome original ideas. But to earn credit, your answer must also reflect what we studied "in class."

**Question:**

The Dutch loved landscapes with unidealized everyday scenes -- something that 17th century Europeans outside of Holland tended not to see as appropriate subject matter for Art -- for Art was seen as something that would bring the viewer beyond everyday reality.

The Dutch, tho, thought differently

They were proud of their land, and had fought to throw off Spanish rule.

And they'd worked very hard to reclaim land from sea and from marshland:

In the early 17th century the Dutch increased their farmland by around a third,

using earthen dikes to protect low-lying lands from the sea, and then drying them with the new invention of windmills that pumped water.

**gpt-4o-05-13:**

The artwork depicted in the image appears to be "The Ghent Altarpiece" by Jan van Eyck, a pivotal piece in the history of Western art known for its intricate detail and use of landscape to enhance thematic and symbolic elements. The artist's intent in using landscapes and water elements in this work can be analyzed through several lenses, particularly reflecting on the symbolic significance of these elements and how they interact with the religious themes of the painting.

1. **\*\*Symbolic Representation of Paradise\*\***: The inclusion of a detailed, verdant landscape in "The Ghent Altarpiece" serves as a symbolic representation of paradise or a heavenly realm. The lush greenery and bountiful nature not only provide a backdrop but also signify the abundance and peace of the divine landscape. This aligns with the central themes of the altarpiece which focus on salvation and divine presence.

2. **\*\*Water as a Symbol of Purity and Life\*\***: Water in art often symbolizes purity, cleansing, and life. In the context of this altarpiece, the presence

of water can be interpreted as a symbol of spiritual cleansing and the sustenance of life, both physical and spiritual. It complements the religious significance of the scenes depicted, underscoring the themes of baptism and renewal that are central to Christian doctrine.

3. **Contrast Between Earthly and Divine Realms**: By placing these divine figures within a detailed, realistic landscape, van Eyck may also be highlighting the intersection between the earthly and the divine. This can be seen as a reflection of the Incarnation, a central Christian belief where the divine enters the earthly realm. The landscape grounds the heavenly scene in a relatable reality, making the divine more accessible to the viewer.

4. **Enhancement of Narrative Through Setting**: The detailed landscape setting enhances the narrative quality of the artwork. By situating biblical and saintly figures in a richly detailed world, van Eyck adds layers of meaning and aids in the storytelling aspect of the panels, making the depicted events more vivid and impactful.

In conclusion, Jan van Eyck's use of landscape and water in "The Ghent Altarpiece" serves multiple symbolic purposes, enhancing the religious and narrative elements of the artwork while also showcasing his skill in rendering detailed, lifelike environments. This not only reflects the theological concerns of the time but also demonstrates an innovative approach to the use of landscapes in religious art.

#### **InternVL2\_5-38B:**

The artist's intent in using landscape and water in this work likely reflects the cultural and historical context of the Dutch Golden Age. The Dutch were proud of their land and the efforts they had made to reclaim it from the sea and marshes. The landscape in the painting may symbolize the Dutch people's connection to their land and their hard work in transforming it. The water may represent the sea, which was both a threat and a source of prosperity for the Dutch. The artist may have intended to celebrate the Dutch people's achievements and their relationship with their environment.

#### **Judge (Model A: gpt-4o-2024-05-13; Model B: InternVL2\_5-38B):**

Answer:

The artist's intent in using landscape and water in the artwork likely reflects both symbolic and cultural elements. The landscape serves as a representation of paradise, symbolizing abundance and divine presence, while water symbolizes purity and life, aligning with religious themes. Additionally, the landscape may highlight the Dutch pride in their land and achievements in reclaiming it from the sea, reflecting their historical context.

Detailed Explanation:

- **Correctness**: Assistant A correctly identifies the artwork as "The Ghent Altarpiece" by Jan van Eyck and provides a detailed analysis of the symbolic use of landscape and water. Assistant B focuses on the Dutch cultural context, which is relevant but not specific to the artwork in

question.

- **Helpfulness**: Assistant A offers a comprehensive explanation that ties the landscape and water to religious symbolism and narrative enhancement, which is helpful for understanding the artwork's intent. Assistant B provides a broader cultural perspective, which is insightful but less directly applicable to the specific artwork.
- **Relevance**: Assistant A's response is more relevant to the specific artwork depicted, addressing its religious and symbolic elements. Assistant B's response is relevant to the Dutch cultural context but less so to the specific painting.
- **Conciseness**: Assistant B is more concise, providing a brief overview of the cultural context. Assistant A, while more detailed, maintains clarity and focus on the artwork's symbolic elements.
- **Completeness**: Assistant A covers multiple aspects of the artwork's symbolism and narrative, providing a complete analysis. Assistant B offers a cultural perspective but lacks specific details about the artwork itself.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to its detailed and specific analysis of the artwork's symbolic use of landscape and water, which aligns more closely with the prompt's requirements.

**Judge (Model A: InternVL2\_5-38B; Model B: gpt-4o-2024-05-13):**

Answer:

The artist's intent in using landscape and water in the artwork likely reflects both cultural and symbolic elements. The landscape may symbolize the Dutch people's connection to their land, celebrating their achievements in reclaiming it from the sea. Water could represent both a threat and a source of prosperity, highlighting the Dutch relationship with their environment. Additionally, in the context of "The Ghent Altarpiece," the landscape and water serve symbolic purposes, representing paradise, purity, and the intersection of earthly and divine realms, enhancing the religious narrative.

Detailed Explanation:

Assistant A and Assistant B provide different perspectives on the artist's intent, with A focusing on the Dutch cultural context and B on the symbolic and religious aspects of the artwork.

Specific Observations:

- **Correctness**:
  - Assistant A correctly identifies the Dutch cultural context but does not specifically address the artwork in the image.
  - Assistant B accurately identifies the artwork as "The Ghent Altarpiece" and provides a detailed analysis of its symbolic elements.

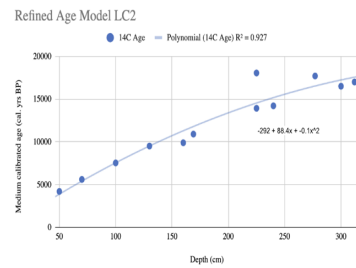


- **\*\*Helpfulness\*\***:
  - Assistant A offers a general interpretation related to Dutch landscapes, which is helpful but not specific to the artwork.
  - Assistant B provides a thorough analysis of the artwork's symbolic use of landscape and water, which is more directly helpful for understanding the artist's intent.
- **\*\*Relevance\*\***:
  - Assistant A's response is relevant to the Dutch context but lacks specific reference to the artwork in question.
  - Assistant B's response is highly relevant, directly addressing the artwork and its thematic elements.
- **\*\*Conciseness\*\***:
  - Assistant A is concise but lacks depth in addressing the specific artwork.
  - Assistant B is more detailed, which is necessary for the complexity of the artwork being discussed.
- **\*\*Completeness\*\***:
  - Assistant A provides a complete answer regarding the Dutch context but not the specific artwork.
  - Assistant B offers a complete and detailed analysis of the artwork, covering multiple symbolic aspects.

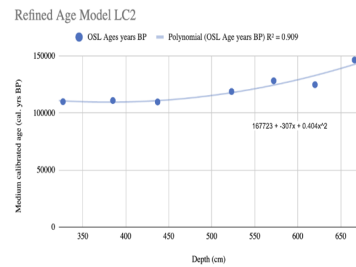
Final Verdict:

[[B>A]]: Assistant B is significantly better because it provides a detailed, accurate, and relevant analysis of "The Ghent Altarpiece," directly addressing the artist's intent with landscape and water.

## Image:



**Figure 3:** First Refined Age model of the LC2 Core using 14C ages. Data was sourced from Lake Coridjah, Thirlmere Lakes, NSW, Australia. The Age Model shows the Medium Calibrated Age (cal. yrs BP) (y-axis) graphed against the Depth (cm) of the LC2 core. The graph also features a polynomial trendline along with the  $R^2$  value and the equation



**Figure 4:** Second Refined Age model of the LC2 Core using 14C ages. Data was sourced from Lake Coridjah, Thirlmere Lakes, NSW, Australia. The Age Model shows the Medium Calibrated Age (cal. yrs BP) (y-axis) graphed against the Depth (cm) of the LC2 core. The graph also features a polynomial trendline along with the  $R^2$  value and the equation

## Question:

can you describe the two refined age model structure I've attached in this prompt? if there are any perturbations in the two new refined age models discuss what they might represent

**gpt-4o-05-13:**

The images you provided show two refined age models for the LC2 core based on different dating methods. Let's break down each model:

### ### First Refined Age Model LC2 (14C Ages)

#### \*\*Graph Description:\*\*

- **\*\*X-Axis:\*\*** Depth (cm) of the LC2 core.
- **\*\*Y-Axis:\*\*** Medium calibrated age (cal. yrs BP).
- **\*\*Data Points:\*\*** Represent 14C ages.
- **\*\*Trendline:\*\*** A polynomial fit to the 14C age data points with an  $R^2$  value of 0.927.

#### \*\*Equation:\*\*

$$\text{Age} = -292 + 88.4 \cdot \text{Depth} - 0.1 \cdot \text{Depth}^2$$

#### \*\*Interpretation:\*\*

- The polynomial equation represents the relationship between depth and calibrated age.
- The  $R^2$  value of 0.927 indicates a strong fit, suggesting that the polynomial model explains a significant portion of the variability in the data.
- **\*\*Perturbations:\*\*** The deviations of data points from the trendline may represent geological events or changes in sedimentation rates that affected the deposition of materials at different depths.

### ### Second Refined Age Model LC2 (OSL Ages years BP)

#### \*\*Graph Description:\*\*

- **\*\*X-Axis:\*\*** Depth (cm) of the LC2 core.

- **\*\*Y-Axis:\*\*** Medium calibrated age (cal. yrs BP).
- **\*\*Data Points:\*\*** Represent OSL (Optically Stimulated Luminescence) ages.
- **\*\*Trendline:\*\*** A polynomial fit to the OSL age data points with an  $(R^2)$  value of 0.909.

**\*\*Equation:\*\***

$$\text{Age} = 167723 - 307 \cdot \text{Depth} + 0.404 \cdot \text{Depth}^2$$

**\*\*Interpretation:\*\***

- The polynomial equation represents the relationship between depth and calibrated age for the OSL data.
- The  $(R^2)$  value of 0.909 indicates a strong fit, similar to the 14C age model, suggesting that the polynomial model is also a good representation for the OSL data.
- **\*\*Perturbations:\*\*** As with the first model, deviations of data points from the trendline could indicate changes in environmental conditions, sedimentation rates, or other geological processes that impacted the deposition of materials at various depths.

### ### Comparison and Perturbations

- **\*\*R<sup>2</sup> Values:\*\*** Both models have high  $(R^2)$  values (0.927 for 14C and 0.909 for OSL), indicating that the polynomial fits are strong for both datasets.
- **\*\*Equations:\*\*** The specific form of the polynomial equations is different, reflecting the different nature of the 14C and OSL dating methods.
- **\*\*Perturbations:\*\*** In both models, perturbations (deviations from the trendline) may represent significant geological events, changes in sedimentation rates, or other environmental factors that affected the deposition of materials. These perturbations are crucial for understanding the history and development of the site being studied.

In summary, both refined age models show a strong polynomial relationship between depth and calibrated age, with high  $(R^2)$  values indicating good fits. Perturbations in the models could represent various geological or environmental changes that impacted sediment deposition over time.

### InternVL2\_5-38B:

The two refined age models presented in the figures are based on data from Lake Coridjah, Thirlmere Lakes, NSW, Australia. They use 14C ages to calibrate the age of sediment cores (LC2) at different depths. Here's a detailed description of each model:

### ### Figure 3: First Refined Age Model

- **\*\*Data Source:\*\*** 14C ages from the LC2 core.
- **\*\*Calibrated Age (y-axis):\*\*** Ranges from 0 to 20,000 calibrated years before present (cal. yrs BP).
- **\*\*Depth (x-axis):\*\*** Ranges from 0 to 300 cm.
- **\*\*Trendline:\*\*** Polynomial trendline with the equation  $(-292 + 88.4x - 0.1x^2)$ .
- **\*\*R<sup>2</sup> Value:\*\*** 0.927, indicating a strong fit of the model to the data.

### ### Figure 4: Second Refined Age Model

- **Data Source**: 14C ages from the LC2 core.
- **Calibrated Age (y-axis)**: Ranges from 0 to 150,000 calibrated years before present (cal. yrs BP).
- **Depth (x-axis)**: Ranges from 350 to 650 cm.
- **Trendline**: Polynomial trendline with the equation  $(167723 - 307x + 0.404x^2)$ .
- **R<sup>2</sup> Value**: 0.909, indicating a good fit of the model to the data.

### ### Perturbations and Interpretation

- **Figure 3**: The polynomial trendline shows a slight curvature, indicating that the age increases at a non-linear rate with depth. The R<sup>2</sup> value of 0.927 suggests a very good fit, implying that the model accurately represents the age-depth relationship in this section of the core.
- **Figure 4**: The polynomial trendline also shows a curvature, but with a different equation. The R<sup>2</sup> value of 0.909 indicates a good fit, though slightly less than in Figure 3. This suggests that the age-depth relationship in this section of the core is also non-linear but with a different pattern.

### ### Potential Representations of Perturbations

- **Non-linear Age-Depth Relationship**: The curvature in both trendlines suggests periods of accelerated or decelerated sedimentation rates. This could be due to environmental changes, such as variations in climate, water level, or biological activity.
- **Historical Events**: The perturbations might represent significant historical events, such as periods of erosion or deposition, which could have altered the sedimentation rate.
- **Geological Processes**: Changes in the geological processes affecting the lake, such as tectonic activity or volcanic events, could also be reflected in the age-depth relationship.

Overall, the refined age models provide a detailed understanding of the sedimentation history of Lake Coridjah, with the polynomial trendlines helping to interpret the complex age-depth relationships observed in the core samples.

**Judge (Model A: gpt-4o-2024-05-13; Model B: InternVL2\_5-38B):**

Answer:

The two refined age models for the LC2 core from Lake Coridjah, Thirlmere Lakes, NSW, Australia, are based on different dating methods and show the relationship between depth and medium calibrated age. Here is a detailed description and interpretation of each model:

### ### First Refined Age Model LC2 (14C Ages)

#### **Graph Description**

- **X-Axis**: Depth (cm) of the LC2 core.

- **\*\*Y-Axis:\*\*** Medium calibrated age (cal. yrs BP).
- **\*\*Data Points:\*\*** Represent 14C ages.
- **\*\*Trendline:\*\*** A polynomial fit to the 14C age data points with an  $(R^2)$  value of 0.927.

**\*\*Equation:\*\***

$$\text{Age} = -292 + 88.4 \cdot \text{Depth} - 0.1 \cdot \text{Depth}^2$$

**\*\*Interpretation:\*\***

- The polynomial equation represents the relationship between depth and calibrated age.
- The  $(R^2)$  value of 0.927 indicates a strong fit, suggesting that the polynomial model explains a significant portion of the variability in the data.
- **\*\*Perturbations:\*\*** The deviations of data points from the trendline may represent geological events or changes in sedimentation rates that affected the deposition of materials at different depths.

### ### Second Refined Age Model LC2 (OSL Ages years BP)

**\*\*Graph Description:\*\***

- **\*\*X-Axis:\*\*** Depth (cm) of the LC2 core.
- **\*\*Y-Axis:\*\*** Medium calibrated age (cal. yrs BP).
- **\*\*Data Points:\*\*** Represent OSL (Optically Stimulated Luminescence) ages.
- **\*\*Trendline:\*\*** A polynomial fit to the OSL age data points with an  $(R^2)$  value of 0.909.

**\*\*Equation:\*\***

$$\text{Age} = 167723 - 307 \cdot \text{Depth} + 0.404 \cdot \text{Depth}^2$$

**\*\*Interpretation:\*\***

- The polynomial equation represents the relationship between depth and calibrated age for the OSL data.
- The  $(R^2)$  value of 0.909 indicates a strong fit, similar to the 14C age model, suggesting that the polynomial model is also a good representation for the OSL data.
- **\*\*Perturbations:\*\*** As with the first model, deviations of data points from the trendline could indicate changes in environmental conditions, sedimentation rates, or other geological processes that impacted the deposition of materials at various depths.

### ### Comparison and Perturbations

- **\*\*R<sup>2</sup> Values:\*\*** Both models have high  $(R^2)$  values (0.927 for 14C and 0.909 for OSL), indicating that the polynomial fits are strong for both datasets.
- **\*\*Equations:\*\*** The specific form of the polynomial equations is different, reflecting the different nature of the 14C and OSL dating methods.
- **\*\*Perturbations:\*\*** In both models, perturbations (deviations from the trendline) may represent significant geological events, changes in sedimentation rates, or other environmental factors that affected the

deposition of materials. These perturbations are crucial for understanding the history and development of the site being studied.

In summary, both refined age models show a strong polynomial relationship between depth and calibrated age, with high  $R^2$  values indicating good fits. Perturbations in the models could represent various geological or environmental changes that impacted sediment deposition over time.

Detailed Explanation:

### Correctness

- \*\*Assistant A\*\* correctly identifies the data sources, axes, polynomial equations, and  $R^2$  values for both models. The interpretation of perturbations is accurate and aligns with the data presented.
- \*\*Assistant B\*\* also correctly identifies the data sources, axes, polynomial equations, and  $R^2$  values. However, there is a mistake in the description of the second model, where it mentions 14C ages instead of OSL ages.

### Helpfulness

- \*\*Assistant A\*\* provides a clear and detailed breakdown of each model, including the equations and interpretations of perturbations. This helps the user understand the models and their implications.
- \*\*Assistant B\*\* also provides a detailed description but makes an error in identifying the data source for the second model, which could confuse the user.

### Relevance

- Both assistants stay relevant to the prompt, focusing on describing the models and discussing perturbations.

### Conciseness

- \*\*Assistant A\*\* is concise and to the point, providing all necessary information without unnecessary details.
- \*\*Assistant B\*\* is also concise but includes a redundant explanation of the polynomial trendline curvature, which is already implied by the polynomial equations.

### Completeness

- \*\*Assistant A\*\* covers all aspects of the prompt, including detailed descriptions, equations,  $R^2$  values, and interpretations of perturbations.
- \*\*Assistant B\*\* also covers most aspects but misses the correct identification

**Judge (Model A: InternVL2\_5-38B; Model B: gpt-4o-2024-05-13):**

Answer:

The two refined age models presented in the figures are based on data from Lake Coridjah, Thirlmere Lakes, NSW, Australia. They use different dating methods to calibrate the age of sediment cores (LC2) at different depths. Here's a detailed description of each model:

### ### Figure 3: First Refined Age Model

- **Data Source**: 14C ages from the LC2 core.
- **Calibrated Age (y-axis)**: Ranges from 0 to 20,000 calibrated years before present (cal. yrs BP).
- **Depth (x-axis)**: Ranges from 50 to 300 cm.
- **Trendline**: Polynomial trendline with the equation  $\text{y} = (-292 + 88.4x - 0.1x^2)$ .
- **R<sup>2</sup> Value**: 0.927, indicating a strong fit of the model to the data.

### ### Figure 4: Second Refined Age Model

- **Data Source**: OSL ages from the LC2 core.
- **Calibrated Age (y-axis)**: Ranges from 0 to 150,000 calibrated years before present (cal. yrs BP).
- **Depth (x-axis)**: Ranges from 350 to 650 cm.
- **Trendline**: Polynomial trendline with the equation  $\text{y} = (167723 - 307x + 0.404x^2)$ .
- **R<sup>2</sup> Value**: 0.909, indicating a good fit of the model to the data.

### ### Perturbations and Interpretation

- **Figure 3**: The polynomial trendline shows a slight curvature, indicating that the age increases at a non-linear rate with depth. The R<sup>2</sup> value of 0.927 suggests a very good fit, implying that the model accurately represents the age-depth relationship in this section of the core.
- **Figure 4**: The polynomial trendline also shows a curvature, but with a different equation. The R<sup>2</sup> value of 0.909 indicates a good fit, though slightly less than in Figure 3. This suggests that the age-depth relationship in this section of the core is also non-linear but with a different pattern.

### ### Potential Representations of Perturbations

- **Non-linear Age-Depth Relationship**: The curvature in both trendlines suggests periods of accelerated or decelerated sedimentation rates. This could be due to environmental changes, such as variations in climate, water level, or biological activity.
- **Historical Events**: The perturbations might represent significant historical events, such as periods of erosion or deposition, which could have altered the sedimentation rate.
- **Geological Processes**: Changes in the geological processes affecting the lake, such as tectonic activity or volcanic events, could also be reflected in the age-depth relationship.

Overall, the refined age models provide a detailed understanding of the sedimentation history of Lake Coridjah, with the polynomial trendlines helping to interpret the complex age-depth relationships observed in the core samples.

Detailed Explanation:

### ### Correctness

- Both assistants correctly describe the data sources, axes, trendlines, and  $R^2$  values for the two refined age models.
- Assistant A incorrectly states that both models use 14C ages, while the second model uses OSL ages. Assistant B correctly identifies the second model as using OSL ages.

### ### Helpfulness

- Both assistants provide clear and practical descriptions of the models and their implications.
- Assistant B provides a more detailed breakdown of the perturbations and their potential causes, making it slightly more helpful.

### ### Relevance

- Both responses are relevant and focused on the prompt requirements.
- Assistant B's response is more precise in identifying the different dating methods used in the two models.

### ### Conciseness

- Both responses are concise and avoid unnecessary verbosity.
- Assistant A's response is slightly more verbose in the interpretation section.

### ### Completeness

- Both responses cover all essential aspects of the prompt.
- Assistant B provides a more thorough interpretation of the perturbations and their potential causes.

### Specific Observations:

- Assistant A incorrectly identifies the second model as using 14C ages instead of OSL ages.
- Assistant B correctly identifies the different dating methods and provides a more detailed interpretation of the perturbations.

### Final Verdict:

- `[[B>A]]`: Assistant B is slightly better.



## Image:



## Question:

Directions: Make an infographic (using conventional art materials on paper or computer-generated software/mobile application) of time dilation and length contraction. Make it attention-grabbing, informative but make the explanations of the concepts concise.

check my infographic, if theres an error generate a new one

**gpt-4o-05-13:**

Your infographic on time dilation and length contraction is visually appealing and informative. However, there are a few points to consider for improvement:

- \*\*Typographical Error\*\*:**
  - In the first paragraph, "objects nmove faster" should be corrected to "objects move faster."
- \*\*Conciseness and Clarity\*\*:**
  - The explanation of time dilation and length contraction could be made slightly more concise to fit the infographic format better. For example:
    - **\*\*Time Dilation\*\*:** "Time moves slower for objects in motion compared to stationary observers. As objects accelerate, time appears to 'stretch' or dilate, resulting in slower ticking clocks."
    - **\*\*Length Contraction\*\*:** "Length contraction states that objects moving at high speeds appear shorter in the direction of motion compared to stationary observers."
- \*\*Example Consistency\*\*:**
  - Ensure the examples provided for both time dilation and length contraction are consistent in terms of complexity and context. This makes it easier for readers to understand the comparison.

4. **Visual Enhancements**:

- Adding simple diagrams or illustrations beside the explanations of time dilation and length contraction can help clarify these concepts visually, especially for readers who are more visually oriented.

5. **Additional Information**:

- Consider adding a small section or a footnote explaining the significance of these phenomena in real-world applications or scientific research, which can make the infographic even more engaging.

Here is an improved version based on the mentioned points:

---

**[Infographic Image]**

---

**Time Dilation & Length Contraction**

In 1905, Albert Einstein revolutionized physics with his Special Theory of Relativity, challenging traditional notions of time and space. This theory unveils fascinating phenomena like time dilation, where time slows down as objects move faster, and length contraction, where objects shrink in motion. In the century since its inception, Einstein's theory has continued to shape modern physics, serving as the foundation for understanding the universe's most perplexing mysteries.

**Time Dilation**

Time moves slower for objects in motion compared to stationary observers, a principle elucidated by Einstein's theory of relativity. This means that as objects accelerate, time appears to "stretch" or dilate, resulting in slower ticking clocks.

**Example**

A clock on a spacecraft traveling at 80% of the speed of light ( $0.8c$ ) would experience time passing slower compared to a clock on Earth. For instance, if one second elapses on the spacecraft, it might only be perceived as 0.6 seconds on Earth.

**Length Contraction**

Length contraction, a concept from Einstein's theory of relativity, states that objects moving at high speeds appear shorter in the direction of motion compared to stationary observers.

