First Multi-Dimensional Evaluation of Flowchart Comprehension for Multimodal Large Language Models

Anonymous ACL submission

Abstract

With the development of Multimodal Large Language Models (MLLMs) technology, its general capabilities are increasingly powerful. To evaluate the various abilities of MLLMs, numerous evaluation systems have emerged. But now there is still a lack of a comprehensive method to evaluate MLLMs in the tasks related to flowcharts, which are very important in daily life and work. We propose the first comprehensive method, FlowCE, to assess MLLMs across various dimensions for tasks related to flowcharts. It encompasses evaluating MLLMs' abilities in Reasoning, Localization Recognition, Information Extraction, Logical Verification, and Summarization on flowcharts. However, we find that even the GPT4o model achieves only a score of 56.63. Among open-source models, Phi-3-Vision obtained the highest score of 49.97. We hope that FlowCE can contribute to future research on MLLMs for tasks based on flowcharts.

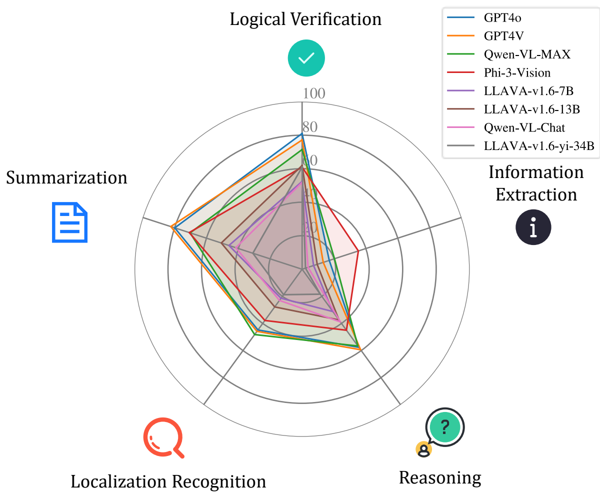


Figure 1: Evaluation results of multimodal large language models on five dimensions of tasks in FlowCE. GPT-4o achieves the highest overall score of 56.63.

1 Introduction

In the modern work environment, flowcharts have become a widely used graphical tool across various industries and fields. Flowcharts provide an intuitive and efficient way to describe and analyze workflows. By representing processes graphically, complex workflows can be simplified into easily understandable steps, thereby facilitating a range of tasks. Currently, leveraging Multimodal Large Language Models (MLLMs) for the understanding and analysis of flowcharts has become a research focus. Represented by models like GPT-4v (Achiam et al., 2023), these large models can comprehend user-input images and perform corresponding question-and-answer tasks. Meanwhile, there have been numerous open-source efforts for MLLMs, such as LLAVA-1.6v (Liu et al., 2023a), QWEN-VL (Bai et al., 2023b), MiniCPM (Hu et al., 2024), phi-3-vision (Abdin et al., 2024), and CogVLM2 (Wang et al., 2023). To evaluate the cross-modal understanding capabilities of existing MLLMs between images and text, various benchmarks have emerged, including MMBench (Liu et al., 2023b), MME (Yin et al., 2023), TextVQA (Singh et al., 2019), MM-Vet (Yu et al., 2023), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), InfographicQA (Mathew et al., 2022), FlowChartQA (Tannert et al., 2023) and so on. Additionally, these evaluation systems measure the capabilities of MLLMs from different perspectives, including the understanding of general images, document-type images, chart-type images, and more.