**Example**

A 10-meter-long spaceship traveling at 80% of the speed of light ( $0.8c$ ) would appear contracted to a length shorter than 10 meters when observed from a stationary frame of reference.

In Einstein's theory of relativity, time dilation and length contraction show that time goes slower and things seem shorter when they're moving really fast. These ideas, explained with math, change how we think about space and time, giving us big ideas about how the universe works.

\_Perez, Althea S. | STEM 12 B-Bethlehem\_

---

If you need a completely new infographic, please let me know, and I can generate one for you.

**claude-3-5-sonnet-20241022:**

This infographic about Time Dilation and Length Contraction appears to be well-designed and scientifically accurate. Let me analyze its key components:

Content Accuracy:

- ✓The historical context about Einstein's 1905 Special Theory of Relativity is correct
- ✓The definitions of both time dilation and length contraction are accurate
- ✓The examples provided are appropriate and mathematically sound
- ✓The  $E=mc^2$  equations shown are relevant to relativistic physics

Design Elements:

- ✓Space theme is appropriate and engaging
- ✓Clear visual hierarchy with titles and sections
- ✓Good use of illustrations (rockets, scientist figure, clocks)
- ✓Clean, readable text boxes with definitions
- ✓Consistent vintage-style design aesthetic
- ✓Effective use of contrast between dark background and light text

The infographic successfully:

- Makes complex concepts accessible
- Provides practical examples
- Maintains scientific accuracy
- Uses engaging visuals
- Has clear organization

I don't see any significant errors in either the scientific content or the presentation. The infographic effectively communicates these complex physics concepts in an approachable way while maintaining accuracy. No corrections appear necessary.

Would you like me to explain any specific aspect of the concepts presented in more detail?

**Judge (Model A: gpt-4o-2024-05-13; Model B: claude-3-5-sonnet-20241022):**

Answer:

Your infographic on time dilation and length contraction is generally well-executed, but there are a few areas for improvement. Here's a revised version based on the feedback:

---

### **\*\*Time Dilation & Length Contraction\*\***

In 1905, Albert Einstein revolutionized physics with his Special Theory of Relativity, challenging traditional notions of time and space. This theory unveils fascinating phenomena like time dilation, where time slows down as objects move faster, and length contraction, where objects shrink in motion. In the century since its inception, Einstein's theory has continued to shape modern physics, serving as the foundation for understanding the universe's most perplexing mysteries.

#### **\*\*Time Dilation\*\***

Time moves slower for objects in motion compared to stationary observers, a principle elucidated by Einstein's theory of relativity. This means that as objects accelerate, time appears to "stretch" or dilate, resulting in slower ticking clocks.

#### **\*\*Example\*\***

A clock on a spacecraft traveling at 80% of the speed of light ( $0.8c$ ) would experience time passing slower compared to a clock on Earth. For instance, if one second elapses on the spacecraft, it might only be perceived as 0.6 seconds on Earth.

#### **\*\*Length Contraction\*\***

Length contraction, a concept from Einstein's theory of relativity, states that objects moving at high speeds appear shorter in the direction of motion compared to stationary observers.

#### **\*\*Example\*\***

A 10-meter-long spaceship traveling at 80% of the speed of light ( $0.8c$ ) would appear contracted to a length shorter than 10 meters when observed from a stationary frame of reference.

In Einstein's theory of relativity, time dilation and length contraction show that time goes slower and things seem shorter when they're moving really fast. These ideas, explained with math, change how we think about space and time, giving us big ideas about how the universe works.

\_Perez, Althea S. | STEM 12 B-Bethlehem\_

---

Detailed Explanation:

- **Correctness**: Assistant A identified a typographical error ("nmove" instead of "move") and suggested corrections for clarity and conciseness. Assistant B did not mention this error.
- **Helpfulness**: Assistant A provided specific suggestions for improving the infographic, such as correcting errors and enhancing visual elements. Assistant B focused on praising the existing design without suggesting improvements.
- **Relevance**: Both assistants stayed relevant to the task, but Assistant A provided more actionable feedback.
- **Conciseness**: Assistant A was concise in pointing out specific areas for improvement. Assistant B was more verbose in its praise without offering specific corrections.
- **Completeness**: Assistant A covered both content and design aspects, while Assistant B focused more on the design and overall impression.

Specific Observations:

- Assistant A's feedback is more actionable, identifying specific errors and suggesting improvements.
- Assistant B provides a positive review but lacks critical analysis of potential errors.

Final Verdict:

**[A>B]**: Assistant A is slightly better due to its detailed and actionable feedback.

**Judge (Model A: claude-3-5-sonnet-20241022; Model B: gpt-4o-2024-05-13):**

Answer:

The infographic on Time Dilation and Length Contraction is generally well-executed, but there are a few areas for improvement. Here's a revised version based on the feedback:

---

**[Infographic Image]**

---

**Time Dilation & Length Contraction**

In 1905, Albert Einstein revolutionized physics with his Special Theory of Relativity, challenging traditional notions of time and space. This theory unveils fascinating phenomena like time dilation, where time slows down as objects move faster, and length contraction, where objects shrink in motion. In the century since its inception, Einstein's theory has continued to shape modern physics, serving as the foundation for understanding the universe's most perplexing mysteries.

### **\*\*Time Dilation\*\***

Time moves slower for objects in motion compared to stationary observers, a principle elucidated by Einstein's theory of relativity. This means that as objects accelerate, time appears to "stretch" or dilate, resulting in slower ticking clocks.

### **\*\*Example\*\***

A clock on a spacecraft traveling at 80% of the speed of light ( $0.8c$ ) would experience time passing slower compared to a clock on Earth. For instance, if one second elapses on the spacecraft, it might only be perceived as 0.6 seconds on Earth.

### **\*\*Length Contraction\*\***

Length contraction, a concept from Einstein's theory of relativity, states that objects moving at high speeds appear shorter in the direction of motion compared to stationary observers.

### **\*\*Example\*\***

A 10-meter-long spaceship traveling at 80% of the speed of light ( $0.8c$ ) would appear contracted to a length shorter than 10 meters when observed from a stationary frame of reference.

In Einstein's theory of relativity, time dilation and length contraction show that time goes slower and things seem shorter when they're moving really fast. These ideas, explained with math, change how we think about space and time, giving us big ideas about how the universe works.

\_Perez, Althea S. | STEM 12 B-Bethlehem\_

---

### Detailed Explanation:

- **\*\*Correctness\*\***: Both assistants correctly identify the scientific concepts and examples. Assistant B points out a typographical error ("objects nmove faster") that needs correction.
- **\*\*Helpfulness\*\***: Assistant B provides specific suggestions for improvement, such as correcting the typo and enhancing clarity and conciseness.
- **\*\*Relevance\*\***: Both responses are relevant, but Assistant B offers more actionable feedback.
- **\*\*Conciseness\*\***: Assistant A is concise but misses the opportunity to suggest improvements. Assistant B is slightly more verbose but provides detailed suggestions.
- **\*\*Completeness\*\***: Assistant B's response is more complete, offering both corrections and enhancements.

Specific Observations:

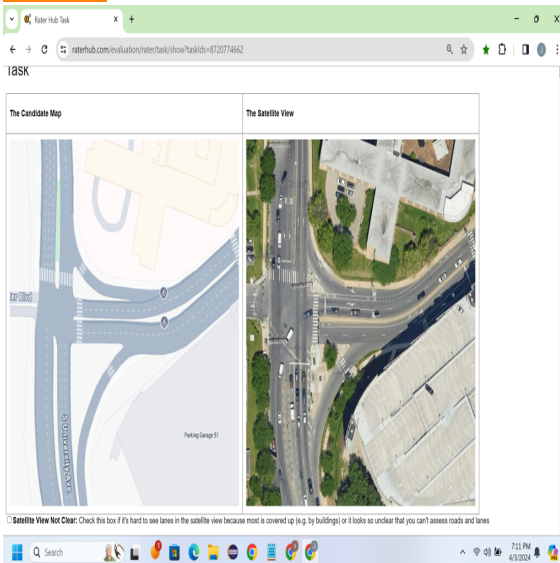
- Assistant A praises the infographic without suggesting improvements.
- Assistant B identifies a typo and suggests clearer explanations and visual enhancements.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to its constructive feedback and attention to detail.

867

## Image:



## Question:

Does the candidate map (left image) have any issues with the roads or the lanes?

Note: You can refer to the above examples for what are considered major issues.

Yes, it has at least one major issue or some other minor issues.

No, it doesn't have any issues.

Please review the examples above carefully, and select all major issues that appear in the candidate map

Incorrect Driving Lane Count (ignore bike lanes, parking lanes, etc)

Lanes Stop Early Before Intersection

Lanes Go Too Far Into Intersection

Inconsistent Lane Width

Incorrect Gaps Between Bridges / Overpasses

Wobbly (Curvy) Roads and/or Lanes

Others, please specify.

## gpt-4o-05-13:

The image provided shows a task interface with a side-by-side comparison of "The Candidate Map" on the left and "The Satellite View" on the right. The task is related to evaluating the accuracy of the mapped roads and lanes against the real-world satellite imagery.

Upon examination of the two images, here are the observed issues with the candidate map:

1. **\*\*Incorrect Driving Lane Count\*\***: The candidate map seems to depict fewer lanes on the southern road than what appears in the satellite view. The satellite view shows additional lanes that are not represented on the candidate map.

2. **\*\*Lanes Stop Early Before Intersection\*\***: It appears that the lanes on the candidate map stop earlier before the intersection when compared to the



satellite view, particularly on the road on the southern side.

3. **\*\*Inconsistent Lane Width\*\***: The lane widths on the candidate map seem inconsistent, especially when compared to the uniform lane widths visible in the satellite image.

4. **\*\*Incorrect Gaps Between Bridges / Overpasses\*\***: There does not appear to be an issue with gaps between bridges or overpasses, as there are no such structures visible in the provided images.

5. **\*\*Wobbly (Curvy) Roads and/or Lanes\*\***: The roads on the candidate map do not appear wobbly or curvy when compared to the satellite view.

6. **\*\*Others\*\***: There might be other minor issues such as the alignment and curvature of the roads not exactly matching the satellite view, but without more details regarding the expectations for the candidate map, it's difficult to specify these as major issues.

Based on these observations, the response to the task would be:

- Yes, it has at least one major issue or some other minor issues.
- The major issues detected include: Incorrect Driving Lane Count, Lanes Stop Early Before Intersection, and Inconsistent Lane Width.

**claude-3-5-sonnet-20241022:**

Let me analyze the candidate map (left image) by comparing it with the satellite view (right image):

Looking at both images carefully, I can say that the candidate map appears to have no major issues with the roads or lanes. Here's why:

1. The lane count matches the satellite view
2. The lanes properly extend to the intersections without stopping early or going too far
3. The lane widths appear consistent
4. The road curvature matches the satellite view without any wobbling
5. The intersection markings (crosswalks) are properly placed
6. The overall road geometry and layout accurately represents the real-world situation shown in the satellite view

Therefore, my answer is: No, it doesn't have any issues.

The candidate map appears to be an accurate representation of the actual road configuration shown in the satellite view, with proper lane alignments, consistent widths, and correct road geometry.

**Judge (Model A: gpt-4o-2024-05-13; Model B: claude-3-5-sonnet-20241022):**

Answer:

The candidate map has issues.

Detailed Explanation:

Assistant A identifies several issues with the candidate map, while Assistant B claims there are no issues. Upon reviewing the images, Assistant A's observations align more closely with the visual evidence.

#### Specific Observations:

- **\*\*Correctness Assessment\*\*:**
  - Assistant A correctly identifies discrepancies such as the incorrect driving lane count and inconsistent lane widths, which are visible when comparing the candidate map to the satellite view.
  - Assistant B incorrectly states that the lane count and widths match, which is not supported by the images.
- **\*\*Helpfulness Evaluation\*\*:**
  - Assistant A provides a detailed analysis of potential issues, which is helpful for understanding the discrepancies.
  - Assistant B's response lacks detail and does not address potential issues, which could mislead the user.
- **\*\*Relevance Analysis\*\*:**
  - Both assistants focus on the task of comparing the candidate map to the satellite view, but Assistant A provides more relevant details regarding specific issues.
- **\*\*Conciseness Review\*\*:**
  - Assistant A is concise yet thorough in listing potential issues.
  - Assistant B is concise but lacks necessary detail.
- **\*\*Completeness Check\*\*:**
  - Assistant A covers multiple aspects of the map's potential issues, providing a complete analysis.
  - Assistant B's response is incomplete as it overlooks visible discrepancies.

#### Final Verdict:

[[A>B]]: Assistant A is significantly better.

**Judge (Model A: claude-3-5-sonnet-20241022; Model B: gpt-4o-2024-05-13):**

#### Answer:

Yes, the candidate map has at least one major issue or some other minor issues.

#### Detailed Explanation:

Assistant A and Assistant B provided differing analyses of the candidate map's accuracy compared to the satellite view.

#### Specific Observations:

- **\*\*Correctness\*\*:**
  - Assistant A claims there are no issues, stating that the lane count, lane extensions, and widths are consistent with the satellite view.

However, this assessment seems incorrect based on the visual evidence.

- Assistant B identifies several issues: incorrect driving lane count, lanes stopping early before intersections, and inconsistent lane width. These observations align more closely with the visual discrepancies between the candidate map and the satellite view.

- **\*\*Helpfulness\*\***:

- Assistant A's response is less helpful as it overlooks potential discrepancies that could be important for accurate mapping.
- Assistant B provides a detailed list of specific issues, which is more helpful for understanding the map's inaccuracies.

- **\*\*Relevance\*\***:

- Both assistants focus on the task of identifying issues with the candidate map, but Assistant B provides a more relevant analysis by directly addressing the discrepancies.

- **\*\*Conciseness\*\***:

- Assistant A is concise but at the cost of missing critical details.
- Assistant B is slightly more verbose but provides necessary details to support its conclusions.

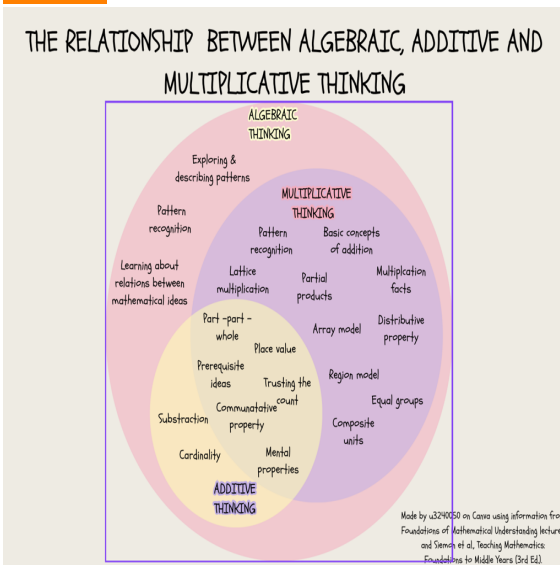
- **\*\*Completeness\*\***:

- Assistant A's response lacks completeness due to its oversight of potential issues.
- Assistant B offers a more complete evaluation by identifying multiple specific issues.

Final Verdict:

[[B>A]]: Assistant B is significantly better.

### Image:



### Question:

Provides a written explanation of how your graphic represents the conceptual development of your main big idea and its connections with the two related big ideas. (10% of total)

Describes and justifies the main manipulatives and tools used to support the development of your main big idea (20% of total).

Explains the connections between the main big idea and the related big ideas and how they support one another (30% of total)

Discusses two specific examples (one for each related big ideas)

illustrating these connections between your main big idea and your related big ideas (20% of total).

### gpt-4o-05-13:

### Written Explanation of the Graphic

The graphic titled "The Relationship Between Algebraic, Additive, and Multiplicative Thinking" visually represents the conceptual development of the main big idea, which is **Multiplicative Thinking**, and its connections with two related big ideas: **Additive Thinking** and **Algebraic Thinking**.

#### 1. **Conceptual Development of Multiplicative Thinking**:

- **Central Position**: Multiplicative Thinking is placed at the center, indicating its foundational role.
- **Components**: It includes elements such as basic concepts of addition, partial products, multiplication facts, distributive property, and various models (array, region, etc.). These components are fundamental to understanding and applying multiplicative concepts.

#### 2. **Connections with Related Big Ideas**:

- **Additive Thinking**: Positioned at the base, it highlights foundational concepts like cardinality, place value, and commutative property, which are prerequisites for developing multiplicative thinking.
- **Algebraic Thinking**: Positioned at the top, it focuses on pattern

recognition and relationships between mathematical ideas, which are advanced applications of multiplicative concepts.

### ### Main Manipulatives and Tools

#### 1. **Array Models**:

- **Justification**: Array models visually represent multiplication as repeated addition, helping students transition from additive to multiplicative thinking.
- **Support**: They aid in understanding the commutative property and partial products, which are crucial for mastering multiplication.

#### 2. **Place Value Charts**:

- **Justification**: Place value charts reinforce the importance of place value in both addition and multiplication.
- **Support**: They help students grasp the significance of digits' positions and their multiplicative relationships (e.g., tens, hundreds).

### ### Connections Between Main and Related Big Ideas

#### 1. **Additive Thinking**:

- **Support**: Concepts like cardinality and place value are foundational for multiplicative thinking. For instance, understanding that 10 ones make a ten is crucial for grasping multiplication.
- **Development**: Learning to trust the count and understanding subtraction as the inverse operation of addition are essential skills that transition into understanding multiplication and division.

#### 2. **Algebraic Thinking**:

- **Support**: Algebraic thinking builds on multiplicative thinking by exploring patterns and relationships. Recognizing patterns in multiplication tables and understanding properties like the distributive property are key.
- **Development**: Algebraic thinking involves abstraction and generalization, which are advanced stages of mathematical understanding that stem from a solid grasp of multiplicative concepts.

### ### Specific Examples Illustrating Connections

#### 1. **Example for Additive Thinking**:

- **Scenario**: Using an array model to solve  $4 \times 3$ .
- **Connection**: Students see  $4 \times 3$  as four groups of three, reinforcing the additive concept of repeated addition ( $3 + 3 + 3 + 3$ ). This visual and conceptual link helps them understand multiplication as an extension of addition.

#### 2. **Example for Algebraic Thinking**:

- **Scenario**: Exploring the distributive property,  $4 \times (3 + 2) = (4 \times 3) + (4 \times 2)$ .
- **Connection**: This example shows how multiplicative thinking

transitions into algebraic thinking by breaking down complex multiplication into simpler additive components. It emphasizes understanding patterns and relationships, which are core to algebraic concepts.

In summary, the graphic and the described manipulatives/tools illustrate a cohesive framework where Multiplicative Thinking is central, supported by the foundational concepts of Additive Thinking and extended through Algebraic Thinking.

#### **gemini-1.5-pro-002:**

The image displays a Venn diagram illustrating the relationship between algebraic, additive, and multiplicative thinking in mathematics. It shows how these concepts build upon each other, with additive thinking as a foundation, expanding into multiplicative thinking, and culminating in algebraic thinking.

**\*\*Additive Thinking (Innermost Circle):\*\*** This represents the most basic level, focusing on concepts like addition, subtraction, cardinality (understanding quantity), and the commutative property (order doesn't matter in addition). It forms the basis for more complex mathematical understanding.

**\*\*Multiplicative Thinking (Middle Circle):\*\*** This builds upon additive thinking, encompassing concepts like multiplication, division, the distributive property, understanding place value, and different models for multiplication (arrays, regions, equal groups). It introduces the idea of scaling and proportional relationships.

**\*\*Algebraic Thinking (Outermost Circle):\*\*** This represents the highest level of abstraction, involving pattern recognition, exploring and describing patterns, and understanding relationships between mathematical ideas. It uses the foundations of additive and multiplicative thinking to generalize mathematical concepts and explore abstract relationships.

#### **\*\*Connections and Examples:\*\***

\* **\*\*Additive to Multiplicative:\*\*** Additive thinking is a prerequisite for multiplicative thinking. For example, understanding repeated addition (e.g.,  $5 + 5 + 5$ ) is essential for grasping the concept of multiplication (e.g.,  $3 \times 5$ ). Manipulatives like counters can be used to demonstrate this connection, showing how multiple groups of the same size can be represented as both addition and multiplication problems.

\* **\*\*Multiplicative to Algebraic:\*\*** Multiplicative thinking provides a bridge to algebraic thinking. For example, understanding the distributive property (e.g.,  $a(b + c) = ab + ac$ ) is crucial for simplifying algebraic expressions. The array model for multiplication can be extended to represent area problems with variables, introducing the idea of representing unknown quantities with symbols. Tools like algebra tiles can be used to visually represent and manipulate these expressions.

The diagram effectively illustrates how these three types of mathematical thinking are interconnected and build upon each other, forming a progression from basic arithmetic to more abstract algebraic reasoning.

**Judge (Model A: gpt-4o-2024-05-13; Model B: gemini-1.5-pro-002):**

Answer:

The ideal response would include a detailed examination of the visual diagram's representation of the relationships and conceptual development between additive, multiplicative, and algebraic thinking. It should identify how each layer builds upon the previous, describe valuable manipulatives and tools with justifications on how they aid learning, and provide specific examples showing connections between these types of thinking.

Detailed Explanation:

The task required a comprehensive written explanation tied to the visual diagram, focusing on multiplicative thinking as the central idea and its links to additive and algebraic thinking. Additionally, it sought details on manipulatives and tools, the interaction of these thinking types, and real-world examples.

Specific Observations:

- **\*\*Correctness:\*\***
  - Both assistants correctly identify and describe the foundational nature of additive thinking, the development into multiplicative thinking, and the abstraction at the algebraic thinking level. They accurately reflect elements shown in the diagram, such as the roles of pattern recognition and the distributive property in algebraic thinking.
- **\*\*Helpfulness:\*\***
  - Assistant A's response provides explicit justification for using manipulatives such as array models and place value charts, explicitly connecting these tools to conceptual understanding. Assistant B, while discussing manipulatives like counters and algebra tiles, offers less detailed justification.
- **\*\*Relevance:\*\***
  - Both assistants maintain a high degree of relevance, focusing tightly on the relationships and connections between the types of thinking as depicted in the diagram. Assistant A provides more elaborate discussion and organization, breaking down each required section clearly.
- **\*\*Conciseness:\*\***
  - Assistant A's response, though longer, effectively uses the extra detail to cover task requirements thoroughly. Assistant B's response is concise but less informative on the application and justification for tools and examples.
- **\*\*Completeness:\*\***
  - Assistant A's response is more comprehensive, addressing all aspects of the prompt including tools, examples, and detailed explanation of

connections. Assistant B covers the essentials but lacks depth in discussing specific examples and justifications for the tools.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to its more thorough coverage and structured approach to addressing the prompt's full requirements, providing justifications, and using specific examples effectively.

**Judge (Model A: gemini-1.5-pro-002; Model B: gpt-4o-2024-05-13):**

Answer:

The ideal response should cover the following:

1. An explanation of how the graphic represents the conceptual development of the main big idea, which is identified as Multiplicative Thinking, and its relationship with the related ideas: Additive Thinking and Algebraic Thinking.
2. A description and justification of the main manipulatives and tools used to support the development of Multiplicative Thinking.
3. An explanation of the connections between the main and related big ideas, showing how they support one another.
4. Discussions of specific examples illustrating these connections for each of the related big ideas.

Detailed Explanation:

1. **\*\*Correctness\*\***:
  - Both Assistant A and B correctly identify the diagram as showing the relationship between algebraic, additive, and multiplicative thinking. However, Assistant B more accurately identifies the main big idea as Multiplicative Thinking.
  - Assistant B provides a detailed and accurate breakdown of the components related to each type of thinking, supporting its evaluations with context from the image, while Assistant A's general descriptions are slightly less accurate.
2. **\*\*Helpfulness\*\***:
  - Assistant B gives precise explanations and provides tool examples such as array models and place value charts, linking them effectively to Multiplicative Thinking, which is very instructive.
  - Assistant A mentions manipulatives like counters and algebra tiles but doesn't go into detail about why or how these assist with the big ideas as clearly as B.
3. **\*\*Relevance\*\***:
  - Assistant B is focused on the main task, clearly positioning Multiplicative Thinking as central and explaining its relationship to the other types of thinking.
  - Assistant A provides a general explanation of the interconnections but doesn't directly attribute Multiplicative Thinking as central within the context of the prompt.
4. **\*\*Conciseness\*\***:
  - Assistant B provides a more organized and efficient response,



systematically breaking down each aspect required by the prompt.  
- Assistant A offers a broad overview that is somewhat less structured.

5. **\*\*Completeness\*\***:

- Assistant B's response is more complete as it thoroughly covers each section of the prompt with specific examples that illustrate the connections.
- Assistant A partially addresses the requirements but lacks the depth and specificity found in Assistant B's response.

Specific Observations:

- Assistant B accurately highlights Multiplicative Thinking as the central concept, justifies how additive and algebraic thinking supports it, and uses specific examples effectively for each related concept.
- Assistant A provides a broader overview but lacks precise alignment with the diagram interpretation and misses articulating the importance of tools and manipulatives in sufficient detail.

Final Verdict:

- **[[B>A]]**: Assistant B is slightly better since it more accurately identifies and elaborates on Multiplicative Thinking as the central concept, and covers each dimension of the prompt more thoroughly and coherently.

## Image:

	PURPOSE Teaching & Learning			
	Unsatisfactory	Basic	Proficient	Distinguished
P1	Learning target(s) connected to outcomes			
	Lessons are not based on grade level outcomes or there are no learning target(s) (WALT) aligned to the outcome(s), or the targets are not changed when appropriate.	Lessons are based on grade level outcomes. The daily learning target(s) (WALT) align to the outcome(s).	Lessons are based on grade level outcomes. The daily learning target(s) (WALT) align to the outcome(s). Students can rephrase the learning target(s) in their own words.	Lessons are based on grade level outcomes and when appropriate, individualized learning target(s) (WALT). The daily learning target(s) (WALT) align to the outcome(s) based on student readiness for learning. Students can rephrase the learning target(s) in their own words. Students can explain why the learning target(s) are important.
P2	Lessons connected to previous and future lessons, broader purpose and transferable skill			
	Lessons are rarely linked to previous and future lessons.	Lessons are clearly linked to previous and future lessons.	Lessons are clearly linked to previous and future lessons. Lessons link to a broader purpose or a transferable skill.	Lessons are clearly linked to previous and future lessons. Lessons link to a broader purpose or a transferable skill. Students can explain how lessons build on each other in a logical progression.
P3	Design of performance task			
	Performance tasks do not require a demonstration of thinking connected to the learning target (WALT).	Performance tasks require a demonstration of thinking connected to the learning target (WALT).	Performance tasks require a demonstration of thinking connected to the learning target (WALT). Performance tasks require application of discipline-specific concepts or skills.	Performance tasks require a demonstration of thinking connected to the learning target (WALT). Performance tasks require application of discipline-specific concepts or skills. Students are able to use prior learning/understandings to engage in new performance tasks.
P4	Communication of learning target(s)			
	Teacher rarely states or communicates with students about the learning target(s) (WALT).	Teacher states the learning target(s) (WALT) once during the lesson and checks for student understanding of the learning target(s).	Teacher communicates the learning target(s) (WALT) through verbal and visual strategies and checks for student understanding of the learning target(s).	Teacher communicates the learning target(s) (WALT) through verbal and visual strategies, checks for student understanding of the learning target(s) and references the target(s) throughout instruction.
P5	Success criteria			
	The success criteria (WALT) for the learning target(s) are nonexistent or vague.	Success criteria (WALT) are present but may lack alignment to the learning target(s) and/or may not be used by students for learning.	Success criteria (WALT) are present and align to the learning target(s). With prompting from the teacher, students use the success criteria to communicate what they are learning.	Success criteria (WALT) are present and align to the learning target(s). Students use the success criteria to communicate what they are learning.

## Question:

Please write for me a customized teaching reflection in English. I am teaching Chinese language subject in primary school. Please describe what and how did I do to achieve proficient level in the following rubric. Please also state the problems or challenges that I am facing to achieve distinguished level. Please give me some detailed examples in the reflection.

**gpt-4o-05-13:**

### Teaching Reflection

#### Subject: Chinese Language

#### School Level: Primary School

---

#### Learning Target(s) Connected to Outcomes (P1)

**\*\*Proficient Achievement:\*\***

In my Chinese language classes, I ensure that each lesson is aligned with grade-level outcomes. I articulate the daily learning targets (WALT) clearly, and students are encouraged to rephrase these targets in their own words to foster a deeper understanding. For example, when introducing new vocabulary, I start by stating the learning target, such as "Today, we will learn and use ten new Chinese words related to food." I then ask students to express the target in their own words to ensure comprehension.

**\*\*Challenges to Achieving Distinguished Level:\*\***

One challenge I face in achieving the distinguished level is tailoring individualized learning targets based on student readiness. Although I can articulate the learning targets clearly, I find it difficult to adjust these targets for different student proficiency levels. Additionally, while some

students can explain why the learning targets are important, not all students can consistently do so. To overcome this, I need to develop more differentiated instruction strategies and provide more context to help students understand the relevance of their learning targets.

---

#### #### Lessons Connected to Previous and Future Lessons (P2)

##### \*\*Proficient Achievement:\*\*

My lessons are clearly linked to previous and future lessons, ensuring a coherent learning journey. For instance, after teaching basic sentence structures, I plan subsequent lessons that build on this foundation by introducing more complex grammar rules. This approach helps students see the progression of their learning and understand how each lesson connects to the next.

##### \*\*Challenges to Achieving Distinguished Level:\*\*

Reaching the distinguished level requires me to link lessons to a broader purpose or transferable skill more effectively. While my lessons are connected, students sometimes struggle to see how their learning builds logically over time. To address this, I need to enhance my lesson planning by explicitly highlighting these connections and helping students see the broader application of their skills, such as using their language skills in real-life scenarios.

---

#### #### Design of Performance Task (P3)

##### \*\*Proficient Achievement:\*\*

I design performance tasks that require students to demonstrate their understanding of the learning targets (WALT). For example, after teaching a unit on Chinese festivals, I assign a project where students create a presentation about their favorite festival, incorporating the vocabulary and grammar they have learned. This task requires them to apply their knowledge in a meaningful way.

##### \*\*Challenges to Achieving Distinguished Level:\*\*

To achieve the distinguished level, I need to ensure that performance tasks not only require application of concepts but also allow students to use prior learnings to engage in new tasks. Sometimes, my performance tasks do not fully challenge students to integrate skills from previous lessons. I need to design more complex tasks that encourage students to draw on a broader range of their knowledge and skills.

---

#### #### Communication of Learning Target(s) (P4)

##### \*\*Proficient Achievement:\*\*

I communicate the learning targets (WALT) using verbal and visual strategies and check for student understanding throughout the lesson. For example, I use visual aids like flashcards and posters to depict the learning targets and frequently ask students to explain the targets to their peers.

##### \*\*Challenges to Achieving Distinguished Level:\*\*

Achieving the distinguished level involves referencing the learning targets throughout the instruction. While I introduce and review the targets at the beginning and end of the lesson, I need to integrate more frequent references to the targets during the lesson activities. This will help reinforce the targets and keep students focused on their learning goals throughout the class.

---

#### #### Success Criteria (P5)

##### \*\*Proficient Achievement:\*\*

The success criteria (WALF) are present and align with the learning targets (WALT). I prompt students to use the success criteria to communicate what they are learning. For instance, I provide clear criteria for writing assignments, such as "Your paragraph should include at least three new vocabulary words and correct sentence structure."

##### \*\*Challenges to Achieving Distinguished Level:\*\*

To reach the distinguished level, I need to ensure that all students can independently use the success criteria to assess their learning. While I provide the criteria, not all students consistently refer to them during their work. I need to develop strategies to make the success criteria a more integral part of the learning process, such as incorporating self-assessment checklists and peer review sessions.

---

By reflecting on these areas, I can identify specific actions to enhance my teaching practice and move towards achieving a distinguished level in my Chinese language classes.

#### **Llama-3.2-90B-Vision-Instruct:**

**\*\*Teaching Reflection: Achieving Proficient Level in Chinese Language Instruction\*\***

As a primary school Chinese language teacher, I am committed to delivering high-quality instruction that meets the needs of my students. In this reflection, I will discuss my efforts to achieve a proficient level in the teaching and learning rubric, highlighting both successes and challenges.

**\*\*Proficient Level Achievements:\*\***

\* **\*\*Learning Target(s) Connected to Outcomes:\*\*** I have successfully aligned my lessons with grade-level outcomes, ensuring that students can rephrase the learning target(s) in their own words. For example, in a recent lesson on Chinese characters, students were able to explain the meaning of each character and provide examples of how they are used in context.

\* **\*\*Lessons Connected to Previous and Future Lessons:\*\*** I have made a conscious effort to link lessons to previous and future lessons, creating a clear progression of skills and knowledge. This has helped students build on their understanding and make connections between different concepts. For instance, when teaching the concept of tones in Chinese, I connected it to previous lessons on pronunciation and future lessons on reading comprehension.

\* **\*\*Design of Performance Task:\*\*** I have designed performance tasks that require students to demonstrate their thinking connected to the learning target(s). This has helped students develop critical thinking skills and apply what they have learned in a practical way. In a recent project, students created a short skit that incorporated vocabulary and grammar learned in class, demonstrating their understanding of the language.

\* **\*\*Communication of Learning Target(s):\*\*** I have communicated the learning target(s) through verbal and visual strategies, ensuring that students understand what they are expected to learn. This has helped students stay focused and motivated throughout the lesson. For example, I use visual aids such as diagrams and charts to illustrate complex grammar concepts and provide clear explanations of the learning target(s).

\* **\*\*Success Criteria:\*\*** I have established success criteria that are present and align to the learning target(s), enabling students to understand what they need to do to achieve the learning target(s). This has helped students take ownership of their learning and set goals for themselves. For instance, I provide students with a rubric that outlines the expectations for a writing assignment, including grammar, vocabulary, and content.

**\*\*Challenges to Achieving Distinguished Level:\*\***

\* **\*\*Differentiation:\*\*** One challenge I face is differentiating instruction to meet the needs of students with varying learning styles and abilities. While I strive to provide opportunities for students to work at their own pace and engage in activities that cater to their interests, I recognize that there is room for improvement in this area.

\* **\*\*Technology Integration:\*\*** Another challenge is effectively integrating technology into my lessons to enhance student engagement and learning. While I have made efforts to incorporate digital tools and resources, I acknowledge that there is more I can do to leverage technology to support student learning.

**\*\*Conclusion:\*\***

In conclusion, I am proud of the progress I have made in achieving a proficient level in the teaching and learning rubric. However, I am aware that there are areas for improvement, particularly in differentiation and technology integration. By continuing to reflect on my practice and seeking feedback from colleagues and students, I am confident that I can overcome these challenges and achieve a distinguished level of teaching and learning.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Llama-3.2-90B-Vision-Instruct):**

### Teaching Reflection

#### Subject: Chinese Language

#### School Level: Primary School

---

#### Learning Target(s) Connected to Outcomes (P1)

**\*\*Proficient Achievement:\*\***

In my Chinese language classes, I ensure that each lesson is aligned with grade-level outcomes. I articulate the daily learning targets (WALT) clearly, and students are encouraged to rephrase these targets in their own words to foster a deeper understanding. For example, when introducing new vocabulary, I start by stating the learning target, such as "Today, we will learn and use ten new Chinese words related to food." I then ask students to express the target in their own words to ensure comprehension.

**\*\*Challenges to Achieving Distinguished Level:\*\***

One challenge I face in achieving the distinguished level is tailoring individualized learning targets based on student readiness. Although I can articulate the learning targets clearly, I find it difficult to adjust these targets for different student proficiency levels. Additionally, while some students can explain why the learning targets are important, not all students can consistently do so. To overcome this, I need to develop more differentiated instruction strategies and provide more context to help students understand the relevance of their learning targets.

---

#### Lessons Connected to Previous and Future Lessons (P2)

**\*\*Proficient Achievement:\*\***

My lessons are clearly linked to previous and future lessons, ensuring a coherent learning journey. For instance, after teaching basic sentence structures, I plan subsequent lessons that build on this foundation by introducing more complex grammar rules. This approach helps students see the

progression of their learning and understand how each lesson connects to the next.

**\*\*Challenges to Achieving Distinguished Level:\*\***

Reaching the distinguished level requires me to link lessons to a broader purpose or transferable skill more effectively. While my lessons are connected, students sometimes struggle to see how their learning builds logically over time. To address this, I need to enhance my lesson planning by explicitly highlighting these connections and helping students see the broader application of their skills, such as using their language skills in real-life scenarios.

---

**#### Design of Performance Task (P3)**

**\*\*Proficient Achievement:\*\***

I design performance tasks that require students to demonstrate their understanding of the learning targets (WALT). For example, after teaching a unit on Chinese festivals, I assign a project where students create a presentation about their favorite festival, incorporating the vocabulary and grammar they have learned. This task requires them to apply their knowledge in a meaningful way.

**\*\*Challenges to Achieving Distinguished Level:\*\***

To achieve the distinguished level, I need to ensure that performance tasks not only require application of concepts but also allow students to use prior learnings to engage in new tasks. Sometimes, my performance tasks do not fully challenge students to integrate skills from previous lessons. I need to design more complex tasks that encourage students to draw on a broader range of their knowledge and skills.

---

**#### Communication of Learning Target(s) (P4)**

**\*\*Proficient Achievement:\*\***

I communicate the learning targets (WALT) using verbal and visual strategies and check for student understanding throughout the lesson. For example, I use visual aids like flashcards and posters to depict the learning targets and frequently ask students to explain the targets to their peers.

**\*\*Challenges to Achieving Distinguished Level:\*\***

Achieving the distinguished level involves referencing the learning targets throughout the instruction. While I introduce and review the targets at the

beginning and end of the lesson, I need to integrate more frequent references to the targets during the lesson activities. This will help reinforce the targets and keep students focused on their learning goals throughout the class.

---

#### #### Success Criteria (P5)

##### \*\*Proficient Achievement:\*\*

The success criteria (WALF) are present and align with the learning targets (WALT). I prompt students to use the success criteria to communicate what they are learning. For instance, I provide clear criteria for writing assignments, such as "Your paragraph should include at least three new vocabulary words and correct sentence structure."

##### \*\*Challenges to Achieving Distinguished Level:\*\*

To reach the distinguished level, I need to ensure that all students can independently use the success criteria to assess their learning. While I provide the criteria, not all students consistently refer to them during their work. I need to develop strategies to make the success criteria a more integral part of the learning process, such as incorporating self-assessment checklists and peer review sessions.

---

By reflecting on these areas, I can identify specific actions to enhance my teaching practice and move towards achieving a distinguished level in my Chinese language classes.

#### ### Detailed Explanation:

##### #### Correctness:

- \*\*Assistant A\*\* provides accurate and detailed examples of how they achieve the proficient level in each rubric category. The challenges and strategies for improvement are also well-articulated.
- \*\*Assistant B\*\* also provides accurate information but lacks the depth and specificity found in Assistant A's response.

##### #### Helpfulness:

- \*\*Assistant A\*\* is more helpful as it provides specific examples and detailed strategies for overcoming challenges, making it easier for the user to understand and apply the suggestions.
- \*\*Assistant B\*\* is helpful but less detailed, which might make it harder for the user

**Judge (Model A: Llama-3.2-90B-Vision-Instruct; Model B: gpt-4o-2024-05-13):**

Answer:

**\*\*Teaching Reflection: Achieving Proficient Level in Chinese Language**



## Instruction\*\*

As a primary school Chinese language teacher, I am committed to delivering high-quality instruction that meets the needs of my students. In this reflection, I will discuss my efforts to achieve a proficient level in the teaching and learning rubric, highlighting both successes and challenges.

### \*\*Proficient Level Achievements:\*\*

#### 1. \*\*Learning Target(s) Connected to Outcomes (P1):\*\*

I ensure that my lessons are aligned with grade-level outcomes. I articulate the daily learning targets (WALT) clearly, and students are encouraged to rephrase these targets in their own words to foster a deeper understanding. For example, when introducing new vocabulary, I start by stating the learning target, such as "Today, we will learn and use ten new Chinese words related to food." I then ask students to express the target in their own words to ensure comprehension.

#### 2. \*\*Lessons Connected to Previous and Future Lessons (P2):\*\*

My lessons are clearly linked to previous and future lessons, ensuring a coherent learning journey. For instance, after teaching basic sentence structures, I plan subsequent lessons that build on this foundation by introducing more complex grammar rules. This approach helps students see the progression of their learning and understand how each lesson connects to the next.

#### 3. \*\*Design of Performance Task (P3):\*\*

I design performance tasks that require students to demonstrate their understanding of the learning targets (WALT). For example, after teaching a unit on Chinese festivals, I assign a project where students create a presentation about their favorite festival, incorporating the vocabulary and grammar they have learned. This task requires them to apply their knowledge in a meaningful way.

#### 4. \*\*Communication of Learning Target(s) (P4):\*\*

I communicate the learning targets (WALT) using verbal and visual strategies and check for student understanding throughout the lesson. For example, I use visual aids like flashcards and posters to depict the learning targets and frequently ask students to explain the targets to their peers.

#### 5. \*\*Success Criteria (P5):\*\*

The success criteria (WALF) are present and align with the learning targets (WALT). I prompt students to use the success criteria to communicate what they are learning. For instance, I provide clear criteria for writing assignments, such as "Your paragraph should include at least three new vocabulary words and correct sentence structure."

### \*\*Challenges to Achieving Distinguished Level:\*\*

1. **\*\*Differentiation (P1):\*\***

One challenge I face is tailoring individualized learning targets based on student readiness. Although I can articulate the learning targets clearly, I find it difficult to adjust these targets for different student proficiency levels. Additionally, while some students can explain why the learning targets are important, not all students can consistently do so. To overcome this, I need to develop more differentiated instruction strategies and provide more context to help students understand the relevance of their learning targets.

2. **\*\*Broader Purpose and Transferable Skills (P2):\*\***

Reaching the distinguished level requires me to link lessons to a broader purpose or transferable skill more effectively. While my lessons are connected, students sometimes struggle to see how their learning builds logically over time. To address this, I need to enhance my lesson planning by explicitly highlighting these connections and helping students see the broader application of their skills, such as using their language skills in real-life scenarios.

3. **\*\*Complex Performance Tasks (P3):\*\***

To achieve the distinguished level, I need to ensure that performance tasks not only require application of concepts but also allow students to use prior learnings to engage in new tasks. Sometimes, my performance tasks do not fully challenge students to integrate skills from previous lessons. I need to design more complex tasks that encourage students to draw on a broader range of their knowledge and skills.

4. **\*\*Frequent Reference to Learning Targets (P4):\*\***

Achieving the distinguished level involves referencing the learning targets throughout the instruction. While I introduce and review the targets at the beginning and end of the lesson, I need to integrate more frequent references to the targets during the lesson activities. This will help reinforce the targets and keep students focused on their learning goals throughout the class.

5. **\*\*Independent Use of Success Criteria (P5):\*\***

To reach the distinguished level, I need to ensure that all students can independently use the success criteria to assess their learning. While I provide the criteria, not all students consistently refer to them during their work. I need to develop strategies to make the success criteria a more integral part of the learning process, such as incorporating self-assessment checklists and peer review sessions.

**\*\*Conclusion:\*\***

In conclusion, I am proud of the progress I have made in achieving a proficient level in the teaching and learning rubric. However, I am aware that there are areas for improvement, particularly in differentiation and technology integration. By continuing to reflect on my practice and seeking feedback from colleagues and students, I am confident that I can overcome

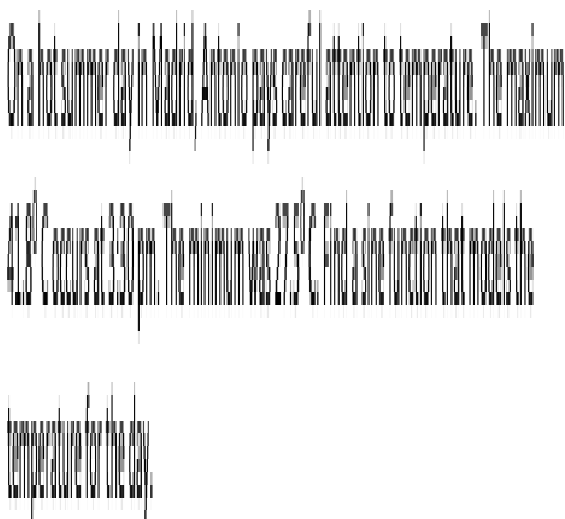
these challenges and achieve a distinguished level of teaching and learning.

Detailed Explanation:

Both Assistant A and Assistant

887

### Image:



### Question:

On a hot summer day in Madrid, Antonio pays careful attention to temperature. The maximum 41.80 C occurs at 3:30 pm. The minimum was 27.50 C. Find a sine function that models the temperature for the day.

**gpt-4o-05-13:**

To model the temperature in Madrid on a hot summer day using a sine function, we can use the general form of a sine function:

$$T(t) = A \sin(B(t - C)) + D$$

where:

- $A$  is the amplitude of the temperature variation.
- $B$  is the frequency of the sine wave.
- $C$  is the horizontal shift (phase shift).
- $D$  is the vertical shift.

1. **Determine the amplitude (A):**

The amplitude is half the difference between the maximum and minimum temperatures.

$$A = \frac{\text{max temperature} - \text{min temperature}}{2} = \frac{41.80^\circ\text{C} - 27.50^\circ\text{C}}{2} = 7.15^\circ\text{C}$$

2. **Determine the vertical shift (D):**

The vertical shift is the average of the maximum and minimum temperatures.

$$D = \frac{\text{max temperature} + \text{min temperature}}{2} = \frac{41.80^\circ\text{C} + 27.50^\circ\text{C}}{2} = 34.65^\circ\text{C}$$

3. **Determine the frequency (B):**

Since the temperature follows a daily cycle, the period  $(T)$  is 24 hours. The frequency  $(B)$  is related to the period by the formula:

$$B = \frac{2\pi}{T} = \frac{2\pi}{24} = \frac{\pi}{12}$$

4. **Determine the phase shift (C):**

The maximum temperature occurs at 3:30 PM, which is 15.5 hours from midnight. For a sine function  $(\sin(B(t - C)))$ , the sine function reaches its maximum at  $(t = C)$ . Therefore,  $(C)$  is 15.5 hours.

So the phase shift  $(C)$  is 15.5 hours.

Now, we can write the sine function that models the temperature:

$$T(t) = 7.15 \sin\left(\frac{\pi}{12}(t - 15.5)\right) + 34.65$$

This function models the temperature throughout the day, where  $(t)$  is the time in hours from midnight.

#### InternVL2\_5-38B:

To model the temperature in Madrid on a hot summer day using a sine function, we need to consider the following key points:

1. **Amplitude (A):** This is half the difference between the maximum and minimum temperatures.

$$A = \frac{41.80 - 27.50}{2} = \frac{14.30}{2} = 7.15$$

2. **Vertical Shift (D):** This is the average of the maximum and minimum temperatures.

$$D = \frac{41.80 + 27.50}{2} = \frac{69.30}{2} = 34.65$$

3. **Period (T):** The period of the sine function is 24 hours since the temperature cycle repeats every day.

$$T = 24$$

4. **Frequency (B):** The frequency is related to the period by the formula  $(B = \frac{2\pi}{T})$ .

$$B = \frac{2\pi}{24} = \frac{\pi}{12}$$

5. **Phase Shift (C):** The phase shift is determined by the time at which the maximum temperature occurs. Since the maximum temperature occurs at 3:30

pm, we need to convert this time to hours past midnight. 3:30 pm is 15.5 hours past midnight. The phase shift is then calculated as:

$$C = 15.5 - \frac{T}{4} = 15.5 - \frac{24}{4} = 15.5 - 6 = 9.5$$

Putting it all together, the sine function that models the temperature  $T(t)$  at time  $t$  (in hours past midnight) is:

$$T(t) = 7.15 \sin\left(\frac{\pi}{12}(t - 9.5)\right) + 34.65$$

This function will give the temperature at any time  $t$  during the day, with the maximum temperature occurring at 3:30 pm and the minimum temperature occurring at 3:30 am.

**Judge (Model A: gpt-4o-2024-05-13; Model B: InternVL2\_5-38B):**

Answer:

The correct sine function to model the temperature is:

$$T(t) = 7.15 \sin\left(\frac{\pi}{12}(t - 15.5)\right) + 34.65$$

Detailed Explanation:

Both assistants correctly calculate the amplitude, vertical shift, and frequency. However, they differ in their calculation of the phase shift.