However, to the best of our knowledge, none of these existing evaluation benchmarks comprehensively assess MLLMs' understanding of flowcharts from multiple perspectives in real-world scenarios. This hinders the development of methods for utilizing MLLMs to understand and analyze flowcharts in open environments. Thus, inspired by previous works such as FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2020), ChartQA

Benchmark	Capability	Real-world Data	Handcrafted Questions	Answer Type	Size	# models
LVLM-eHub	General Multi-Modality	✓	▲	MC/OE	332k	8
MME	General Multi-Modality	✓	✓	MC	2,194	10
MMBench	General Multi-Modality	✓	▲	MC	2,974	14
TextVQA	Text Recognition and Contextual Reasoning	✓	✓	OE	45.3k	6
InfographicVQA	Integrated Document Visual and Textual Reasoning	✓	✓	OE	30k	1
ChartQA	Chart Understanding and Analysis	✓	✓	OE	9.6k	4
EgoThink	First-Person Thinking	✓	✓	OE	700	21
MathVista	Mathematical Reasoning	✓	✓	MC	6141	11
FlowchartQA	Geometirc and Topological Information of Flowcharts	✗	✗	MC	6M	1
FlowCE(ours)	Comprehensive Understanding of Flowcharts	✓	✓	OE	505	19

Table 1: Comparison of recent comprehensive evaluation benchmarks of MLLMs and our proposed benchmark FlowCE. MC/OE indicate multi-choice and open-ended question-answering respectively. "▲" indicates that there are both handcrafted questions and questions generated using templates.

(Masry et al., 2022) and FlowchartQA (Tannert et al., 2023), and motivated by the successful development of MLLMs. We propose a novel benchmark, FlowCE, for the first time: comprehensively assessing the understanding capabilities of multimodal large language models on flowcharts in real-world scenarios. FlowCE evaluates the understanding capabilities of existing MLLMs on flowcharts from multiple dimensions, including Reasoning, Information Extraction, Localization Recognition, Summarization, and Logical Verification. We have carefully designed diverse question-answer pairs for various dimensional tasks in open environments. Additionally, the flowchart images in FlowCE are sourced from a variety of real-world scenarios and styles. We have carefully designed diverse question-answer pairs for various dimensional tasks in open environments. Additionally, the flowchart images in FlowCE are sourced from a variety of real-world scenarios and styles.

We conducted evaluations on all mainstream MLLMs, both open-source and proprietary, using FlowCE. The evaluation results for some parts on FlowCE are shown in Figure 1. The results indicate that even the highly performant GPT4o achieves only a score of 56.63, with the best performance among open-source models being achieved by Phi-3-Vision (Abdin et al., 2024), scoring 49.97. Our main contributions are as follows:

- We introduce FlowCE to comprehensively evaluate the understanding capabilities of MLLMs on flowcharts. It encompasses evaluation tasks and methodologies across dimensions such as reasoning, information extraction, localization recognition, summarization, and logical verification.

- We extensively evaluate mainstream open-source and proprietary models using FlowCE. Through detailed analysis of these MLLMs' performance across different dimensional tasks, we uncovered their strengths and limitations in understanding flowcharts. Additionally, we proposed some improvement suggestions for existing models to facilitate future research and development.

- We are open-sourcing our resources to foster future advancements in this field.

2 Related Work

2.1 Multimodal Large Language Models

Inspired by the remarkable success of LLMs such as internVL (Chen et al., 2024), llama3 (Touvron et al., 2023), Yi-chat (Young et al., 2024), Qwen (Bai et al., 2023a), and Vicuna (Zheng et al., 2024), recent MLLMs have incorporated these advanced LLMs as their primary backbone. Examples include the LLAVA-V1.6 (Liu et al., 2024b) series, ShareGPT4 (Chen et al., 2023) series, Qwen (Bai et al., 2023b) series, Cogvlm (Wang et al., 2023) series and so on. Initially, MLLMs leverage vast datasets consisting of image-text pairs (Alayrac et al., 2022; Zhu et al., 2024) or an arbitrarily combination of visual and textual data for pre-training (Li et al., 2023; Liu et al., 2024a). Moreover, the availability of extensive image-text instruction datasets facilitate recent studies (Dai et al., 2024; Li et al., 2023; Liu et al., 2024c; Ye et al., 2023; Chen et al., 2023) to implement instruction tuning. This fine-tuning process enhances the ability of MLLMs to produce high-quality responses. This two-phase training strategy (Li et al., 2022; Yang et al., 2022) enables recent MLLMs to achieve outstanding performance in downstream vision-language tasks (Antol et al., 2015; Hudson and Manning, 2019; Lin et al., 2014; Plummer et al., 2015).