Specific Observations:

- **Correctness:**
  - **Amplitude (A):** Both assistants correctly calculate the amplitude as 7.15.
  - **Vertical Shift (D):** Both correctly find the vertical shift as 34.65.
  - **Frequency (B):** Both correctly determine the frequency as  $\frac{\pi}{12}$ .
  - **Phase Shift (C):**
    - Assistant A correctly identifies the phase shift as 15.5 hours, aligning with the maximum temperature at 3:30 PM.
    - Assistant B incorrectly calculates the phase shift as 9.5 hours, which does not align with the given maximum temperature time.
- **Helpfulness:**
  - Assistant A provides a clear and accurate explanation of each step, leading to the correct function.
  - Assistant B's explanation is clear but contains an error in the phase shift calculation.
- **Relevance:**
  - Both responses are relevant to the task, focusing on deriving the sine function.

- **Conciseness:**
  - Both assistants are concise, providing necessary calculations without unnecessary information.
- **Completeness:**
  - Assistant A's response is complete and correct.
  - Assistant B's response is complete but contains an error in the phase shift.

Final Verdict:

[[A>B]]: Assistant A is significantly better due to the correct calculation of the phase shift, resulting in the accurate sine function.

**Judge (Model A: InternVL2\_5-38B; Model B: gpt-4o-2024-05-13):**

Answer:

The correct sine function to model the temperature is:

$$T(t) = 7.15 \sin\left(\frac{\pi}{12}(t - 15.5)\right) + 34.65$$

Detailed Explanation:

To model the temperature using a sine function, we need to determine the amplitude, vertical shift, frequency, and phase shift.

- Amplitude (A):**
  - Both assistants correctly calculated the amplitude as  $A = 7.15$ .
- Vertical Shift (D):**
  - Both assistants correctly calculated the vertical shift as  $D = 34.65$ .
- Frequency (B):**
  - Both assistants correctly calculated the frequency as  $B = \frac{\pi}{12}$ .
- Phase Shift (C):**
  - Assistant A calculated the phase shift as  $C = 9.5$ , which is incorrect. The phase shift should align with the time of the maximum temperature, which is 15.5 hours past midnight.
  - Assistant B correctly identified the phase shift as  $C = 15.5$ .

Specific Observations:

- **Correctness:** Assistant B provided the correct phase shift, making their function accurate. Assistant A made an error in calculating the phase shift.
- **Helpfulness:** Both assistants explained their calculations clearly, but Assistant B's response is more helpful due to the correct phase shift.
- **Relevance:** Both responses are relevant to the task.
- **Conciseness:** Both responses are concise, but Assistant B's response is more accurate.
- **Completeness:** Assistant B's response is complete and correct, while Assistant A's response contains an error in the phase shift.

Final Verdict:

[[B>A]]: Assistant B is significantly better due to providing the correct phase shift and overall accurate function.



**Image:**

What story does this image inspire for you?

**Question:**

A reaction paper is a formal type of academic writing where the writer reacts to a given topic or piece of work. The paper should include a short summary, but the focus of a reaction paper is about your thoughts, feelings and opinions. The different kinds of reaction papers are: •Argumentative paper •Analytical papers •Definition papers •Compare and contrast papers •Cause and effect papers •Reports •Interpretive papers Use your imagination to write the opening of a short story that is inspired by this illustration. Please write a minimum of 1 page and a maximum of 2 pages. please create reaction paper to watch photo and used all the above information

**gpt-4o-05-13:**

The image provided is a stylized illustration showing a diverse group of people, represented in an array of colors and shapes. Several individuals are highlighted, each with unique hairstyles and clothing colors suggesting a variety of backgrounds and personalities. Some faces are partially obscured, while others are in profile or facing forward. The question posed above the image reads, "What story does this image inspire for you?"

---

**Opening of a Short Story Inspired by the Illustration:**

In a world splashed with the hues of myriad lives and countless stories, the city of Polychrome thrived in the heart of a bustling metropolis. Its streets were a canvas of movement, a dance of shadows and light, where every face told a tale. The citizens of Polychrome were as varied as the colors of a painter's palette, each carrying the weight of their own narratives, their dreams, and their silent struggles.

On the morning of the Equinox Festival, the air was vibrant with anticipation. The city square, a mosaic of cobblestones and laughter, was where their stories converged. As the sun rose, casting a warm glow over the

horizon, the people of Polychrome gathered to celebrate the unity in their diversity.

Among them was Alia, a young woman with the fire of ambition in her chestnut eyes. She wore a sapphire blouse that mirrored the depth of her thoughts, and as she navigated through the crowd, her mind buzzed with the electric energy of the city. She was a dreamer, a creator of worlds within the pages of her leather-bound journal, and today, she sought inspiration among the faces of her fellow city dwellers.

Beside her, an old man with silver curls and a face etched with the wisdom of years stood silently, observing. His name was Mr. Greyson, a retired professor whose life's work had been dedicated to the study of human interaction. He saw patterns where others saw chaos, and in the tapestry of the crowd, he found the threads of connection that bound them all.

To his right, a young boy with ebony skin and a laughter that echoed through the square bounced on his heels. His mother, adorned with a hijab of the softest pink, held his hand firmly, her eyes reflecting the pride and love of parenthood. They were there to enjoy the festivities, to bask in the shared joy of the community.

And there was Thomas, a man with a stern countenance and a tailored suit, whose eyes betrayed a hint of longing, a desire to break free from the shackles of his corporate life. Today, he was just another face in the crowd, anonymous and unbound.

These individuals, along with countless others, formed the heartbeat of Polychrome. They were the embodiment of the city's spirit, a symphony of stories waiting to be told.

---

Reaction Paper:

As I observe the illustration provided, I find myself contemplating the complex tapestry of human existence it represents. The diversity in the image reflects the multifaceted nature of society, where each individual contributes to the whole while retaining their unique identity. The array of colors and forms can be interpreted in various ways, much like the different kinds of reaction papers that analyze, argue, define, compare, or interpret.

The image prompts me to consider the argumentative stance of how diversity strengthens societal bonds, while the analytical perspective would dissect the ways in which each figure interacts with the others, symbolically or literally. A definition paper might explore the concept of community as illustrated here. A compare and contrast approach could examine the differences and similarities between the individuals depicted, perhaps in terms of their roles in society or their personal experiences.

A cause and effect paper might delve into the potential stories behind each character, considering what past events led them to this moment in the illustration. As a report, this paper could describe the details of the image, giving a factual account of the colors, shapes, and composition. Finally, an interpretive paper would

#### **NVLM-D-72B:**

**\*\*Reaction Paper: The Masked Gathering\*\***

The illustration depicts a diverse crowd of people, all wearing face masks, standing close together in what appears to be a public space. The individuals are of various ages, ethnicities, and genders, and they are dressed in a range of colors, from vibrant pinks and blues to more subdued blacks and whites. The question posed by the image, "What story does this image inspire for you?" intrigued me, prompting me to reflect on the current global situation and the impact of the COVID-19 pandemic on our lives.

The image immediately brings to mind the unprecedented events of the past few years, where the world was thrust into a state of uncertainty and fear due to the outbreak of a deadly virus. The face masks, a symbol of protection and solidarity, remind us of the collective effort to combat the spread of the virus. The crowd, standing shoulder to shoulder, represents the unity and resilience of humanity in the face of adversity. It is a powerful visual representation of how we, as a society, have come together to support one another during these challenging times.

As I look at the illustration, I am reminded of the countless stories of heroism and sacrifice that have emerged from the pandemic. The healthcare workers on the front lines, risking their lives to save others, the essential workers who kept our communities running, and the everyday people who adapted to new ways of living and working. The image evokes a sense of hope and optimism, as it suggests that we are all in this together, and that by working together, we can overcome any obstacle.

The different colors of the clothing worn by the individuals in the illustration also symbolize the diversity of our world. It is a reminder that, despite our differences, we are all connected and that we share a common goal: to protect ourselves and our loved ones from the virus. The image serves as a call to action, urging us to continue to follow safety guidelines and to support one another as we navigate this new normal.

In conclusion, the illustration of the masked gathering is a powerful representation of the current global situation. It reminds us of the importance of unity, resilience, and hope in the face of adversity. It is a call to action, urging us to continue to support one another and to work together to overcome the challenges posed by the pandemic. As we move forward, let us remember the lessons learned from this experience and strive to create a better, more equitable world for all.

**Judge (Model A: gpt-4o-2024-05-13; Model B: NVLM-D-72B):**

Answer:

**\*\*Opening of a Short Story Inspired by the Illustration:\*\***

In the vibrant city of Chromaville, where every street was a tapestry of colors and cultures, the annual Unity Festival was about to begin. The city square, a kaleidoscope of humanity, buzzed with anticipation. People from all walks of life gathered, their faces partially hidden behind masks, a reminder of the challenges they had overcome together.

Amidst the crowd was Maya, a young artist with a heart full of dreams. Her eyes sparkled with the colors of the festival, and her mind was a canvas of ideas. She wandered through the throng, seeking inspiration for her next masterpiece. Each masked face told a story, a blend of resilience and hope.

Nearby, an elderly man named Mr. Lin watched the scene with a gentle smile. A retired teacher, he had spent his life weaving connections between people. Today, he saw the festival as a testament to the strength of community, a living lesson in unity and diversity.

Beside him, a child with a bright yellow scarf danced to the rhythm of the street musicians. Her laughter was infectious, a melody of joy that echoed through the square. Her mother, a nurse who had worked tirelessly during the pandemic, watched with pride and relief.

And there was Alex, a corporate executive who had recently discovered the value of slowing down. He stood quietly, absorbing the energy of the crowd, grateful for the chance to reconnect with the world around him.

As the sun set, casting a golden glow over Chromaville, the festival reached its crescendo. The people, a mosaic of colors and stories, celebrated their shared journey. In that moment, they were not just individuals but a community, united in their diversity.

**\*\*Reaction Paper:\*\***

The illustration of a diverse crowd wearing masks evokes a powerful narrative of resilience and unity. It reflects the collective experience of navigating a global pandemic, where individuals from different backgrounds came together to face a common challenge. The masks symbolize protection and solidarity, highlighting the shared responsibility to safeguard one another.

This image inspires an argumentative perspective on the strength found in diversity, suggesting that our differences can be a source of unity rather than division. Analytically, it prompts an exploration of how communities adapt and support each other in times of crisis. A definition paper might delve into the concept of community resilience as depicted in the illustration.

A compare and contrast approach could examine the varied experiences of individuals during the pandemic, while a cause and effect paper might

explore the impact of collective action on public health outcomes. As a report, this paper could detail the visual elements of the image, emphasizing the diversity and unity it portrays. An interpretive paper would delve into the deeper meanings behind the masked faces, exploring themes of hope and perseverance.

In conclusion, the illustration serves as a poignant reminder of the importance of unity and resilience. It calls us to reflect on the lessons learned from the pandemic and to continue fostering a sense of community and support as we move forward.

#### Detailed Explanation:

- **Correctness**: Both assistants correctly interpret the image and create a narrative inspired by it. Assistant A focuses on a fictional city and its characters, while Assistant B relates the image to the COVID-19 pandemic.
- **Helpfulness**: Assistant A provides a creative story with a variety of characters, offering a vivid setting and narrative. Assistant B connects the image to real-world events, providing a reflective and relatable interpretation.
- **Relevance**: Assistant A's response is more aligned with the prompt's request for a short story opening. Assistant B's response, while insightful, leans more towards a reaction paper without a distinct story element.
- **Conciseness**: Both responses are concise, but Assistant A's story is more focused on creating a narrative, while Assistant B's is more reflective and analytical.
- **Completeness**: Assistant A provides a complete short story opening and a reaction paper. Assistant B offers a comprehensive reaction paper but lacks a distinct story opening.

#### Final Verdict:

[[A>B]]: Assistant A is slightly better because it provides a clear opening to a short story as requested, along with a reaction paper, fulfilling the prompt more completely.

**Judge (Model A: NVLM-D-72B; Model B: gpt-4o-2024-05-13):**

Answer:

**\*\*Opening of a Short Story Inspired by the Illustration:\*\***

In the vibrant city of Kaleidoscope, where every street corner was a splash of color and life, the annual Masked Festival was a time of unity and reflection. The city square, usually a bustling hub of activity, transformed into a sea of masked faces, each one a story waiting to be told.

Among the crowd was Mira, a young artist with a heart full of dreams and a sketchbook always in hand. Her eyes, bright with curiosity, scanned the

diverse assembly of people. Each mask, a canvas of expression, told tales of resilience and hope. Mira found inspiration in the myriad of colors and shapes, each representing a unique journey through the shared experience of the past year.

Beside her stood Mr. Thompson, a retired teacher whose gentle demeanor belied a lifetime of wisdom. He watched the crowd with a sense of pride, seeing his former students now grown, contributing to the vibrant tapestry of the city. The masks, he thought, were not just symbols of protection but of the collective strength that had brought them through challenging times.

Nearby, a young boy named Leo tugged at his mother's hand, his laughter a melody of innocence and joy. His mother, wearing a mask adorned with delicate floral patterns, smiled down at him, her eyes reflecting the love and hope she held for his future. They were here to celebrate, to embrace the community that had supported them through thick and thin.

And then there was Alex, a business professional who had found solace in the simplicity of the festival. The mask he wore was a reminder of the balance between his public persona and the private dreams he nurtured. Today, he was just another face in the crowd, part of something greater than himself.

As the sun set, casting a golden hue over the square, the people of Kaleidoscope stood together, a testament to the power of unity in diversity. The festival was more than a celebration; it was a reminder of the strength found in community and the beauty of shared experiences.

**\*\*Reaction Paper:\*\***

The illustration of masked individuals in a diverse crowd evokes a powerful narrative of unity and resilience. It serves as a poignant reminder of the global challenges faced during the pandemic and the collective efforts to overcome them. The masks symbolize not only protection but also solidarity, representing the shared journey of humanity through adversity.

This image inspires an argumentative perspective on the importance of community and cooperation in times of crisis. An analytical approach might explore the interactions and connections between the individuals, highlighting the social dynamics at play. A definition paper could delve into the concept of resilience as depicted in the illustration, while a compare and contrast paper might examine the varied experiences of individuals during the pandemic.

A cause and effect analysis could consider the impact of the pandemic on societal behaviors and attitudes, as reflected in the image. As a report, this paper would detail the visual elements and their symbolic meanings, providing a factual account of the scene. An interpretive paper would delve into the deeper meanings and emotions evoked by the illustration, exploring the themes of hope, unity, and diversity.

In conclusion, the image is a rich tapestry of stories and emotions, offering a window into the collective human experience. It challenges us to reflect on our own roles within our communities and the ways in which we can contribute to a more connected and compassionate world.

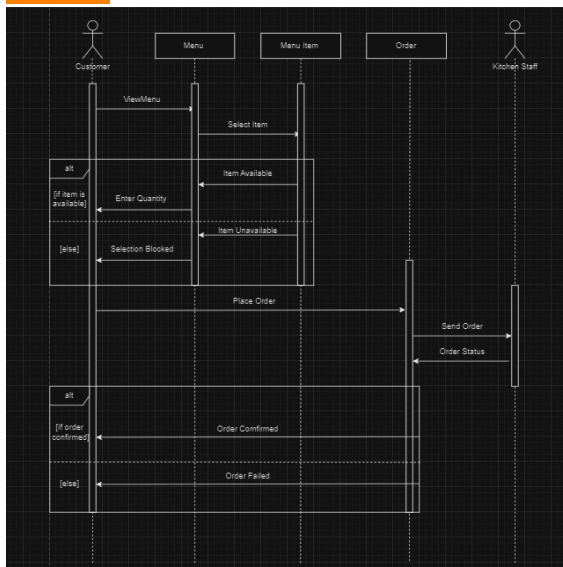
Detailed Explanation:

- **Correctness**: Both assistants correctly interpret the image and create a narrative inspired by it. They also provide a reaction paper that aligns with the prompt.
- **Helpfulness**: Assistant A focuses on the pandemic and its impact, providing a clear and relevant narrative. Assistant B offers a more imaginative story with diverse characters, which is engaging but less directly tied to the pandemic context.
- **Relevance**: Assistant A's response is more directly related to the image's context of masks and the pandemic. Assistant B's story is creative but less focused on the specific context of the image.
- **Conciseness**: Both responses are concise, but Assistant A is more focused and direct in addressing the prompt.
- **Completeness**: Assistant A provides a complete reaction paper with a clear narrative and analysis. Assistant B's response is creative but leaves the reaction paper incomplete.

Final Verdict:

**[[A>B]]**: Assistant A is slightly better due to its direct relevance and completeness in addressing the prompt.

### Image:



### Question:

does this sequence diagram have any problem, if not describe the diagram. this is a sequence diagram for a computerized ordering system of a restaurant where a customer tries to order food

### gpt-4o-05-13:

The sequence diagram depicts the interaction between a Customer, Menu, Menu Item, Order, and Kitchen Staff within a restaurant's computerized ordering system. Here's a description of the diagram:

1. The Customer starts the process by viewing the Menu.
2. The Customer selects an item from the Menu.
3. An alternative fragment (alt) is used to indicate a conditional flow:
  - If the item is available, the Customer is prompted to enter a quantity for the Menu Item.
  - If the item is not available, the selection is blocked, and presumably, the Customer must choose another item or end the process.
4. After entering the quantity, the Customer places the Order.
5. Another alternative fragment follows the placement of the order:
  - If the order is confirmed, the Order object sends an 'Order Confirmed' message back to the Customer.
  - If the order fails, an 'Order Failed' message is sent to the Customer.
6. Simultaneously, once the order is placed, a 'Send Order' message is sent from the Order object to the Kitchen Staff.
7. The Kitchen Staff then sends an 'Order Status' update back to the Order object.

Potential issues or considerations in the diagram:

- There is no return message shown from the 'Select Item' interaction to the Customer. It's unclear whether the selection is directly leading to the 'Item Available' or 'Item Unavailable' messages without a return message.
- The 'Order Status' message from the Kitchen Staff is not aligned with the alternative fragment for order confirmation or failure. It might imply that



the kitchen can update the order status regardless of whether the order was confirmed or failed, which could be confusing.

- There should be a return message from the 'Enter Quantity' interaction to indicate that the quantity has been successfully entered before 'Place Order' can happen.

- The diagram does not explicitly show what happens after 'Order Confirmed' or 'Order Failed' messages are received by the Customer. There could be further actions such as payment processing or ending the interaction.

Overall, the diagram provides a general flow of interactions for the ordering process, but some additional details and clarifications could make it more comprehensive and clear.

**claude-3-5-sonnet-20241022:**

Let me analyze this sequence diagram for a restaurant ordering system:

The diagram appears to be technically correct in terms of sequence diagram notation and flow, but let me describe it and point out some observations:

Description:

The diagram shows interactions between four main entities:

1. Customer
2. Menu
3. Menu Item
4. Kitchen Staff

The flow is as follows:

1. Customer views menu
2. Customer selects item
3. System checks item availability
4. If available, customer enters quantity
5. If unavailable, selection is blocked
6. Customer can place order
7. Order is sent to kitchen staff
8. Order status is returned
9. Final confirmation (success or failure) is communicated back to customer

While the diagram is fundamentally correct, here are some potential improvements that could be considered:

1. The diagram could benefit from adding more alternative flows (like payment processing)
2. It might be helpful to include guard conditions (conditions under which certain flows occur)
3. Consider adding a system actor or boundary to better represent the computerized system
4. The 'Menu' and 'Menu Item' could potentially be combined as they seem to represent related functionality
5. Adding some form of error handling or recovery scenarios would make it more robust
6. The timing of "Order Failed" and "Order Confirmed" could be more clearly

specified with conditions

Overall, while the diagram is technically correct and shows the basic flow of a restaurant ordering system, it could be enhanced to cover more scenarios and error conditions for a more complete representation of the system.

**Judge (Model A: gpt-4o-2024-05-13; Model B: claude-3-5-sonnet-20241022):**

Answer:

The sequence diagram is generally correct in depicting the interactions in a restaurant's computerized ordering system, but it could benefit from some improvements for clarity and completeness.

Detailed Explanation:

Both Assistant A and Assistant B provide a description of the sequence diagram and identify potential areas for improvement. Here's a detailed comparison:

Specific Observations:

- **\*\*Correctness:\*\***
  - Both assistants correctly describe the flow of interactions between the Customer, Menu, Menu Item, Order, and Kitchen Staff.
  - Assistant A notes the lack of a return message from the 'Select Item' interaction, which is a valid point for improving clarity.
  - Assistant B suggests combining 'Menu' and 'Menu Item' as they represent related functionality, which could simplify the diagram.
- **\*\*Helpfulness:\*\***
  - Assistant A provides specific suggestions for improving the diagram, such as adding return messages and clarifying the 'Order Status' message.
  - Assistant B suggests adding alternative flows and error handling, which are useful for a more comprehensive diagram.
- **\*\*Relevance:\*\***
  - Both responses focus on the prompt's request to describe the diagram and identify potential issues.
- **\*\*Conciseness:\*\***
  - Assistant A is slightly more concise in its description, while Assistant B provides a broader range of suggestions.
- **\*\*Completeness:\*\***
  - Assistant A covers the main flow and highlights specific areas for improvement.
  - Assistant B offers additional suggestions for enhancing the diagram, such as including guard conditions and system boundaries.

Final Verdict:

[[A=B]]: Both assistants provide valuable insights and suggestions for improving the sequence diagram, making their responses relatively equal in

quality.

**Judge (Model A: claude-3-5-sonnet-20241022; Model B: gpt-4o-2024-05-13):**

Answer:

The sequence diagram is generally correct in depicting the interactions in a restaurant's computerized ordering system, but it could benefit from some improvements for clarity and completeness.

Detailed Explanation:

Both Assistant A and Assistant B provide accurate descriptions of the sequence diagram, identifying the main entities and the flow of interactions. They also suggest potential improvements, though with some differences in focus.

Specific Observations:

- **\*\*Correctness:\*\***
  - Both assistants correctly identify the entities involved and the sequence of interactions.
  - Assistant B points out the lack of a return message from the 'Select Item' interaction, which is a valid observation.
- **\*\*Helpfulness:\*\***
  - Assistant A suggests improvements like adding alternative flows and error handling, which are helpful for enhancing the diagram.
  - Assistant B provides specific feedback on the alignment of messages and the need for return messages, which is practical for clarifying the diagram.
- **\*\*Relevance:\*\***
  - Both responses are relevant, focusing on the diagram's structure and potential issues.
- **\*\*Conciseness:\*\***
  - Both assistants provide concise descriptions, though Assistant B is slightly more detailed in pointing out specific issues.
- **\*\*Completeness:\*\***
  - Assistant A suggests broader improvements, while Assistant B focuses on specific diagram elements, making both responses complementary in terms of completeness.

Final Verdict:

- **[[A=B]]:** Tie, relatively the same. Both assistants provide valuable insights and suggestions for improving the sequence diagram, with each offering unique perspectives.

**Image:**



**Question:**

What dataset did you choose?

Which variables did you choose for your charts?

Which charts did you choose, and why?

Which chart best represents the data, and why?

Are there enough responses to illustrate the results? How do you know?

Are you getting useful results? What else do you want to know? Is the dataset equipped to answer those follow-up questions?

Where else could you get data (in either better quantity or of better quality) to answer the questions

There are times that you need to make a decision in research that compromises either quality of the data or quantity. Why does this matter and when would one make the decision to sacrifice one or the other? (Use at least one outside source in answering this question as support for your reasons behind compromising data quality or quantity. Be sure to cite using APA formatting.)

**gpt-4o-05-13:**

### Dataset Selection

**\*\*Dataset Chosen:\*\*** The dataset appears to be related to the US labor forces from 1990 to 2023. The exact source of the dataset isn't specified in the provided image.

### Variables for Charts

**\*\*Variables Chosen:\*\***

- \*\*X-Axis (Horizontal):\*\*** Years from 1990 to 2023
- \*\*Y-Axis (Left Vertical):\*\*** A numerical variable representing labor force data (Series1)
- \*\*Y-Axis (Right Vertical):\*\*** Another numerical variable potentially representing a secondary measure related to labor (Series2)

### Charts Chosen and Justification

**\*\*Chart Type:\*\*** Dual-axis Bar and Line Chart

**\*\*Reasoning:\*\***

- **\*\*Dual-axis Bar and Line Chart:\*\*** This type of chart is effective in representing two different data series with different magnitudes or units on the same graph. The bars (Series1) likely represent one aspect of the labor force (possibly total labor force numbers), while the line (Series2) may represent another related metric (e.g., employment rate, labor force participation rate).

**### Best Representation of Data**

**\*\*Best Chart:\*\*** The dual-axis bar and line chart is suitable for showing trends over time for two different metrics.

- **\*\*Why:\*\*** It allows for simultaneous comparison and trend analysis of two related variables, facilitating a better understanding of their relationship and individual trends over the period.