2.2 Benchmarks for MLLMs

To evaluate the capabilities of Vision-Language Models (MLLMs), various downstream language tasks are employed. General benchmarks, such as MMBench (Liu et al., 2023b), MME (Yin et al., 2023), and LVLM-ehub (Xu et al., 2023), provide a comprehensive assessment of model performance. Domain-specific benchmarks, such as TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2021), evaluate the fine-grained transcription capabilities of MLLMs on low-resolution images. MathVista (Lu et al., 2024) examines the ability of MLLMs to integrate visual and mathematical logic. ChartQA (Masry et al., 2022) aims to evaluate direct chart understanding and analysis, while InfographicQA (Mathew et al., 2022) addresses logical questions about data visualizations and charts. EgoThink (Cheng et al., 2024) elaborate on the capabilities of MLLMs to think from a first-person perspective. General benchmarks offer a broad and consistent evaluation framework (Xu et al., 2023; Yin et al., 2023; Liu et al., 2023b), whereas domain-specific benchmarks enable detailed assessment of model capabilities and promote advancements in specific research areas.

In Table 1, we compare FlowCE with various existing benchmarks. FlowCE comprehensively assesses for the first time the ability of MLLMs to understand flowcharts. Specifically, compared to FlowchartQA (Tannert et al., 2023), we not only introduce tasks across more dimensions but also create real-world flowchart data and open-scenario question-answer pairs.

3 FlowCE

In this section, we elaborate first on the evaluation tasks across various dimensions in FlowCE. Then, we introduce the process of manually constructing diverse open-scenario question-answer pairs. Finally, we present the evaluation methodologies for tasks across different dimensions.

3.1 Tasks across different dimensions

As shown in Figure 2, we establish tasks across five dimensions in real flowchart scenarios, including reasoning, information extraction, localization recognition, summarization, and logical verification, for quantitative evaluation.

Logical Verification Upon receiving a process diagram, users provide the logical relationships between different nodes or boxes in the diagram, and MLLMs are tasked with evaluating these relationships. Figure 2(a) shows an example of Logical Verification. This process entails a comprehensive analysis of the structure and content of the process diagram to ensure the accuracy and coherence of the logical relationships. MLLMs assess whether the provided process logic aligns with the actual scenario by considering the interactions among individual nodes and their roles throughout the entire process.

Information Extraction The task entails MLLMs receiving flowchart images and extracting corresponding textual information based on user queries. We have categorized the questions into two main types based on the content of the flowchart: the first type involves prompting MLLMs to extract all textual information from the flowchart, while the second type entails extracting specific textual information based on the characteristics of the flowchart. An example of Information Extraction is shown in Figure 2(b).

Localization Recognition Users will inquire about the positional relationships between different nodes or boxes in the flowchart, an example of Localization Recognition is illustrated in Figure 2(c), thereby assessing whether MLLMs have an accurate understanding of the positional relationships of nodes and boxes in the flowchart.

Reasoning For an example of Reasoning, as shown in Figure 2(d), the task refers to MLLMs making decisions in response to user inquiries based on the content of the flowchart images after being provided with them. Here, we formulate more natural questions based on the content of the flowchart, which require judgment and reasoning considering aspects such as conditional relationships within the flowchart to answer, rather than relying solely on the direction of the arrows in the flowchart.

Summarization MLLMs provide a summarized abstraction of the content depicted in process diagrams, elucidating the conveyed information. They accomplish this task by analyzing the logical relationships among various nodes within the diagram, identifying key steps and critical information, and integrating them into a concise yet comprehensive summary. Through understanding and encapsulating the process diagram, MLLMs generate the primary flow of the process and key decision points, thereby assisting users in better comprehending the process or system represented by the diagram, as shown in Figure 2(e).