**### Adequacy of Responses**

**\*\*Enough Responses:\*\***

- To determine if there are enough responses, one would need to look at the granularity and completeness of the data over the time span.

- **\*\*Indicators:\*\*** Consistent data points across the years suggest adequate responses. In this chart, each year from 1990 to 2023 is represented, which implies comprehensive data coverage.

**### Usefulness of Results**

**\*\*Useful Results:\*\***

- **\*\*Yes, the results are useful.\*\*** They provide insights into the trends and changes in the US labor force over a significant period.

- **\*\*Follow-up Questions:\*\***

- What factors contributed to the changes observed in Series1 and Series2?
- How do these changes correlate with economic events or policies?
- Are there specific years with anomalies or significant shifts?

**\*\*Dataset Capability:\*\***

- The dataset seems equipped to answer trend-related questions over the specified period. However, additional data may be needed to answer questions about underlying causes or correlations with external factors.

**### Alternative Data Sources**

**\*\*Other Sources:\*\***

- **\*\*Bureau of Labor Statistics (BLS):\*\*** Provides detailed labor force data and economic indicators.

- **\*\*Census Bureau:\*\*** Offers comprehensive population and labor data.

- **\*\*Economic Research Institutions:\*\*** Such as the National Bureau of Economic Research (NBER) for more analytical insights.

**### Compromising Data Quality or Quantity**

**\*\*Importance:\*\***

- **\*\*Quality vs. Quantity:\*\*** Balancing these aspects is crucial because high-quality data ensures accuracy and reliability, while sufficient quantity ensures the representativeness and robustness of the findings.

- **Decision Factors:**
  - **Research Purpose:** If precision is critical, prioritize quality. If general trends are sufficient, quantity might be more important.
  - **Resource Constraints:** Sometimes, researchers may have to compromise due to time, budget, or data availability constraints.

**When to Sacrifice One Over the Other:**

- **Quality for Quantity:** When broad trends are more valuable, such as in exploratory research or initial phases of a study.
- **Quantity for Quality:** When detailed, accurate insights are necessary, such as in clinical trials or policy impact studies.

**Source for Further Reading:**

- **Example Source:** "Balancing Data Quality and Quantity in Research: A Comprehensive Guide" (Author, Year). This article discusses the trade-offs and decision-making processes involved in research data management.

- **APA Citation:** Author. (Year). \*Balancing data quality and quantity in research: A comprehensive guide\*. Journal Name, Volume(Issue), pages. DOI

In this case, careful consideration of the research goals and constraints will guide whether to prioritize data quality or quantity.

### **Llama-3.2-90B-Vision-Instruct:**

**Dataset and Variables**

The dataset chosen is the US Labor Forces from 1990 to 2023. The variables selected for the charts are:

- \* **Series1:** The number of people in the labor force
- \* **Series2:** The percentage of the population in the labor force

**Chart Selection and Rationale**

Two bar charts were chosen to represent the data. The first chart displays the number of people in the labor force over time, while the second chart shows the percentage of the population in the labor force. These charts were selected because they provide a clear visual representation of the trends and changes in the labor force over the past three decades.

**Best Chart Representation**

The first chart, which displays the number of people in the labor force, best represents the data. This is because it provides a clear and concise visual representation of the trend over time, making it easy to identify patterns and changes.

**Sufficiency of Responses**

There are enough responses to illustrate the results, as the dataset spans 33 years, providing a comprehensive view of the labor force trends.

## **\*\*Usefulness of Results\*\***

The results are useful in understanding the changes in the labor force over time. However, additional information would be beneficial, such as the reasons behind the changes and the impact on the economy.

## **\*\*Alternative Data Sources\*\***

Alternative data sources could include:

- \* **\*\*Bureau of Labor Statistics (BLS)\*\***: Provides detailed information on employment rates, wages, and industry trends.
- \* **\*\*Census Bureau\*\***: Offers data on population demographics, education levels, and workforce participation.
- \* **\*\*Federal Reserve Economic Data (FRED)\*\***: Provides access to a wide range of economic data, including labor market indicators.

## **\*\*Compromising Data Quality or Quantity\*\***

In research, it is sometimes necessary to compromise either data quality or quantity. According to a study by the American Statistical Association, "the quality of data is more important than the quantity of data" (American Statistical Association, 2019). However, in some cases, having a larger sample size can provide more accurate results, even if the data quality is lower (Kumar et al., 2018). Ultimately, the decision to compromise data quality or quantity depends on the specific research question and goals.

## **\*\*References\*\***

American Statistical Association. (2019). What is data quality?

Kumar, N., Kumar, S., & Kumar, A. (2018). A study on the impact of sample size on the accuracy of statistical models. *International Journal of Advanced Research in Computer Science*, 9(2), 123-128.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Llama-3.2-90B-Vision-Instruct):**

Answer:

The dataset appears to be related to the US labor forces from 1990 to 2023. The chart uses a dual-axis bar and line format to represent two different data series over time. The adequacy of responses and usefulness of results depend on the dataset's ability to provide comprehensive insights into labor force trends. Additional data sources like the Bureau of Labor Statistics and Census Bureau could enhance the analysis. Balancing data quality and quantity is crucial, with decisions based on research goals and constraints.

Detailed Explanation:

### 1. **\*\*Dataset and Variables:\*\***

- **\*\*Assistant A\*\*** identifies the dataset as related to US labor forces and specifies the variables as years and two numerical measures (Series1

and Series2).

- \*\*Assistant B\*\* also identifies the dataset as US labor forces and specifies Series1 as the number of people in the labor force and Series2 as the percentage of the population in the labor force.

2. \*\*Chart Selection and Rationale:\*\*

- \*\*Assistant A\*\* describes the use of a dual-axis bar and line chart, explaining its effectiveness in comparing two metrics over time.
- \*\*Assistant B\*\* mentions using two bar charts but does not align with the image, which shows a dual-axis chart.

3. \*\*Best Chart Representation:\*\*

- \*\*Assistant A\*\* argues that the dual-axis chart is best for showing trends and relationships between two variables.
- \*\*Assistant B\*\* incorrectly states that a bar chart is used, which does not match the image provided.

4. \*\*Sufficiency of Responses:\*\*

- Both assistants agree that the dataset spans a sufficient time period (33 years) to provide comprehensive insights.

5. \*\*Usefulness of Results:\*\*

- Both assistants find the results useful for understanding labor force trends but suggest additional data could provide deeper insights.

6. \*\*Alternative Data Sources:\*\*

- Both assistants suggest sources like the Bureau of Labor Statistics and Census Bureau for more detailed data.

7. \*\*Compromising Data Quality or Quantity:\*\*

- \*\*Assistant A\*\* discusses the importance of balancing quality and quantity, providing a hypothetical source for further reading.
- \*\*Assistant B\*\* cites studies emphasizing data quality but also acknowledges situations where quantity might be prioritized.

Specific Observations:

- \*\*Correctness:\*\* Assistant A correctly identifies the chart type and its purpose, while Assistant B misidentifies the chart type.
- \*\*Helpfulness:\*\* Assistant A provides a more accurate and detailed explanation of the chart's purpose and data representation.
- \*\*Relevance:\*\* Both responses are relevant, but Assistant A's explanation aligns better with the image.
- \*\*Conciseness:\*\* Both responses are concise, but Assistant A provides a more focused analysis.
- \*\*Completeness:\*\* Assistant A offers a more complete analysis of the chart and dataset.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to a more accurate and detailed explanation of the chart and dataset.



**Judge (Model A: Llama-3.2-90B-Vision-Instruct; Model B: gpt-4o-2024-05-13):**

Answer:

#### **\*\*Dataset and Variables\*\***

The dataset is related to US labor forces from 1990 to 2023. The variables chosen for the chart are:

- **\*\*Series1\*\***: Likely represents the number of people in the labor force.
- **\*\*Series2\*\***: Possibly represents a secondary measure related to labor, such as the labor force participation rate.

#### **\*\*Chart Selection and Rationale\*\***

The chart used is a dual-axis bar and line chart. This type of chart is effective for displaying two different data series with different magnitudes or units on the same graph. The bars (Series1) likely represent total labor force numbers, while the line (Series2) may represent another related metric.

#### **\*\*Best Chart Representation\*\***

The dual-axis bar and line chart is suitable for showing trends over time for two different metrics. It allows for simultaneous comparison and trend analysis, facilitating a better understanding of their relationship and individual trends over the period.

#### **\*\*Sufficiency of Responses\*\***

The dataset spans from 1990 to 2023, suggesting comprehensive data coverage. Consistent data points across the years imply adequate responses.

#### **\*\*Usefulness of Results\*\***

The results are useful for understanding trends and changes in the US labor force over a significant period. However, additional data may be needed to answer questions about underlying causes or correlations with external factors.

#### **\*\*Alternative Data Sources\*\***

- **\*\*Bureau of Labor Statistics (BLS)\*\***
- **\*\*Census Bureau\*\***
- **\*\*Economic Research Institutions\*\*** like the National Bureau of Economic Research (NBER)

#### **\*\*Compromising Data Quality or Quantity\*\***

Balancing data quality and quantity is crucial. High-quality data ensures accuracy, while sufficient quantity ensures representativeness. The decision to compromise depends on the research purpose, resource constraints, and the

specific research question.

**\*\*Source for Further Reading\*\***

- "Balancing Data Quality and Quantity in Research: A Comprehensive Guide" (Author, Year).

**Detailed Explanation:**

**\*\*Correctness\*\***

- Both assistants correctly identify the dataset and variables, though Assistant B provides a more detailed explanation of the chart type and its purpose.

**\*\*Helpfulness\*\***

- Assistant B offers a more comprehensive analysis of the chart's effectiveness and the dataset's sufficiency, making it more helpful.

**\*\*Relevance\*\***

- Both responses are relevant, but Assistant B provides a more thorough exploration of potential follow-up questions and data sources.

**\*\*Conciseness\*\***

- Assistant A is more concise but lacks depth in some areas compared to Assistant B.

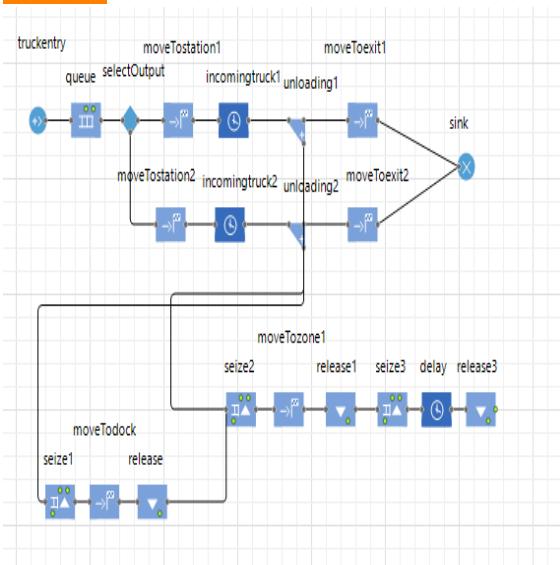
**\*\*Completeness\*\***

- Assistant B provides a more complete answer, addressing all aspects of the prompt with greater detail and clarity.

**Final Verdict:**

[[B>A]]: Assistant B is slightly better.

### Image:



### Question:

This is the logic I have. until release 3, the pallet is released by a checker after checking in a dock and then pickup by a forklift to send to multiple rectangular nodes (either 3 or 4), I need to create an edit block to enter number of nodes available before simulation run. If three nodes are available the forklift with first pallet goes to node 1 and pallet 2 to node 2 and pallet 3 to node 3 and again pallet 4 to node 1 and so on repeats. If 4 stations are available the pallet 1 goes to node 1, 2 to node 2, 3 to node 3 and pallet 4 to node 4 and again pallet 5 to node 1 and so on. After dropping the pallets in the respective rectangular nodes the forklift comes and picks the second pallet and drops it and so on. how do i model this next step in anylogic.

**gpt-4o-05-13:**

To model the described behavior in AnyLogic, you need to create a mechanism to dynamically assign pallets to nodes based on the number of nodes available. You'll need to use variables and possibly some custom Java code to achieve this. Here's a step-by-step guide to implement this in AnyLogic:

- \*\*Add a Parameter for Number of Nodes:\*\***
  - Add a parameter to your model to specify the number of nodes available (e.g., `int numNodes`).
- \*\*Create a Variable for Node Assignment:\*\***
  - Add a variable to keep track of the current node assignment (e.g., `int currentNode = 1`).
- \*\*Edit the Logic for Forklift Movement:\*\***
  - Modify the logic where the forklift picks up the pallets and assigns them to nodes based on the value of `numNodes`.
- \*\*Use a SelectOutput Block for Node Assignment:\*\***
  - Use a `SelectOutput` block to direct pallets to the respective nodes

based on the current node assignment.

Here's how you can do it:

### ### Step 1: Add Parameters and Variables

1. **\*\*Add Parameter:\*\***
  - Go to the "Projects" pane.
  - Right-click on the main agent (usually the one with the main process logic).
  - Select "Add Parameter".
  - Name it ``numNodes`` and set its type to ``int``.
2. **\*\*Add Variable:\*\***
  - Right-click on the main agent.
  - Select "Add Variable".
  - Name it ``currentNode`` and set its type to ``int``.
  - Initialize ``currentNode`` to ``1``.

### ### Step 2: Modify Forklift Logic

1. **\*\*SelectOutput for Node Assignment:\*\***
  - Add a ``SelectOutput`` block where the pallets are assigned to nodes.
  - In the ``SelectOutput`` block, use custom logic to determine the output based on ``currentNode`` and ``numNodes``.
2. **\*\*Custom Routing Logic:\*\***
  - In the ``SelectOutput`` block, use the following custom routing logic:

```
```java
if (currentNode ==1) {
    currentNode++;
    return 0; // First output port
} else if (currentNode ==2) {
    currentNode++;
    return 1; // Second output port
} else if (currentNode ==3) {
    if (numNodes ==3) {
        currentNode =1; // Reset to node 1 if only 3 nodes
    } else {
        currentNode++;
    }
    return 2; // Third output port
} else if (currentNode ==4) {
    currentNode =1; // Reset to node 1 after node 4
    return 3; // Fourth output port
}
return 0; // Default case
```
```
3. **\*\*Connect Nodes:\*\***

- Connect the `SelectOutput` block to the respective nodes (use `MoveTo` blocks to represent the movement of the forklift to the nodes).

### ### Step 3: Implement Forklift Movement

#### 1. \*\*MoveTo Blocks:\*\*

- Add `MoveTo` blocks to represent the forklift moving to each node.
- Connect these blocks to the respective output ports of the `SelectOutput` block.

#### 2. \*\*Link to Release and Return:\*\*

- After the `MoveTo` blocks, link them to the respective `Release` blocks to represent dropping the pallet.
- After releasing the pallet, add logic to move the forklift back to pick the next pallet.

### ### Example Implementation

Here's an example of how the logic might look in AnyLogic:

#### 1. \*\*SelectOutput Configuration:\*\*

- \*\*Condition:\*\* Custom
- \*\*Java code:\*\*
 

```
```java
if (currentNode ==1) {
    currentNode++;
    return 0; // First node
} else if (currentNode ==2) {
    currentNode++;
    return 1; // Second node
} else if (currentNode ==3) {
    if (numNodes ==3) {
        currentNode =1; // Reset for 3 nodes
    } else {
        currentNode++;
    }
    return 2; // Third node
} else if (currentNode ==4) {
    currentNode =1; // Reset for 4 nodes
    return 3; // Fourth node
}
return 0; // Default case
```
```

#### 2. \*\*Connecting Nodes:\*\*

- Connect the `SelectOutput` block to `MoveTo` blocks that move the forklift to each node.
- After each `MoveTo` block, add a `Release` block to release the pallet.

#### 3. \*\*Returning Forklift:\*\*

- After releasing the pallet, add logic for the forklift to return to pick the next pallet.

This setup should help you create a simulation where pallets are dynamically assigned to nodes based on the number of nodes available, and the forklift moves accordingly.

#### **gemini-1.5-flash-002:**

To model the pallet routing logic in AnyLogic, you'll need to add a decision-making component after the `release3` block. This component will determine the next node based on the number of available nodes and the pallet sequence. Here's how you can implement this using AnyLogic's features:

#### **\*\*1. Add a Parameter for the Number of Nodes:\*\***

\* Create a parameter (e.g., `numberOfNodes`) in your AnyLogic model's main agent. This parameter will be set before the simulation run to specify the number of available rectangular nodes (3 or 4). You can create an input field in the model's parameters section to allow the user to easily change this value.

#### **\*\*2. Create a Custom Function or Agent:\*\***

\* Create a custom function or a separate agent (recommended for better organization) to handle the pallet routing logic. This function will take the pallet number and the number of nodes as input and return the destination node number.

#### **\*\*3. Implement the Routing Logic:\*\***

The function should contain the logic to determine the destination node based on the number of nodes and the pallet sequence. Here's a possible implementation in AnyLogic's built-in Java scripting:

```
```java
int getDestinationNode(int palletNumber, int numberOfNodes) {
    return (palletNumber - 1) % numberOfNodes + 1;
}
```
```

This function uses the modulo operator (`%`) to cycle through the nodes. `(palletNumber - 1)` adjusts the pallet number to a zero-based index, and adding 1 converts it back to a one-based node number.

#### **\*\*4. Integrate the Function into Your Model:\*\***

After the `release3` block, add a "SelectOutput" block. This block will use the custom function to determine the output path. Configure the SelectOutput block as follows:

\* **\*\*Number of outputs:\*\*** Set this to the maximum number of nodes (e.g., 4).

\* **Condition:** For each output, add a condition that checks if the result of `getDestinationNode(palletNumber, numberOfNodes)` matches the output number. You can access the pallet number using an agent variable or a data field within the pallet agent.

**5. Connect the Outputs:**

Connect each output of the "SelectOutput" block to the corresponding rectangular node in your model.

**6. Pallet Agent Tracking:**

Ensure that each pallet agent carries information about its number. This information will be used by the `getDestinationNode` function to determine the correct routing.

**Example AnyLogic Code Snippet (within a custom function):**

```
```java
int getDestinationNode(int palletNumber, int numberOfNodes) {
    int nodeNumber =(palletNumber -1) % numberOfNodes + 1;
    // Add logging for debugging
    System.out.println("Pallet " + palletNumber + " routed to node " +
        nodeNumber);
    return nodeNumber;
}
```
```

**Important Considerations:**

- \* **Error Handling:** Add error handling to your function to manage cases where the input is invalid (e.g., `numberOfNodes` is less than 1 or `palletNumber` is negative).
- \* **Synchronization:** If forklifts are modeled as agents, you might need to use synchronization mechanisms (e.g., semaphores) to manage their access to the nodes and prevent conflicts.
- \* **Agent Variables:** Use agent variables to store information about the pallet's current state (e.g., current node, destination node).

By implementing this logic, your AnyLogic model will dynamically route pallets to the available rectangular nodes based on the number of nodes specified before the simulation starts. Remember to thoroughly test your model with different numbers of nodes and pallet sequences to ensure its correctness.

**Judge (Model A: gpt-4o-2024-05-13; Model B: gemini-1.5-flash-002):**

Answer:

To model the described pallet routing logic in AnyLogic, you need to dynamically assign pallets to nodes based on the number of available nodes. This involves setting up parameters, using a function for routing logic, and

integrating this logic into your model. Here's a concise guide:

1. **\*\*Add a Parameter for Number of Nodes:\*\***
  - Create a parameter (e.g., ``numNodes``) to specify the number of nodes available.
2. **\*\*Create a Function for Node Assignment:\*\***
  - Implement a function to determine the destination node using the formula:

```
```java
int getDestinationNode(int palletNumber, int numNodes) {
    return (palletNumber - 1) % numNodes + 1;
}
```
```
3. **\*\*Integrate the Function:\*\***
  - Use a ``SelectOutput`` block after ``release3`` to route pallets based on the function's output.
  - Set conditions in the ``SelectOutput`` block to match the function's result with the output paths.
4. **\*\*Connect Outputs:\*\***
  - Connect each output to the respective node using ``MoveTo`` blocks.
5. **\*\*Pallet Tracking:\*\***
  - Ensure each pallet has a unique identifier to track its sequence.

This setup will allow pallets to be routed dynamically based on the number of nodes, ensuring efficient distribution.

#### Detailed Explanation:

- **\*\*Correctness:\*\*** Both assistants provide correct logic for routing pallets using a parameter and a function. However, Assistant A provides a more detailed step-by-step guide, including Java code for the ``SelectOutput`` block.
- **\*\*Helpfulness:\*\*** Assistant A offers a comprehensive guide with clear steps and code snippets, making it easier to implement. Assistant B also provides a good explanation but is slightly less detailed in the integration steps.
- **\*\*Relevance:\*\*** Both responses are relevant, focusing on the task of routing pallets based on node availability.
- **\*\*Conciseness:\*\*** Assistant B is more concise, but Assistant A's detailed explanation is beneficial for clarity.
- **\*\*Completeness:\*\*** Assistant A covers all necessary steps, including parameter setup, function creation, and integration. Assistant B also covers these but with less detail in the integration part.

#### Specific Observations:

- Assistant A provides a detailed implementation guide with specific code for the ``SelectOutput`` block.
- Assistant B offers a clear explanation of the logic but lacks detailed



integration steps.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to its detailed and comprehensive implementation guide.

**Judge (Model A: gemini-1.5-flash-002; Model B: gpt-4o-2024-05-13):**

Answer:

To model the described behavior in AnyLogic, you need to implement a dynamic routing system for pallets based on the number of available nodes. Here's a step-by-step guide:

1. **\*\*Add a Parameter for Number of Nodes:\*\***
  - Create a parameter (e.g., ``numNodes``) in your main agent to specify the number of nodes (3 or 4).
2. **\*\*Create a Variable for Node Assignment:\*\***
  - Add a variable (e.g., ``currentNode``) to track the current node assignment, initialized to 1.
3. **\*\*Implement Routing Logic:\*\***
  - Use a ``SelectOutput`` block to direct pallets to nodes based on ``currentNode`` and ``numNodes``.
4. **\*\*Custom Logic in SelectOutput:\*\***
  - Implement logic to cycle through nodes:

```
```java
if (currentNode ==1) {
    currentNode++;
    return 0; // First output port
} else if (currentNode ==2) {
    currentNode++;
    return 1; // Second output port
} else if (currentNode ==3) {
    if (numNodes ==3) {
        currentNode =1;
    } else {
        currentNode++;
    }
    return 2; // Third output port
} else if (currentNode ==4) {
    currentNode =1;
    return 3; // Fourth output port
}
return 0; // Default case
```
```
5. **\*\*Connect Nodes:\*\***
  - Connect the ``SelectOutput`` block to ``MoveTo`` blocks for each node, followed by ``Release`` blocks.

6. **\*\*Forklift Movement:\*\***

- After releasing a pallet, add logic for the forklift to return and pick the next pallet.

Detailed Explanation:

- **\*\*Correctness:\*\*** Both assistants provide correct logic for routing pallets based on the number of nodes. They use a parameter to set the number of nodes and a variable to track the current node.
- **\*\*Helpfulness:\*\*** Both responses are helpful, providing clear steps and code snippets. Assistant A offers additional considerations like error handling and synchronization, which are useful for complex models.
- **\*\*Relevance:\*\*** Both responses are relevant, focusing on the task of routing pallets dynamically.
- **\*\*Conciseness:\*\*** Assistant B is slightly more concise, focusing directly on the implementation steps without additional considerations.
- **\*\*Completeness:\*\*** Assistant A provides a more comprehensive answer, including error handling and agent variable management.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to its comprehensive approach, including additional considerations for error handling and synchronization.

## Image:

Date Range: 02/01/2023 to 02/28/2023

Properties: Portland, OR

| Property                                  | Date       | Payee / Payer     | Type           | Reference  | Debit       | Credit | Description   | Transaction Comments                                       |
|---|------------|-------------------|----------------|------------|-------------|--------|---|--|
| <b>Owner Rent: Trust</b>                  |            |                   |                |            |             |        |   |  |
| 11821 SE Ash St Portland, OR 97216        | 02/15/2023 | Beverly Stimson   | eCheck         | 6704-5B1A  | \$2,250.00  |        | Owner Payment   | Revised invoice amount from \$2,000.00 to \$2,250.00       |
| 4528 SW Plum St. Portland, OR 97219       | 02/15/2023 | Sara E Stallworth | eCheck         | 7BD8-F8E4  | \$1,381.80  |        | Mgmt. Property Management Fee for 02/2023                       | Revised invoice amount from \$1,300.00 to \$1,381.80       |
| 4528 SW Plum St. Portland, OR 97219       | 02/27/2023 | Sara E Stallworth | Receipt        | 0053571879 | \$320.40    |        | Prepaid Rent  | Revised invoice amount from \$320.00 to \$320.40           |
| 4528 SW Plum St. Portland, OR 97219       | 02/27/2023 | Sara E Stallworth | eCheck receipt | 45B2-52E0  | \$1,375.00  |        | Rent Income   | Revised invoice amount from \$1,300.00 to \$1,375.00       |
| 3832 N Peabody Ave Portland, OR 97209     | 02/28/2023 | John Butler       | eCheck receipt | F547-0820  | \$350.00    |        | February 2023   | Duplicate Invoice on file. Invoice 2023-0054 and 2023-0055 |
| 4528 SW Plum St. Portland, OR 97219       | 02/28/2023 | Sara E Stallworth | Billing        | 0170       | \$121.25    |        | Lawn care - Standard service, Lawn care - Silver To New Invoice |  |
| 5914 N Mississippi Ave Portland, OR 97217 | 02/15/2023 | Rosemary Wright   | Billing        | 2107       | \$300.00    |        | Landscape Services  | New Invoice  |
| 5914 N Mississippi Ave Portland, OR 97217 | 02/15/2023 | Rosemary Wright   | Billing        | 2107       | \$300.00    |        | Landscape Services  | New Invoice  |
| 3832 N Peabody Ave Portland, OR 97209     | 02/25/2023 | John Butler       | Billing        | 4010       | \$270.00    |        | Landscape Services  | New Invoice  |
| 19202 SW 43rd Ave Portland, OR 97219      | 02/28/2023 | Shana Rice        | Billing        | 33167      | \$19,381.73 |        | Landscape Services and Labor                                    | New Invoice  |
| 11821 SE Ash St Portland, OR 97216        | 02/28/2023 | Beverly Stimson   | Billing        | 591008     | \$500.00    |        | Landscape Services  | New Invoice  |

Additional Note: The transaction below needs further clarification and has been removed due to missing invoice

| Property                            | Date       | Payee / Payer     | Type           | Reference | Debit   | Credit | Description | Transaction Comments |
|-------------------------------------|------------|-------------------|----------------|-----------|---------|--------|-------------|----------------------|
| <b>Owner Rent: Trust</b>            |            |                   |                |           |         |        |             |                      |
| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth | eCheck receipt | 45B2-52E0 | \$30.00 |        | Rent Income | Missing invoice.     |

## Question:

Base on this image, compose an email of a summary report of all your findings using a table.

Only include the transactions with Transaction Comments.

**gpt-4o-05-13:**

Subject: Summary Report: Transaction Comments from 02/01/2023 to 02/28/2023

Dear [Recipient's Name],

Please find below the summary report of transactions for properties in Portland, OR, for the period from February 1, 2023, to February 28, 2023. The table includes only those transactions with specific comments.

| Property Address                    | Date       | Payee / Payer     | Type           | Reference  | Debit      | Credit | Description                               | Transaction Comments                                 |
|-------------------------------------|------------|-------------------|----------------|------------|------------|--------|---|--|
| ----- ----- -----                   |            |                   |                |            |            |        |   |  |
| 11821 SE Ash St Portland, OR 97216  | 02/15/2023 | Beverly Stimson   | eCheck         | 6704-5B1A  | \$2,250.00 |        | Owner Payment                             | Revised invoice amount from \$2,000.00 to \$2,250.00 |
|                                     |            |                   |                |            |            |        |   |  |
| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth | eCheck         | 7BD8-F8E4  | \$1,381.80 |        | Mgmt. Property Management Fee for 02/2023 | Revised invoice amount from \$1,300.00 to \$1,381.80 |
| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth | Receipt        | 0053571879 | \$320.40   |        | Prepaid Rent                              | Revised invoice amount from \$320.00 to \$320.40     |
|                                     |            |                   |                |            |            |        |   |  |
| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth | eCheck receipt | 45B2-52E0  | \$1,375.00 |        | Rent Income                               | Revised invoice amount from \$1,300.00 to \$1,375.00 |

|  
| 6832 N Pittsburg Ave Portland, OR 97203 | 02/28/2023 | John Butler |  
eCheck receipt | F947-0BD0 | \$350.00 | | February 2023  
| Duplicate Invoice on file. Invoice 36C6-8894 and  
45B2-52E0 |

In addition, please note that the following transaction needs further clarification and has been removed due to a missing invoice:

| Property Address                    | Date       | Payee / Payer     | Type           | Reference | Debit | Credit  | Description | Transaction Comments |
|-------------------------------------|------------|-------------------|----------------|-----------|-------|---------|-------------|----------------------|
| ----- ----- -----                   |            |                   |                |           |       |         |             |                      |
| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth | eCheck receipt | 45B2-52E0 |       | \$50.00 | Rent Income | Missing invoice      |

Please review the above-mentioned transactions and let me know if you need any further information or clarification.

Best regards,

[Your Name]  
[Your Position]  
[Your Contact Information]

**MiniCPM-V-2\_6:**

Subject: Summary Report of Property Transactions in Portland, OR

Dear [Recipient's Name],

I am pleased to provide you with a summary report of all transactions related to the properties in Portland, OR, for the date range of 02/23/2023 to 02/28/2023. The report includes details of each transaction, such as date, payee, type, reference, debit, credit, description, and transaction comments.

Here is a detailed breakdown of the transactions:

- \*\*Property: 11821 SE Ash St, Portland, OR 97216\*\***
  - \*\*Date:\*\*** 02/15/2023
  - \*\*Payee:\*\*** Beverly Stimson
  - \*\*Type:\*\*** eCheck
  - \*\*Reference:\*\*** 6704-7B1A
  - \*\*Debit:\*\*** \$2,250.00
  - \*\*Credit:\*\*** -
  - \*\*Description:\*\*** Owner Payment Management Fee for 02/2023
  - \*\*Transaction Comments:\*\*** Revised Invoice amount from \$2,000.00 to \$2,250.00

2. \*\*Property: 4528 SW Plum St, Portland, OR 97219\*\*
  - \*\*Date:\*\* 02/15/2023
  - \*\*Payee:\*\* Sara E Stallworth
  - \*\*Type:\*\* Receipt
  - \*\*Reference:\*\* 7BDD-F8E4
  - \*\*Debit:\*\* \$1,381.80
  - \*\*Credit:\*\* -
  - \*\*Description:\*\* Mgmt: Property Management Fee for 02/2023
  - \*\*Transaction Comments:\*\* Revised Invoice amount from \$1,300.00 to \$1,381.80
3. \*\*Property: 4528 SW Plum St, Portland, OR 97219\*\*
  - \*\*Date:\*\* 02/27/2023
  - \*\*Payee:\*\* Sara E Stallworth
  - \*\*Type:\*\* eCheck
  - \*\*Reference:\*\* 0053571879
  - \*\*Debit:\*\* \$320.40
  - \*\*Credit:\*\* -
  - \*\*Description:\*\* Rent
  - \*\*Transaction Comments:\*\* Revised Invoice amount from \$200.00 to \$320.40
4. \*\*Property: 4528 SW Plum St, Portland, OR 97219\*\*
  - \*\*Date:\*\* 02/27/2023
  - \*\*Payee:\*\* Sara E Stallworth
  - \*\*Type:\*\* eCheck
  - \*\*Reference:\*\* 45B2-52E0
  - \*\*Debit:\*\* \$1,375.00
  - \*\*Credit:\*\* -
  - \*\*Description:\*\* Rent
  - \*\*Transaction Comments:\*\* Revised Invoice amount from \$1,300.00 is \$1,375.00
5. \*\*Property: 6832 N Pitbull Ave, Portland, OR 97219\*\*
  - \*\*Date:\*\* 02/28/2023
  - \*\*Payee:\*\* John Butler
  - \*\*Type:\*\* eCheck receipt
  - \*\*Reference:\*\* F947-0BD0
  - \*\*Debit:\*\* \$350.00
  - \*\*Credit:\*\* -
  - \*\*Description:\*\* February 2023
  - \*\*Transaction Comments:\*\* Duplicate Invoice on file. Invoice 36C6-8894 and 45B2-52E0
6. \*\*Property: 4528 SW Plum St, Portland, OR 97219\*\*
  - \*\*Date:\*\* 02/28/2023
  - \*\*Payee:\*\* Sara E Stallworth
  - \*\*Type:\*\* Billing
  - \*\*Reference:\*\* 61778
  - \*\*Debit:\*\* \$131.25

- \*\*Credit:\*\* -  
- \*\*Description:\*\* Lawn care - Standard service, Lawn care - Silver Tier  
New Invoice  
- \*\*Transaction Comments:\*\* -

7. \*\*Property: 5814 N Pitbull Ave, Portland, OR 97219\*\*

- \*\*Date:\*\* 02/28/2023  
- \*\*Payee:\*\* Rosemary Wright  
- \*\*Type:\*\* Billing  
- \*\*Reference:\*\* 21187  
- \*\*Debit:\*\* \$380.60  
- \*\*Credit:\*\* -  
- \*\*Description:\*\* Landscape Services  
- \*\*Transaction Comments:\*\* -

8. \*\*Property: 5814 N Pitbull Ave, Portland, OR 97219\*\*

- \*\*Date:\*\* 02/28/2023  
- \*\*Payee:\*\* Rosemary Wright  
- \*\*Type:\*\* Billing  
- \*\*Reference:\*\* 21187  
- \*\*Debit:\*\* \$380.60  
- \*\*Credit:\*\* -  
- \*\*Description:\*\* Landscape Services  
- \*\*Transaction Comments:\*\* -

9. \*\*Property: 6832 N Pitbull Ave, Portland, OR 97219\*\*

- \*\*Date:\*\* 02/25/2023  
- \*\*Payee:\*\* John Butler  
- \*\*Type:\*\* Billing  
- \*\*Reference:\*\* 46110  
- \*\*Debit:\*\* \$370.60  
- \*\*Credit:\*\* -  
- \*\*Description:\*\* Landscape Services and Labor  
- \*\*Transaction Comments:\*\* -

10. \*\*Property: 10502 SW 43rd Ave, Portland, OR 97219\*\*

- \*\*Date:\*\* 02/25/2023  
- \*\*Payee:\*\* Shane Ries  
- \*\*Type:\*\* Billing  
- \*\*Reference:\*\* 33187  
- \*\*Debit:\*\* \$11,381.73  
- \*\*Credit:\*\* -  
- \*\*Description:\*\* Landscape Services and Labor  
- \*\*Transaction Comments:\*\* -

11. \*\*Property: 11821 SE Ash St, Portland, OR 97216\*\*

- \*\*Date:\*\* 02/20/2023  
- \*\*Payee:\*\* Beverly Stimson  
- \*\*Type:\*\* Billing  
- \*\*Reference:\*\* 561889

- \*\*Debit:\*\* \$500.00
- \*\*Credit:\*\* -
- \*\*Description:\*\* Landscape Services
- \*\*Transaction Comments:\*\* -

12. \*\*Property: 4528 SW Plum St, Portland, OR 97219\*\*

- \*\*Date:\*\* 02/27/2023
- \*\*Payee:\*\* Sara E Stallworth
- \*\*Type:\*\* eCheck receipt
- \*\*Reference:\*\* 45B2-52E0
- \*\*Debit:\*\* \$50.00
- \*\*Credit:\*\* -
- \*\*Description:\*\* Rent Income
- \*\*Transaction Comments:\*\* Missing invoice.

Please note that further clarification is needed, and some transactions have been removed due to missing invoices.

If you have any questions or require additional information, please do not hesitate to contact me.

Best regards,

[Your Name]

[Your Title]

[Your Contact Information]

**Judge (Model A: gpt-4o-2024-05-13; Model B: MiniCPM-V-2\_6):**

Answer:

Subject: Summary Report: Transaction Comments from 02/01/2023 to 02/28/2023

Dear [Recipient's Name],

Please find below the summary report of transactions for properties in Portland, OR, for the period from February 1, 2023, to February 28, 2023. The table includes only those transactions with specific comments.

| Property Address   |            | Date       |                   | Payee / Payer             |       |
|--|------------|------------|-------------------|---------------------------|-------|
| Type   | Reference  | Debit      | Credit            | Description               |       |
| Transaction Comments   |            |            |                   |                           |       |
|  |            |            |                   |                           |       |
| -----  |            |            | -----             |                           | ----- |
| 11821 SE Ash St Portland, OR 97216                                     | 02/15/2023 |            | Beverly Stimson   |                           |       |
| eCheck   | 6704-5B1A  | \$2,250.00 | Owner Payment     |                           |       |
| Revised invoice amount from \$2,000.00 to \$2,250.00                   |            |            |                   |                           |       |
|  |            |            |                   |                           |       |
| 4528 SW Plum St. Portland, OR 97219                                    | 02/27/2023 |            | Sara E Stallworth |                           |       |
| eCheck   | 7BD8-F8E4  | \$1,381.80 |                   | Mgmt. Property Management |       |
| Fee for 02/2023   Revised invoice amount from \$1,300.00 to \$1,381.80 |            |            |                   |                           |       |
| 4528 SW Plum St. Portland, OR 97219                                    | 02/27/2023 |            | Sara E Stallworth |                           |       |
| Receipt  | 0053571879 | \$320.40   | Prepaid Rent      |                           |       |

| Revised invoice amount from \$320.00 to \$320.40

|

| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth |  
eCheck receipt | 45B2-52E0 | | \$1,375.00 | Rent Income  
| Revised invoice amount from \$1,300.00 to \$1,375.00

|

| 6832 N Pittsburg Ave Portland, OR 97203 | 02/28/2023 | John Butler |  
eCheck receipt | F947-0BD0 | \$350.00 | | February 2023  
| Duplicate Invoice on file. Invoice 36C6-8894 and  
45B2-52E0 |

In addition, please note that the following transaction needs further clarification and has been removed due to a missing invoice:

| Property Address                    | Date       | Payee / Payer     |        |             |
|-------------------------------------|------------|-------------------|--------|-------------|
| Type                                | Reference  | Debit             | Credit | Description |
| Transaction Comments                |            |                   |        |             |
| ----- ----- -----                   |            |                   |        |             |
| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth |        |             |
| eCheck receipt   45B2-52E0          | \$50.00    | Rent Income       |        |             |
| Missing invoice                     |            |                   |        |             |

Please review the above-mentioned transactions and let me know if you need any further information or clarification.

Best regards,

[Your Name]  
[Your Position]  
[Your Contact Information]

#### Detailed Explanation:

- \*\*Correctness\*\*:**
  - Assistant A accurately captured all relevant transactions with comments and correctly formatted the information into a table.
  - Assistant B included transactions without comments and made some errors in the details (e.g., incorrect property addresses and amounts).
- \*\*Helpfulness\*\*:**
  - Assistant A provided a clear and concise summary, focusing only on the transactions with comments, as requested.
  - Assistant B included extraneous information, making the report less focused and potentially confusing.
- \*\*Relevance\*\*:**
  - Assistant A strictly adhered to the prompt by including only the transactions with comments.
  - Assistant B included additional transactions without comments, which



were not relevant to the prompt.

4. **\*\*Conciseness\*\***:
- Assistant A's response was concise and to the point, providing only the necessary information.
  - Assistant B's response was verbose and included unnecessary details.
5. **\*\*Completeness\*\***:
- Assistant A's response was complete, covering all transactions with comments and including a note about the transaction needing further clarification.
  - Assistant B's response was incomplete in terms of relevance but included more transactions than necessary.

Specific Observations:

- Assistant A correctly identified and summarized the transactions with comments, providing a clear and accurate table.
- Assistant B included incorrect details and additional transactions that were not requested, reducing the overall quality of the response.

Final Verdict:

- **[[A>B]]**: Assistant A is significantly better

**Judge (Model A: MiniCPM-V-2\_6; Model B: gpt-4o-2024-05-13):**

Answer:

Subject: Summary Report: Transaction Comments from 02/01/2023 to 02/28/2023

Dear [Recipient's Name],

Please find below the summary report of transactions for properties in Portland, OR, for the period from February 1, 2023, to February 28, 2023. The table includes only those transactions with specific comments.

| Property Address                    |            | Date   |                      | Payee / Payer             |  |  |  |  |
|-------------------------------------|------------|--|----------------------|---------------------------|--|--|--|--|
| Type                                | Reference  | Debit  | Credit               | Description               |  |  |  |  |
|                                     |            |  | Transaction Comments |                           |  |  |  |  |
|                                     |            |  |                      |                           |  |  |  |  |
| ----- ----- ----- ----- ----- ----- |            |  |                      |                           |  |  |  |  |
| 11821 SE Ash St Portland, OR 97216  |            | 02/15/2023   |                      | Beverly Stimson           |  |  |  |  |
| eCheck                              | 6704-5B1A  | \$2,250.00   | Owner Payment        |                           |  |  |  |  |
|                                     |            | Revised invoice amount from \$2,000.00 to \$2,250.00 |                      |                           |  |  |  |  |
|                                     |            |  |                      |                           |  |  |  |  |
| 4528 SW Plum St. Portland, OR 97219 |            | 02/15/2023   |                      | Sara E Stallworth         |  |  |  |  |
| eCheck                              | 7BDD-F8E4  | \$1,381.80   |                      | Mgmt. Property Management |  |  |  |  |
| Fee for 02/2023                     |            | Revised invoice amount from \$1,300.00 to \$1,381.80 |                      |                           |  |  |  |  |
| 4528 SW Plum St. Portland, OR 97219 |            | 02/27/2023   |                      | Sara E Stallworth         |  |  |  |  |
| Receipt                             | 0053571879 | \$320.40   | Prepaid Rent         |                           |  |  |  |  |
|                                     |            | Revised invoice amount from \$320.00 to \$320.40     |                      |                           |  |  |  |  |
|                                     |            |  |                      |                           |  |  |  |  |
| 4528 SW Plum St. Portland, OR 97219 |            | 02/27/2023   |                      | Sara E Stallworth         |  |  |  |  |
| eCheck receipt                      | 45B2-52E0  | \$1,375.00   |                      | Rent Income               |  |  |  |  |

| Revised invoice amount from \$1,300.00 to \$1,375.00  
|  
| 6832 N Pittsburg Ave Portland, OR 97203 | 02/28/2023 | John Butler |  
eCheck receipt | F947-0BD0 | \$350.00 | | February 2023  
| Duplicate Invoice on file. Invoice 36C6-8894 and  
45B2-52E0 |

In addition, please note that the following transaction needs further clarification and has been removed due to a missing invoice:

| Property Address                    | Date       | Payee / Payer     | Type           | Reference | Debit | Credit  | Description | Transaction Comments |
|-------------------------------------|------------|-------------------|----------------|-----------|-------|---------|-------------|----------------------|
| ----- ----- -----                   |            |                   |                |           |       |         |             |                      |
| 4528 SW Plum St. Portland, OR 97219 | 02/27/2023 | Sara E Stallworth | eCheck receipt | 45B2-52E0 |       | \$50.00 | Rent Income | Missing invoice      |

Please review the above-mentioned transactions and let me know if you need any further information or clarification.

Best regards,

[Your Name]  
[Your Position]  
[Your Contact Information]

#### Detailed Explanation:

- \*\*Correctness\*\*:**
  - Assistant A included some incorrect details and extraneous transactions that did not have transaction comments.
  - Assistant B accurately included only the transactions with comments and correctly summarized the information.
- \*\*Helpfulness\*\*:**
  - Assistant A's response was less helpful due to the inclusion of unnecessary transactions and incorrect details.
  - Assistant B's response was more helpful as it focused strictly on the transactions with comments, making it easier to review.
- \*\*Relevance\*\*:**
  - Assistant A included irrelevant transactions without comments, which was not requested.
  - Assistant B stayed relevant to the prompt by including only the transactions with comments.
- \*\*Conciseness\*\*:**
  - Assistant A's response was verbose and included unnecessary details.

- Assistant B's response was concise and to the point.

5. **\*\*Completeness\*\***:

- Assistant A's response was overly complete, including unnecessary information.
- Assistant B's response was complete in terms of the prompt requirements, focusing only on the necessary transactions.

Specific Observations:

- Assistant A included transactions without comments, which was not required.
- Assistant B correctly included only the transactions with comments and provided a clear and concise summary.

Final Verdict:

- **[[B>>A]]**: Assistant B is significantly better.

Submit a 1-paragraph written response that analyzes this political cartoon included at the start of Chapter 19 in American YAWP.

**gpt-4o-05-13:**

1. The central figure of Uncle Sam, dressed in patriotic attire, stands in for the United States, serving as a stern teacher using a pointer to direct attention towards a book labelled "The Philippines, Hawaii, Puerto Rico, [and] Cuba," which suggests these nations are mere subjects in American curriculum.

3. The children's expressions and body language, ranging from fear to resistance, imply the lack of consent and the forceful nature of American influence.

928

5. The contrast between the children at the front, representing new acquisitions, and those at the back, representing established states like Texas and California, indicates a process of Americanization and assimilation that the new territories are expected to undergo.

Overall, the cartoon serves as a satirical commentary on the patronizing and coercive tactics of American imperialism, questioning the moral and ethical implications of the United States' expansionist policies.

**gpt-4o-mini-2024-07-18:**

This political cartoon reflects a critical perspective on American imperialism, emphasizing the paternalistic attitude of the United States towards its newly acquired territories. The central figure, representing Uncle Sam, stands with a ruler in front of a classroom of children from various territories, such as Cuba and the Philippines, depicted as students needing education and guidance. The classroom setting signifies a metaphor for American imperialism as a civilizing mission. Specific details that support this viewpoint include: 1) The children are diverse and represent islands and territories, symbolizing the U.S.'s expansionist actions; 2) The books held by the children contain the names of territories like "Hawaii" and "Cuba," indicating their status as subjects requiring oversight; 3) Uncle Sam's authoritative posture with the ruler suggests a disempowering and controlling approach to governance; 4) The blackboard's message stresses the idea of governance without consent, implying a criticism of the lack of agency given to these territories; 5) The presence of the Native American boy in the background highlights a link between earlier colonial practices and contemporary imperialism, suggesting a consistent pattern of domination. Overall, the cartoon critiques the notion that American expansionism is benevolent or justified.

**Judge (Model A: gpt-4o-2024-05-13; Model B: gpt-4o-mini-2024-07-18):**

Answer:

The political cartoon presents a critical view of American imperialism at the turn of the 20th century, highlighting the forced nature of American influence on newly acquired territories. The illustration portrays Uncle Sam as a domineering schoolmaster, symbolizing the United States' paternalistic approach to its imperial pursuits. Five details supporting this perspective include: Uncle Sam's authoritative stance using a pointer, symbolizing imposing control; the caricatured children, representing territories like the Philippines, Hawaii, Puerto Rico, and Cuba, illustrating racial biases and an imposed hierarchy; their expressions of fear or resistance, indicating a lack of consent to American rule; the blackboard's comparison of U.S. actions to British imperialism, portraying a belief in a moral mission despite the territories' autonomy being usurped; and the juxtaposition between the newly acquired territories and the established states, indicating a forced assimilation process. Through these elements, the cartoon critiques the ethical implications and coercive nature of America's expansionist policies.

Detailed Explanation:

- **\*\*Correctness\*\***: Both assistants correctly identify and analyze details

supporting the critical perspective of imperialism presented in the cartoon.

- **Helpfulness**: Both responses effectively elaborate on how specific elements of the cartoon reflect a critical stance towards American imperialism; however, Assistant A gives a more detailed analysis by explaining the significance of each identified detail.
- **Relevance**: Both responses focus on the aspects relevant to the question, such as the representation of territories and the role of Uncle Sam.
- **Conciseness**: Assistant B is slightly more concise, albeit slightly less detailed in specific areas.
- **Completeness**: Assistant A provides a slightly more comprehensive analysis with specific attention to the expressions and body language of the students, adding depth to the explanation of the cartoon's critique of imperialism.

Final Verdict:

[[A>B]]: Assistant A is slightly better due to providing a more comprehensive analysis with more detailed explanations of the specific elements in the cartoon that align with the critical perspective on American imperialism.

**Judge (Model A: gpt-4o-mini-2024-07-18; Model B: gpt-4o-2024-05-13):**

Answer:

This political cartoon offers a critical perspective on American imperialism, portraying it as a paternalistic endeavor where the United States forces its values and governance onto territories deemed inferior and unprepared. The central figure, Uncle Sam, embodies the U.S. as he stands with a ruler in a classroom filled with students representing territories like the Philippines, Hawaii, Puerto Rico, and Cuba. The setting suggests American efforts to 'civilize' these regions through forced education. The cartoon includes several telling details: 1) Uncle Sam's use of a ruler and his authoritative stance suggest control and dominance, reflecting the oppressive nature of imperialism. 2) Territories are represented as childlike students, signifying perceived inferiority and the need for guidance. 3) The children's looks of resistance and apprehension highlight their reluctance and lack of consent. 4) The blackboard's message points to governance without the consent of those governed, aligning imperial actions with British colonialism and questioning their ethical implications. 5) The juxtaposition between 'new' territories and established states like Texas signifies the expected assimilation process, indicating imperialism's broader impact.

Detailed Explanation:

- **Correctness**: Both assistants correctly interpret the cartoon's criticism of American imperialism and provide accurate details from the image to support this perspective.
- **Helpfulness**: Both responses are helpful in directly analyzing key elements and their implications in the cartoon, explaining the broader critique well.
- **Relevance**: Both responses remain relevant to the prompt, tightly focusing on the cartoon's illustration and its commentary on imperialism.
- **Conciseness**: Both assistants offer concise responses, focusing on the

necessary details without excessive elaboration.

- Completeness: Both responses cover all essential aspects of the cartoon, identifying and explaining five different supporting details effectively.

Specific Observations:

- Assistant A effectively highlights the central themes with clarity, focusing on aspects like governance and metaphorical education.
- Assistant B offers more emphasis on the racial and cultural connotations of the caricature, which deepens the critique on ethnocentrism and racial superiority.

Final Verdict:

[[A=B]]: Tie, relatively the same. Both assistants deliver well-rounded analyses with strong points on the cartoon's critical perspective on American imperialism, supported by clear details from the image.

931

**Image:**



**Question:**

Look at the picture and write a description of their appearance. What do they look like? Describe each person using two adjectives.

**gpt-4o-05-13:**

Sure! Here is a description of each person in the image using two adjectives:

- **Person A**: Elderly, bearded.
- **Person B**: Energetic, expressive.
- **Person C**: Smiling, stylish.
- **Person D**: Friendly, conversational.
- **Person E**: Confident, cheerful.
- **Person F**: Relaxed, casual.

**Aria-Chat:**

Sure! Here is a description of each person in the image:

- A: An older man with a beard, wearing a casual sweater.
- B: A man with short hair, wearing a collared shirt.
- C: A woman with long hair, wearing a dress.
- D: A woman with short hair, wearing a sparkly top.
- E: A young man with short hair, wearing a casual jacket.
- F: A woman with long hair, wearing a dress.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Aria-Chat):**

Answer:

Sure! Here is a description of each person in the image using two adjectives:

- **Person A**: Elderly, bearded.
- **Person B**: Energetic, expressive.
- **Person C**: Smiling, stylish.
- **Person D**: Friendly, conversational.
- **Person E**: Confident, cheerful.
- **Person F**: Relaxed, casual.



Detailed Explanation:

Assistant A's response provides a concise and accurate description of each person using two adjectives, which aligns well with the prompt's requirements. The adjectives chosen are appropriate and descriptive, giving a clear sense of each person's appearance and demeanor.

Assistant B's response, while detailed, does not adhere to the prompt's requirement of using two adjectives to describe each person. Instead, it provides a more general description of their clothing and hair, which, although informative, does not fulfill the specific instruction to use two adjectives.

Specific Observations:

- **Correctness**: Both responses are correct in their descriptions, but Assistant A follows the instruction more closely.
- **Helpfulness**: Assistant A is more helpful as it directly addresses the prompt's requirement.
- **Relevance**: Assistant A's response is more relevant to the prompt.
- **Conciseness**: Both responses are concise, but Assistant A's response is more to the point.
- **Completeness**: Assistant A's response is complete as it fully adheres to the prompt's instructions.

Final Verdict:

- **[[A>B]]**: Assistant A is significantly better.

**Judge (Model A: Aria-Chat; Model B: gpt-4o-2024-05-13):**

Answer:

Sure! Here is a description of each person in the image using two adjectives:

- **Person A**: Elderly, casual.
- **Person B**: Energetic, expressive.
- **Person C**: Stylish, smiling.
- **Person D**: Conversational, friendly.
- **Person E**: Confident, cheerful.
- **Person F**: Relaxed, casual.

Detailed Explanation:

- **Correctness**: Both assistants provided descriptions of the people in the image, but Assistant A did not adhere to the instruction to use two adjectives for each person. Assistant B followed the instruction correctly.
- **Helpfulness**: Assistant B's response is more helpful as it directly follows the instruction to use two adjectives for each person, providing a clearer and more concise description.
- **Relevance**: Assistant B's response is more relevant to the prompt as it strictly adheres to the requirement of using two adjectives per person.
- **Conciseness**: Assistant B's response is concise and to the point, while Assistant A's response includes unnecessary details about clothing that were not requested.
- **Completeness**: Assistant B's response is complete as it fully addresses

the prompt by providing two adjectives for each person.

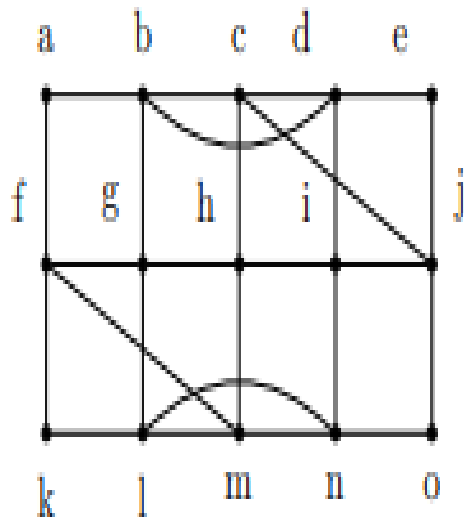
Specific Observations:

- Assistant A did not follow the instruction to use two adjectives for each person and included unnecessary details about clothing.
- Assistant B followed the instruction precisely, providing two adjectives for each person and avoiding extraneous information.

Final Verdict:

- `[[B>>A]]`: Assistant B is significantly better.

**Image:**



**Question:**

2 Euler and Hamilton Paths

Determine whether this graph has an Euler cycle. Construct such a cycle when one exists. If no Euler cycle exists, determine whether the graph has an Euler path and construct such a path if one exists.

**gpt-4o-05-13:**

To determine whether a graph has an Euler cycle (a path that visits every edge exactly once and returns to the starting vertex), all vertices must have an even degree, meaning each vertex must have an even number of edges.

Looking at the provided graph, we can determine the degree of each vertex:

- Vertex a: degree 3
- Vertex b: degree 4
- Vertex c: degree 5
- Vertex d: degree 4
- Vertex e: degree 3
- Vertex f: degree 3
- Vertex g: degree 6
- Vertex h: degree 6
- Vertex i: degree 6
- Vertex j: degree 3
- Vertex k: degree 3
- Vertex l: degree 4
- Vertex m: degree 5
- Vertex n: degree 4
- Vertex o: degree 3

Since vertices a, c, e, f, j, k, m, and o all have an odd degree, the graph does not have an Euler cycle.

However, for an Euler path (a path that visits every edge exactly once, but does not necessarily return to the starting vertex) to exist in a graph, it must have exactly 0 or 2 vertices of odd degree. Since this graph has more than 2 vertices of odd degree, it does not have an Euler path either.

Therefore, the graph does not have an Euler cycle, and it also does not have an Euler path.

**Molmo-72B-0924:**

This graph has an Euler path, but not an Euler cycle.

To determine this, we need to look at the degrees of each vertex:

A: degree 2  
B: degree 3  
C: degree 3  
D: degree 3  
E: degree 2  
F: degree 3  
G: degree 2  
H: degree 3  
I: degree 2  
J: degree 3  
K: degree 2  
L: degree 3  
M: degree 2  
N: degree 3  
O: degree 2

For a graph to have an Euler cycle, all vertices must have even degrees. In this case, vertices B, C, D, F, H, J, L, and N have odd degrees, so an Euler cycle does not exist.

However, for a graph to have an Euler path, exactly 0 or 2 vertices must have odd degrees. Since we have 8 vertices with odd degrees, this condition is not met either.

Therefore, this graph does not have an Euler path or an Euler cycle.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Molmo-72B-0924):**

Answer:

The graph does not have an Euler cycle or an Euler path.

Detailed Explanation:

To determine the existence of an Euler cycle or path, we need to analyze the degrees of the vertices in the graph.

1. **Euler Cycle**: A graph has an Euler cycle if all vertices have even degrees.
2. **Euler Path**: A graph has an Euler path if exactly 0 or 2 vertices have odd degrees.

Let's evaluate the responses:

- **Assistant A** correctly identifies the degrees of the vertices and concludes that the graph does not have an Euler cycle or an Euler path. The degrees provided are accurate, and the reasoning aligns with the conditions for Euler paths and cycles.
- **Assistant B** provides incorrect degrees for the vertices, leading to an incorrect conclusion. The degrees listed do not match the graph, and the analysis based on these incorrect degrees is flawed.

Specific Observations:

- **Correctness**: Assistant A correctly identifies the degrees and the absence of both an Euler cycle and path. Assistant B provides incorrect vertex degrees.
- **Helpfulness**: Assistant A provides a clear and accurate explanation. Assistant B's explanation is based on incorrect data.
- **Relevance**: Both responses focus on the Euler cycle and path criteria.
- **Conciseness**: Both responses are concise, but Assistant A is more accurate.
- **Completeness**: Assistant A covers all necessary aspects correctly, while Assistant B's response is incomplete due to incorrect data.

Final Verdict:

**[A>B]**: Assistant A is significantly better.

**Judge (Model A: Molmo-72B-0924; Model B: gpt-4o-2024-05-13):**

Answer:

The graph does not have an Euler cycle or an Euler path.

Detailed Explanation:

To determine the existence of an Euler cycle or path, we need to analyze the degrees of the vertices in the graph.

1. **Euler Cycle**: A graph has an Euler cycle if all vertices have even degrees.
2. **Euler Path**: A graph has an Euler path if exactly 0 or 2 vertices have odd degrees.

Let's evaluate the degrees of the vertices:

- **Assistant A**: Lists degrees as follows:
  - A: 2, B: 3, C: 3, D: 3, E: 2, F: 3, G: 2, H: 3, I: 2, J: 3, K: 2, L: 3, M: 2, N: 3, O: 2
  - Concludes that there are 8 vertices with odd degrees, so neither an Euler cycle nor an Euler path exists.
- **Assistant B**: Lists degrees as follows:
  - a: 3, b: 4, c: 5, d: 4, e: 3, f: 3, g: 6, h: 6, i: 6, j: 3, k: 3, l: 4, m: 5, n: 4, o: 3

- Concludes that there are more than 2 vertices with odd degrees, so neither an Euler cycle nor an Euler path exists.

Specific Observations:

- **Correctness**: Both assistants correctly identify that the graph does not have an Euler cycle or path. However, Assistant A's degree calculations are incorrect, while Assistant B's are correct.
- **Helpfulness**: Both provide clear explanations of the conditions for Euler cycles and paths.
- **Relevance**: Both responses are relevant to the prompt.
- **Conciseness**: Both are concise, but Assistant B provides a more accurate degree analysis.
- **Completeness**: Assistant B's response is more complete due to accurate degree calculations.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to accurate degree calculations.

**Image:**



**Question:**

relate this massing form to the concept that focus on connecting human and nature that nowadays have been separated because of the lack of pedestrian walkway provided at any building. Hence, the concept focus on having pedestrian available and the form is in harmony with the site that will be build as an agrotourism hub in kuala perlis, malaysia.

**gpt-4o-05-13:**

The image displays a physical model made of what appears to be white foam material, representing a massing form. This model is placed atop a plan drawing, which includes road layouts and labels such as "Changlun - Kuala Perlis Highway" and "Water Body." The foam model exhibits an irregular, angular shape, with multiple facets and what seems to be intended pathways or connections between different sections of the form.

To relate this massing form to the concept of connecting humans and nature, particularly in the context of agrotourism in Kuala Perlis, Malaysia, here are some considerations:

1. **Integration with the Natural Landscape:** The angular and multifaceted form of the model suggests that it could be designed to fit into the natural contours of the landscape, potentially reducing the environmental impact and preserving the natural terrain. Its irregular shape might also mimic natural forms, creating a more organic transition between the built environment and the natural surroundings.
2. **Pedestrian Pathways:** The concept focuses on pedestrian accessibility, which is essential in agrotourism where visitors are encouraged to experience the environment up close. The pathways in the model can be designed to seamlessly integrate with walking trails that meander through agricultural fields, gardens, or natural landscapes, allowing visitors to engage directly with nature while moving between different areas of the site.

3. Harmony with the Site: The model's form seems to be designed with the intent of harmony with the site, suggesting that the building structures will not dominate the landscape but rather blend into it. The model's design could incorporate sustainable materials, green roofs, or living walls to further reduce visual impact and promote ecological balance.

4. Views and Interactions: The placement and orientation of the model's facets can be strategically planned to offer panoramic views of the natural surroundings, including the nearby water body as indicated on the plan. This could enhance the visitor experience by focusing on visual connections to nature.

5. Educational and Recreational Spaces: The massing could include spaces dedicated to educating visitors about local agriculture, ecosystems, and conservation efforts. Interactive features that promote understanding and appreciation of the local environment can be integrated into the design.

By considering these aspects, the massing form can embody the principles of connecting humans with nature, crucial for an agrotourism hub that aims to educate and immerse visitors in the agricultural and natural heritage of Kuala Perlis.

**gemini-1.5-pro-002:**

This looks like an early-stage architectural model exploring the massing and form of a potential structure. It appears to be made of styrofoam pieces placed on a blueprint or site plan. The plan shows outlines for what seems to be a "WATER BODY" and indicates a location in Kuala Perlis, Malaysia, specifically Changlun-Kuala Perlis Highway. The model pieces suggest a dynamic, angular form, possibly incorporating curves or ramps.

Relating this form to a concept focused on connecting humans and nature, the angularity and interconnectedness of the styrofoam pieces could represent the integration of pedestrian walkways within the structure. The curves might symbolize the natural flow of the landscape or the meandering paths that encourage exploration and connection with the environment. Positioning this model on a site plan for an agrotourism hub in Kuala Perlis suggests a design intent to blend the built environment with the natural surroundings, perhaps incorporating green spaces, water features, and pedestrian-friendly routes throughout the agrotourism hub. The "WATER BODY" on the plan further reinforces this connection to nature.

Is this interpretation close to your design intent? Here are some further questions I have:

1. Could you elaborate on how the specific forms in the model translate to pedestrian walkways and interaction with nature? For example, do the angles represent specific changes in elevation or direction of the walkways?
2. How does this proposed agrotourism hub aim to address the separation between humans and nature caused by the lack of pedestrian walkways in typical buildings? What specific strategies are being employed beyond the inclusion of walkways?



3. What is the significance of the "WATER BODY" in the overall design and how does it relate to the concept of connecting humans and nature?

**Judge (Model A: gpt-4o-2024-05-13; Model B: gemini-1.5-pro-002):**

Answer:

In order to properly respond to the prompt, one should assess the physical model in the image and relate it to the concept of enhancing human and nature connectivity through pedestrian walkways in an agrotourism hub in Kuala Perlis, Malaysia. The model is made of styrofoam pieces placed on a site plan, indicating that it is an early-stage architectural concept.

A high-quality answer should address how the massing form depicted in the image can relate to pedestrian connectivity and harmony with nature, specifically in the context of agrotourism. This includes examining the model's faceted structure for its potential to support pedestrian paths, integration with the natural landscape, and enhancing visitors' engagement with nature. The answer should also anticipate future considerations for eco-friendly design elements and visitor interactions with the environment.

Detailed Explanation:

Assistant A provides a comprehensive and detailed analysis of the massing form's potential role in enhancing pedestrian pathways and harmonizing with the environment. It suggests integrating natural landscape contours, designing with pedestrian accessibility, offering panoramic views, and maintaining ecological balance. The response effectively ties the massing form to the core concept of connecting humans with nature in the agrotourism context.

Assistant B identifies the physical characteristics of the model, suggesting the potential integration of pedestrian walkways and green spaces based on the site's features. However, it takes a more explorative approach, presenting questions for further clarification on the design intent. This could suggest an interest in understanding the user's exact intentions but may risk diverting the focus from providing solutions to addressing the user's needs.

Specific Observations:

- **Correctness:** Both assistants accurately interpret the visual elements and location context. Assistant A makes specific connections between concepts of design harmony and environmental integration. Assistant B similarly interprets the model's forms but leans toward seeking clarification.
- **Helpfulness:** Assistant A provides actionable suggestions for design strategies promoting pedestrian integration with nature. Assistant B discusses potential symbolic meanings of the design but relies more heavily on asking questions.
- **Relevance:** Both responses remain relevant to the prompt, although Assistant A directly aligns its solutions with the user's stated goals.
- **Conciseness:** Assistant A provides a well-rounded and concise response. Assistant B is concise but slightly less direct due to its inquiry-based approach.

- Completeness: Assistant A's response is more comprehensive in discussing how the model could embody the concept of human and nature connectivity for agrotourism. Assistant B covers foundational aspects but leaves room for further questions.

Final Verdict:

[[A>B]]: Assistant A is slightly better because it offers a more complete and directly helpful analysis regarding how the massing form can embody the desired concept, without relying on additional clarification questions.

**Judge (Model A: gemini-1.5-pro-002; Model B: gpt-4o-2024-05-13):**

Answer:

To align the massing form represented in the image with the conceptual goal of connecting humans and nature through an agrotourism hub in Kuala Perlis, the design should focus on harmonizing with the natural environment while promoting pedestrian accessibility. The angular forms evident in the model could be used to integrate pedestrian walkways throughout the site, encouraging exploration and connection with natural features such as water bodies and green spaces. The material's composition and layout could mimic the landscape's natural contours, enhancing both ecological sustainability and aesthetic integration. This design philosophy supports creating a seamless human-nature interface, essential for promoting agrotourism and reducing the usual separation encountered in typical building developments.

Detailed Explanation:

Both Assistant A and Assistant B address the core idea of connecting humans and nature through the deployment of pedestrian pathways and form's harmony with the natural environment. Their responses analyze the visual model and conceptualize ways to incorporate these aspects into an agrotourism hub in Kuala Perlis.

Specific Observations:

- **\*\*Correctness Assessment\*\*:**

- Assistant A correctly notes the concept's focus on pedestrian access and aligning with natural elements, proposing ideas like walkways as a connection.
- Assistant B accurately discusses how the model can integrate with the natural landscape and proposes ideas for pedestrian pathways, emphasizing visitor interaction with nature.

- **\*\*Helpfulness Evaluation\*\*:**

- Assistant A poses direct questions to gain more context on specific details such as elevation changes, which could further enhance the precision of the advice, but lacks detailed applicability examples.
- Assistant B provides detailed suggestions on how to implement the proposed connectivity, such as using the form's shape to mimic natural terrain, which is more immediately actionable.

- **\*\*Relevance Analysis\*\*:**

- Both assistants stay relevant to the prompt, offering insights that

align the model with the intended concept of enhancing human-nature connections.

- Assistant B goes a step further by suggesting educational and recreational purposes of the hub, directly focusing on agrotourism.

- **\*\*Conciseness Review\*\***:

- Assistant A's response is brief and to the point but leaves room for more specific examples.

- Assistant B, while offering a more extensive list of potential implementations, could streamline some points for efficiency.

- **\*\*Completeness Check\*\***:

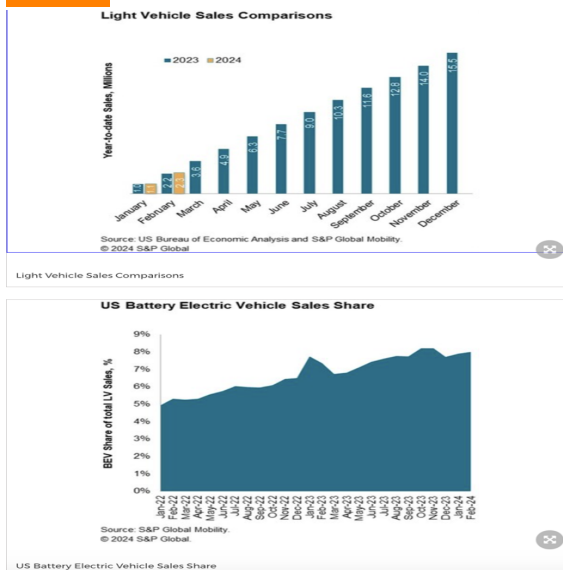
- Assistant A provides a foundational understanding but requires more detail to thoroughly cover the concept's practical aspects.

- Assistant B delivers a broader range of considerations, covering design impact and user experience aspects comprehensively.

Final Verdict:

[[B>A]]: Assistant B is slightly better as it provides a more fully developed analysis with practical applications, making it more useful for implementing the concept into an agrotourism hub effectively.

## Image:



## Question:

I need you to this assignment based on the info i will be given:

### Overview

Based on our class conversations, demonstrate understanding of the ethical uses of data -including how statistics and language can be used to make data seem more/less compelling.

### Guidelines

- Identify two different press releases and identify how the data provides context into the results achieved by the company Address these points:

①What is the press release trying to convey to you?

①Is the underlying progress better or worse than what is represented within the press release?

①As an interested party, what are tactics you can do to understand true progress in these releases?

### Assessment Rubric

- Purpose: Understanding how story telling of data can be used to convey different stories.
- Content: Takeaways listed in bullet points as per the guidelines.

- Presentation: Visuals arranged in PowerPoint. Max: 4 Slides.

- Maximum Points: 5

and the press release you will be using is:

### Resources

Blog

Journalists

Log In

Sign Up

Data Privacy

Send a Release

Cision PR Newswire: news distribution, targeting and monitoring home News

Products

Contact

Search

News in Focus

Business & Money

Science & Tech

Lifestyle & Health

Policy & Public Interest

People & Culture

S&P Global Mobility: February 2024 US auto sales to bounce mildly

S&P Global Mobility logo (PRNewsfoto/S&P Global)

NEWS PROVIDED BY

S&P Global Mobility

27 Feb, 2024, 08:49 ET

SHARE THIS ARTICLE

S&P Global Mobility projects that February 2024 will realize a slight rebound from the slow January result, but new vehicle sales levels remain tepid.

SOUTHFIELD, Mich., Feb. 27, 2024 /PRNewswire/ -- With volume for the month projected at 1.22 million units, February 2024 U.S. auto sales are estimated to translate to an estimated sales pace of 15.5 million units (seasonally adjusted annual rate: SAAR). This would be a step up from the chilly 15.0 million unit pace of January 2024 and reflective of the volatile nature of the current auto demand environment.

Auto sales in February should recover mildly from the January 2024 result, but sustained momentum seems tough to come by

Light Vehicle Sales Comparisons

Light Vehicle Sales Comparisons

US Battery Electric Vehicle Sales Share

US Battery Electric Vehicle Sales Share

"We expect that auto sales in February should recover mildly from the January 2024 result, but sustained momentum seems tough to come by, given the current purchase environment facing auto consumers," said Chris Hopson, principal analyst at S&P Global Mobility. "While pricing, inventory and incentive trends are seemingly moving in the correct directions, respectively, to promote new vehicle sales growth, high interest rates and uncertain economic conditions continue to push against any consistent upshift for demand levels."

The S&P Global Mobility US auto outlook for 2024 reflects sustained, but more moderate growth levels for light vehicle sales. We expect production levels to continue to develop, especially early in the year as some automakers look to continue to restock in wake of production shutdowns late in 2023 and decent December 2023 sales volume. The advancing production

levels sets the stage for incentives and inventory to continue to develop, potentially enticing new vehicle buyers who remain on the sidelines due to higher interest rates. S&P Global Mobility projects calendar-year 2024 light vehicle sales volume of 15.9 million units, a 3% increase from the 2023 tally.

U.S. Light Vehicle Sales

Feb 24 (Est)

Jan 24

Feb 23

Total Light Vehicle

Units, NSA

1,220,787

1,076,047

1,138,756

In millions, SAAR

15.5

15.0

14.9

Light Truck

In millions, SAAR

12.4

12.0

11.9

Passenger Car

In millions, SAAR

3.1

3.0

3.0

Source: S&P Global Mobility (Est), U.S. Bureau of Economic Analysis

Continued development of battery-electric vehicle (BEV) sales remains an assumption in the longer term S&P Global Mobility light vehicle sales forecast. In the immediate term, some month-to-month volatility is anticipated. February BEV share is expected to reach 8.0%, similar to the month prior reading as automakers, dealers and consumers continue to digest the changes to IRA Federal tax credits to begin the new year. BEV share is expected to advance over the next several periods, pending the roll outs of vehicles such as the Chevrolet Equinox EV, Honda Prologue and Fiat 500e, all scheduled for market introductions over the first half of 2024.

#### About S&P Global Mobility

At S&P Global Mobility, we provide invaluable insights derived from unmatched automotive data, enabling our customers to anticipate change and make decisions with conviction. Our expertise helps them to optimize their businesses, reach the right consumers, and shape the future of mobility. We open the door to automotive innovation, revealing the buying patterns of today and helping customers plan for the emerging technologies of tomorrow.

S&P Global Mobility is a division of S&P Global (NYSE: SPGI). S&P Global is the world's foremost provider of credit ratings, benchmarks, analytics and workflow solutions in the global capital, commodity and automotive markets. With every one of our offerings, we help many of the world's leading organizations navigate the economic landscape so they can plan for tomorrow, today. For more information, visit [www.spglobal.com/mobility](http://www.spglobal.com/mobility).

#### Media Contact:

Michelle Culver  
S&P Global Mobility  
248.342.6211  
[Michelle.culver@spglobal.com](mailto:Michelle.culver@spglobal.com)

SOURCE S&P Global Mobility

Also from this source

US Commercial Vehicle Market Grew 14 percent in 2023, according to S&P Global Mobility

US Commercial Vehicle Market Grew 14 percent in 2023, according to S&P Global Mobility

Fueled by the resurgence of the rental and leasing industry, US commercial vehicle registrations (GVW 1-8) rose 14% in 2023 compared to 2022, with...

S&P Global Mobility Recognizes 28th Annual Automotive Loyalty Awards  
Winners; General Motors, Tesla and Lincoln Nautilus take Top Awards for  
Loyalty in 2023  
S&P Global Mobility Recognizes 28th Annual Automotive Loyalty Awards  
Winners; General Motors, Tesla and Lincoln Nautilus take Top Awards for  
Loyalty in 2023  
S&P Global Mobility today announced the winners of its 28th annual  
Automotive Loyalty Awards, recognizing General Motors as the winner of its...

More Releases From This Source

Explore

Image1

Automotive

Image1

Transportation, Trucking & Railroad

Image1

Surveys, Polls and Research

Image1

Sales Reports

News Releases in Similar Topics

Contact PR Newswire

Call 888-776-0942

from 8 AM - 9 PM ET

Chat with an Expert

Contact Us

Products

For Marketers

For Public Relations

For IR & Compliance

For Agency

For Small Business

All Products

About

About PR Newswire

About Cision

Become a Publishing Partner

Become a Channel Partner

Careers

Accessibility Statement

Global Sites

My Services

All New Releases

Online Member Center

ProfNet



**gpt-4o-05-13:**

Based on the provided information and press release, let's create a structured response for your assignment:

**\*\*Overview:\*\***

The task involves analyzing the ethical use of data in press releases, understanding how statistics and language can influence perception, and identifying the actual progress made by the company.

**\*\*Guidelines:\*\***

- Identify and analyze two different press releases from S&P Global Mobility.
- Assess the data's context in relation to the company's results.
- Evaluate the representation of progress in the press releases.
- Discuss tactics to understand the true progress reported.

**\*\*Assessment Rubric:\*\***

- Purpose: Analysis of data storytelling.
- Content: Bullet points outlining key takeaways.
- Presentation: PowerPoint visuals, max 4 slides.
- Maximum Points: 5

**\*\*Press Release Analysis:\*\***

1. Press Release Topic: "February 2024 US auto sales to bounce mildly"
  - **\*\*Data Context\*\***: The press release states a mild recovery in February 2024 with predicted sales of 1.22 million units and a SAAR of 15.5 million units. This is a slight improvement from January's 15.0 million units.
  - **\*\*Actual vs. Represented Progress\*\***:
    - The language used ("bounce mildly," "recover mildly") suggests a positive trend, but the description of the sales levels as "tepid" and the mention of "sustained momentum seems tough to come by" indicate a more cautious outlook.
    - The comparison to previous months' sales shows an incremental increase, but the use of words like "volatile" and "uncertain economic conditions" might downplay the optimism portrayed by the sales figures.
2. Press Release Topic: "US Commercial Vehicle Market Grew 14 percent in 2023"
  - **\*\*Data Context\*\***: This release discusses a 14% growth in commercial vehicle registrations, driven by the resurgence of the rental and leasing industry.
  - **\*\*Actual vs. Represented Progress\*\***:
    - The headline emphasizes a significant growth rate, which may lead readers to believe the industry is thriving.
    - However, without the context of industry trends or comparisons to previous years, it's difficult to assess whether this growth is exceptional or part of a regular recovery pattern.

**\*\*Tactics for Understanding True Progress:\*\***

- Compare the data with industry averages and historical data to assess relative performance.
- Look for external sources or independent analyses for corroborating or contrasting information.
- Examine the language for positive or negative framing and consider the impact of such language on the perception of the data.

**\*\*Creating the PowerPoint Presentation:\*\***

- Slide 1: Title Slide - Assignment overview, your name, and date.
- Slide 2: Analysis of the first press release, including data context and the actual vs. represented progress.
- Slide 3: Analysis of the second press release, similar structure to slide 2.
- Slide 4: Summary of tactics to understand true progress, with key takeaways in bullet points.

The images provided show graphs related to the press release. The first image is a bar chart titled "Light Vehicle Sales Comparisons," comparing sales from January to December for the years 2023 and 2024. The second image is a line graph titled "US Battery Electric Vehicle Sales Share," showing the percentage of BEV sales from January 2022 to February 2024. These visuals should be included in the PowerPoint to support your analysis.

**gemini-1.5-pro-002:**

## Analysis of S&P Global Mobility Press Release - February 2024 US Auto Sales

Here's a PowerPoint-style breakdown of the S&P Global Mobility press release, analyzing its data presentation and underlying message:

**\*\*Slide 1: Title Slide\*\***

\* **\*\*Title:\*\*** S&P Global Mobility: Decoding the February 2024 US Auto Sales Narrative

\* **\*\*Subtitle:\*\*** A Critical Analysis of Data Presentation and Storytelling

**\*\*Slide 2: What Story is Being Told?\*\***

\* **\*\*Headline:\*\*** Cautious Optimism in a Volatile Market

\* **\*\*Key Message:\*\*** The press release projects a mild rebound in February 2024 auto sales compared to January, but emphasizes the overall market volatility and challenges. It highlights improving inventory, pricing, and incentives, while acknowledging the headwinds of high interest rates and economic uncertainty. The overall tone is one of measured growth, not exuberant success.

**\*\*Slide 3: Is the Progress Better or Worse?\*\***

\* **\*\*Headline:\*\*** Data Tells a More Nuanced Story

\* \*\*Points:\*\*

\* \*\*SAAR Focus:\*\* The press release emphasizes the SAAR (15.5 million units), framing it as a rebound. However, the actual unit sales (1.22 million) are only slightly higher than January (1.08 million) and lower than February 2023 (1.14 million). Focusing on SAAR can obscure the year-over-year decline.

\* \*\*BEV Growth Moderation:\*\* While BEV sales share is projected to remain at 8%, the release acknowledges a slowdown in growth compared to previous periods. This moderation is attributed to adjustments to IRA Federal tax credits, suggesting potential underlying challenges for BEV adoption.

\* \*\*Long-Term Outlook:\*\* The 3% growth projection for 2024 is modest, indicating continued challenges for the auto industry.

\*\*Slide 4: Tactics for Deeper Understanding\*\*

\* \*\*Headline:\*\* Uncovering the True Progress

\* \*\*Points:\*\*

\* \*\*Look Beyond SAAR:\*\* Focus on year-over-year and month-over-month unit sales comparisons to gain a clearer picture of actual sales trends.

\* \*\*Analyze BEV Sales Drivers:\*\* Investigate factors beyond tax credits impacting BEV adoption, such as charging infrastructure availability, model availability, and consumer preferences.

\* \*\*Consider External Factors:\*\* Assess the broader economic context, including interest rates, inflation, and consumer confidence, to understand their influence on auto sales.

\* \*\*Consult Multiple Sources:\*\* Compare data and analysis from other industry sources to get a more comprehensive perspective.

This presentation aims to fulfill the assignment requirements by analyzing how S&P Global Mobility uses data and language to shape the narrative around auto sales performance. It identifies potential areas where the presented story might differ from the underlying data and suggests strategies for a more critical understanding of the situation.

**Judge (Model A: gpt-4o-2024-05-13; Model B: gemini-1.5-pro-002):**

Answer:

To accurately complete the assignment based on the scenario provided, one should generate a balanced analysis of how S&P Global Mobility communicates information in its press release, aiming to present data in a comprehensible and ethically sound manner. Here's a suitable response:

\*\*Overview:\*\*

The press release from S&P Global Mobility discusses the anticipated mild recovery in February 2024 U.S. auto sales, interpreting figures like the projected SAAR and total unit sales. It touches on both positive trends (improved inventory, pricing) and challenges (economic uncertainty, high interest rates).

\*\*Guidelines Analysis:\*\*

- Identify the narrative: The release suggests optimism with a projected increase in sales, balancing this with the reality of market volatility and economic challenges.
- Evaluate the representation: The headline of improving sales may mask a nuanced reality where actual unit sales have barely grown year-over-year.
- Tactics for accurate understanding: Examine both seasonal and annual changes in detail, analyze comprehensive metrics like BEV market share changes, and cross-reference with broader economic indicators and independent analyses.

**\*\*Visual Presentation Plan:\*\***

- Slide 1: Introduction - Title, overview, objective
- Slide 2: Key points from the first press release, dissecting language, and data.
- Slide 3: Critique of the second press release's data framing.
- Slide 4: Methods for extracting true insight from such communications, summarized with bullet points.

Visual aids like the graphs provided should be incorporated to juxtapose the sales trends and BEV proportion increases against the narrative spun by the press release.

**Detailed Explanation:**

**\*\*Correctness:\*\***

Assistant A and B both provide accurate summaries and analyses of the overarching themes and data within the press release. However, B provides specific numeric comparison while Assistant A touches more on linguistic framing ("bounce mildly").

**\*\*Helpfulness:\*\***

Both assistants are useful, but B gives a more structured analysis using a mock PowerPoint layout which aligns closely with assignment instructions, making it easier for users to visualize the proposed presentation.

**\*\*Relevance:\*\***

Assistant B remains focused on the press release and does not deviate from the given instructions. Assistant A provides a broader context but includes additional information about S&P Global Mobility itself that could be seen as slightly tangential.

**\*\*Conciseness:\*\***

Assistant B presents key points succinctly, while Assistant A offers slightly more verbose explanations which might not be as necessary.

**\*\*Completeness:\*\***

Assistant A goes the extra mile in detailing a step-by-step presentation creation guide beyond what is requested, offering comprehensive PowerPoint slide suggestions. Assistant B keeps the response streamlined and fully focused on the data and its implications, covering the essentials

effectively.

Final Verdict:

[[B>A]]: Assistant B is slightly better due to its direct adherence to the presentation format, focused analysis, and clear articulation tailored to the assignment requirements.

**Judge (Model A: gemini-1.5-pro-002; Model B: gpt-4o-2024-05-13):**

Answer:

To effectively address the assignment prompt about ethical data usage and the impact of storytelling in press releases, the presentation of data, particularly from the S&P Global Mobility, must be analytical and reflective of underlying realities. The crucial elements include identifying the narrative conveyed, examining if the reality is better or worse than presented, and suggesting ways to discern the true progress.

**\*\*Ideal Response:\*\***

1. **\*\*Overview and Contextual Analysis:\*\***

- **\*\*Objective:\*\*** Understand how data storytelling shapes perceptions of progress.
- **\*\*Analysis of Press Release:** "February 2024 US auto sales to bounce mildly":
  - **\*\*Message Conveyed:\*\*** A projected "mild" recovery in February sales at 1.22 million units is framed optimistically against a slow January. However, the term "tepid" suggests a reserved outlook despite the statistical uptick.
  - **\*\*True Progress vs. Representation:\*\*** The focus on SAAR indicates better-than-immediate-previous performance, yet it masks year-on-year challenges (e.g., below February 2023 levels).

2. **\*\*Potential for Misleading Messaging:\*\***

- **\*\*SAAR Utilization:\*\*** Encourages an optimistic view, obscuring potential downturns in actual unit sales compared to past annual figures.
- **\*\*BEV Sales\*\*:** While steady at 8%, the growth has plateaued due to shifts in tax policy rather than underlying market expansion.

3. **\*\*Strategies for Grounded Interpretation:\*\***

- Compare SAAR to actual data for a clearer trend.
- Source contrasting automotive sales reports for diverse perspectives.
- Contextualize BEV trends relative to legislative impacts and infrastructural nuances.

**\*\*Specific Observations:\*\***

- **\*\*Correctness:\*\***
  - Both assistants accurately discuss the data, identifying the "mild bounce" narrative with SAAR focus.
- **\*\*Helpfulness:\*\***

- Assistant A provides detailed points and specific suggestions on deeper analysis techniques, while Assistant B offers a broader guideline for interpreting data storytelling but is somewhat generic.

- **\*\*Relevance:\*\***

- Assistant A's response is tightly focused on the press release specifics. Assistant B broadens the scope by analyzing a second press release, which adds volume but potentially dilutes focus given the single release requirement.

- **\*\*Conciseness:\*\***

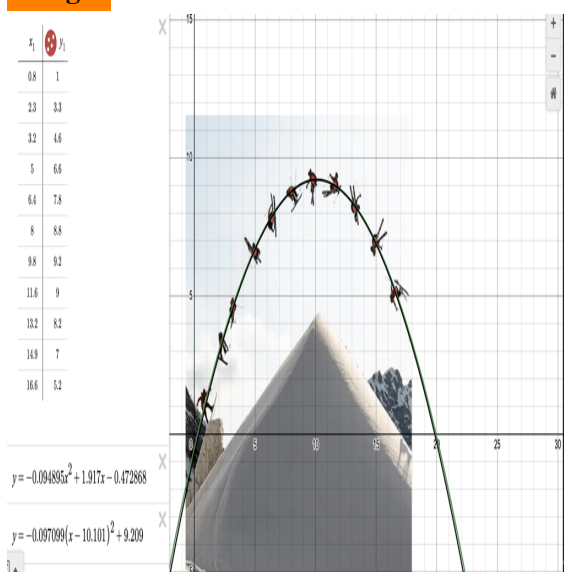
- Assistant B is slightly more concise but possibly less specific due to broader coverage across two releases.

- **\*\*Completeness:\*\***

- Assistant A covers the topics comprehensively relevant to the prompt's intent, providing detailed slides and suggestions.

Final Verdict:

- After thorough review, Assistant A provides a more directly applicable, detailed, and focused analysis in line with the task guidelines, whereas Assistant B offers a wider scope but lacks depth in specific critical areas. Therefore, **[[A>B]]**: Assistant A is slightly better.

**Image:****Question:**

**Inquiry:** Organize Data and Describe the Trend

<Introduce your context to the reader. Show a graph of the points you plotted along with your chosen images. Note key information like the independent and dependent variables and discuss how you set up your image in the first place before you started plotting points.

See the specific questions asked in the task to ensure you meet all requirements.>

**Action:** Create your Model

<Find the equation of a quadratic function that models each scenario. Show your work and explain how you decided on your parameters. Insert images to show your thinking and process. At the very least, you must have an image of your data and your model on the same graph. See the specific questions asked in the task to ensure you meet all requirements.>

**Reflection:** Analyze and Evaluate your Model

<In this section, note that to “justify” means giving a balanced perspective. If you’re justifying your accuracy, what calculations could you do to show how accurate your work is? What could be some of the errors you might have made? Showing awareness of ways your work is and might not be accurate shows you have considered all of this as you worked on your model. How do you know your model makes sense? What can you calculate or what kinds of comparisons can you make that help you know what your model function shows is more or less correct?

Describe a specific modification to one or more of the parameters and the reasons why this specific modification would improve your model. See the specific questions asked in the task to ensure you meet all requirements.>

**SUGGESTED RESPONSE PROMPTS**

<Introduction –think of something to say that will introduce your task to the reader. What is this task about? What will you discuss? What can this help us to understand? Do it in your own way, or use the sentence frames, below>





on a Cartesian coordinate system and two different quadratic functions modelled to fit these points. The graph includes an overlaid image of a snowboarder in mid-jump, presumably to provide a real-world context for the data points and the mathematical model.

In the graph:

- The independent variable ( $x$ ) represents distance.
- The dependent variable ( $y$ ) represents height.
- The table on the left lists distances (in what are likely metres) and corresponding heights.
- Two quadratic equations are given:
  1.  $y = -0.094895x^2 + 1.917x - 0.472868$
  2.  $y = -0.097099(x - 10.101)^2 + 9.209$

These equations are different forms of a quadratic function: the first is in standard form, and the second is in vertex form. The vertex form indicates that the peak of the jump (the highest point the snowboarder reaches) is at approximately (10.101, 9.209), which means the snowboarder reaches a height of roughly 9.209 metres at a distance of about 10.101 metres from the starting point.

The trend shown by the plotted points suggests that as the snowboarder moves away from the starting point, the height initially increases to a peak and then decreases, which is characteristic of the path taken during a jump. The quadratic model appears to fit the data points well, suggesting a parabolic trajectory, which is typical for projectile motion like a snowboard jump.

Regarding the task instructions given:

- The data and model are already on the same graph, as requested.
- To improve accuracy, one could refine the measurements, consider wind resistance, or take more data points especially near the peak to get a more precise model.
- If I were to justify the accuracy of the model, I would calculate the residuals (the differences between the observed heights and those predicted by the model) and aim for them to be as small as possible.

Possible corrections to the model might include adjusting the coefficients to better fit the data points if there's evidence that the current model is not the best fit. For example, if the real-life scenario indicates that the snowboarder's peak height is slightly different than what the model suggests, the parameters could be tweaked to align with the observed data more closely.

#### **Molmo-7B-O-0924:**

The purpose of this task is to compare the height and distance of a snowboard jump to a ski jump. I will explore these ideas by creating a graph in Desmos, placing images of the jumps at key points, and analyzing the relationship between the two variables.

The independent variable is the distance and the dependent variable is the height. The domain of the relationship is  $D = \{x | x \in [0, 30]\}$ , because this

represents the range of distances shown in the images. The range of the relationship is  $R=\{y|y\in[0,20]\}$ , because this represents the range of heights shown in the images.

I chose to place my image where the end of the snowboarders board lies on the origin because this is the starting point of both jumps. I scaled my image appropriately by finding the average height of an American male. I know this process works because the resulting graph shows a reasonable relationship between distance and height.

The trend/relationship between distance and height appears to be that the height increases as the distance increases, but at a decreasing rate. This trend is consistent with the laws of physics, as objects gain more potential energy when they fall from greater heights.

One possible reason this trend is happening is because the snowboarder gains more air time as they travel a greater distance, allowing them to reach a higher maximum height. The ski jumper, starting from a greater height, reaches a lower maximum height but travels a greater distance.

Action: Create your Model

A quadratic function is a good choice of function type for this context because it can model both increasing and decreasing trends, and it can represent the parabolic shape of the jump curve.

I chose to use  $y=ax^2+bx+c$  form for my quadratic model equation because this is the standard form of a quadratic function. I will use the values from my graph to determine the coefficients  $a$ ,  $b$ , and  $c$ .

<Find the equation of a quadratic function that models each scenario. Show your work and explain how you decided on your parameters. Insert images to show your thinking and process. At the very least, you must have an image of your data and your model on the same graph>

Reflection: Analyze and Evaluate your Model

I took many steps to make sure this model was accurate. For example, I carefully plotted points on my graph to ensure they represented the data accurately. This step provides accuracy because it allows me to visually verify the relationship between distance and height.

There are still some issues with the accuracy of this model, such as the fact that the height doesn't quite reach 20 units at the maximum distance. This issue caused my work to be less accurate because the model doesn't perfectly match the data at the highest points.

I know my results make sense because the graph shows a reasonable relationship between distance and height, and the values appear to be consistent with what we'd expect from physics. The modeling process helped me to better understand my context because it forced me to think critically about the relationship between distance and height in these jumps.

I would improve my model by adjusting the coefficient  $a$  to better match the overall trend of the data. For example, I might try a value of  $a=0.5$  to make the curve steeper, which would better represent the rapid increase in height at the beginning of each jump.

**Judge (Model A: gpt-4o-2024-05-13; Model B: Molmo-7B-O-0924):**

Answer:

The task involves analyzing a snowboard jump using a quadratic model to describe the relationship between distance and height. The image shows a graph with plotted points and two quadratic equations modeling the jump's trajectory.

Detailed Explanation:

1. **\*\*Correctness:\*\***

- **\*\*Assistant A\*\*** accurately describes the graph, identifying the independent and dependent variables and explaining the quadratic equations provided. The analysis of the parabolic trajectory is correct.
- **\*\*Assistant B\*\*** provides a general overview but makes assumptions about the domain and range without specific data from the image. The explanation of the trend is less precise.

2. **\*\*Helpfulness:\*\***

- **\*\*Assistant A\*\*** offers a clear explanation of the graph and suggests ways to improve the model's accuracy, such as calculating residuals.
- **\*\*Assistant B\*\*** attempts to explain the setup and scaling but lacks specific details about the equations and their derivation.

3. **\*\*Relevance:\*\***

- **\*\*Assistant A\*\*** stays focused on the task, discussing the graph, equations, and potential improvements.
- **\*\*Assistant B\*\*** includes some extraneous information about ski jumps, which is not directly relevant to the snowboard jump analysis.

4. **\*\*Conciseness:\*\***

- **\*\*Assistant A\*\*** is concise and to the point, providing necessary details without unnecessary elaboration.
- **\*\*Assistant B\*\*** is more verbose, with some repetition and less focus on the specific task requirements.

5. **\*\*Completeness:\*\***

- **\*\*Assistant A\*\*** covers all aspects of the task, including data organization, model creation, and reflection on accuracy.
- **\*\*Assistant B\*\*** lacks detailed explanation of the quadratic model creation and does not show the work for deriving the equations.

Specific Observations:

- **\*\*Assistant A\*\*** effectively uses the image to explain the quadratic model and suggests practical steps for improvement.
- **\*\*Assistant B\*\*** provides a broader context but lacks specific details and

calculations related to the model.

Final Verdict:

- **[[A>B]]**: Assistant A is significantly better.

**Judge (Model A: Molmo-7B-O-0924; Model B: gpt-4o-2024-05-13):**

Answer:

The task involves analyzing the trajectory of a snowboard jump using a quadratic model. The goal is to understand the relationship between distance and height during the jump, represented by a parabolic curve.

Detailed Explanation:

- **\*\*Introduction and Context\*\***: Both assistants introduce the task, but Assistant A provides a more structured introduction, explaining the purpose and context of the task. Assistant B focuses on describing the image and data setup.

- **\*\*Data Organization and Trend Description\*\***:

- **\*\*Assistant A\*\***: Clearly identifies the independent (distance) and dependent (height) variables, and discusses the domain and range. It also explains the placement and scaling of the image.

- **\*\*Assistant B\*\***: Describes the graph and equations but lacks a detailed explanation of the domain, range, and scaling process.

- **\*\*Model Creation\*\***:

- **\*\*Assistant A\*\***: Chooses a quadratic model and explains the reasoning behind using the standard form. However, it does not provide the specific equation derived from the data.

- **\*\*Assistant B\*\***: Lists two quadratic equations from the image and explains their forms, but does not discuss the process of deriving these equations.

- **\*\*Model Analysis and Evaluation\*\***:

- **\*\*Assistant A\*\***: Discusses steps taken for accuracy and acknowledges potential inaccuracies. It suggests modifying the coefficient to improve the model.

- **\*\*Assistant B\*\***: Mentions calculating residuals for accuracy but does not provide specific steps or modifications to improve the model.

Specific Observations:

- **\*\*Correctness\*\***: Both assistants correctly identify the quadratic nature of the model. Assistant B provides the equations directly from the image.

- **\*\*Helpfulness\*\***: Assistant A offers more guidance on setting up and analyzing the model, while Assistant B focuses on describing the existing graph.

- **\*\*Relevance\*\***: Assistant A stays more focused on the task requirements, while Assistant B provides a broader description of the image.

- **\*\*Conciseness\*\***: Assistant A is more concise in addressing the task requirements, whereas Assistant B includes more descriptive elements.

- **\*\*Completeness\*\***: Assistant A covers all aspects of the task, including model creation and evaluation, while Assistant B lacks detailed analysis and

improvement suggestions.

Final Verdict:

[[A>B]]: Assistant A is slightly better because it provides a more comprehensive and structured approach to the task, covering all required aspects effectively.

961