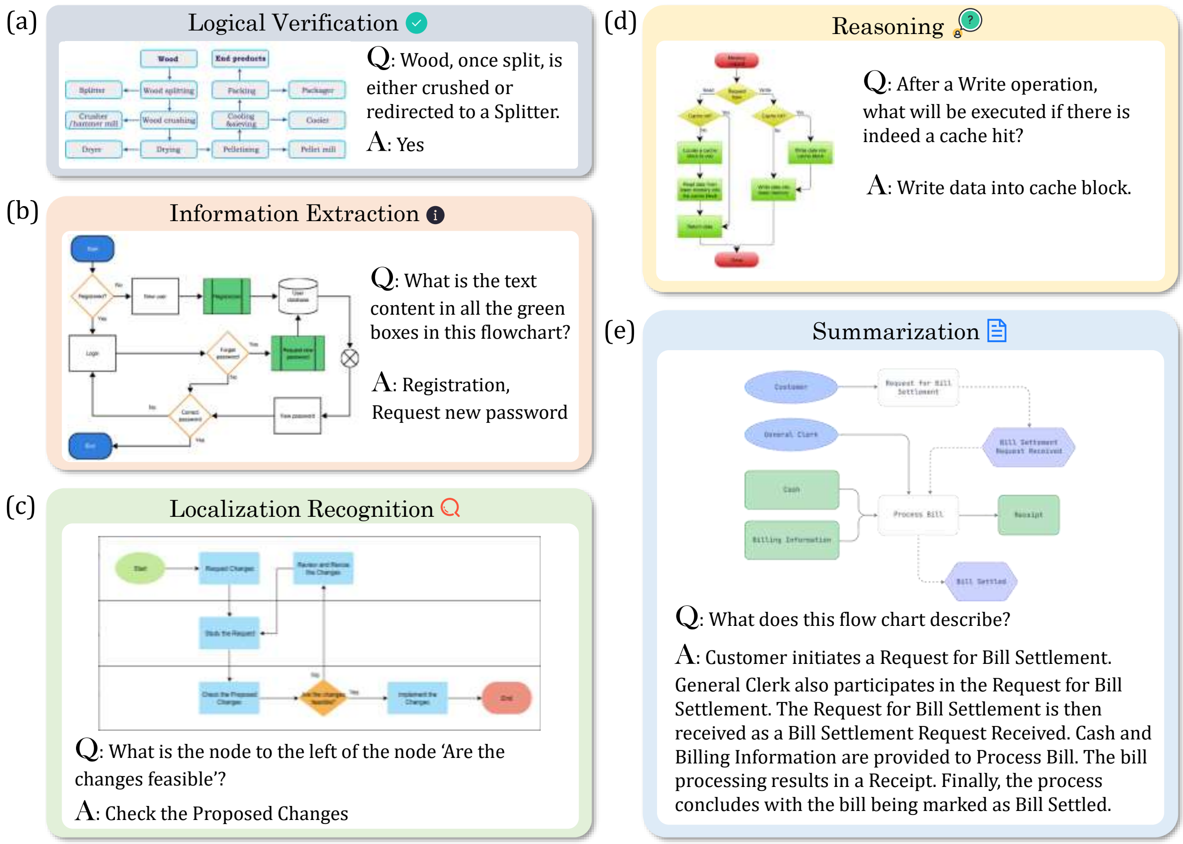


Figure 2: Data samples of FlowCE, which covers 5 evaluation dimensions. Each evaluation dimension contains human-annotated question-answer pairs.

3.2 Data construction

In this section, we introduce the data of FlowCE and elaborate on the detailed process of constructing FlowCE.

FlowCE-data FlowCE is built upon 500 real-world flowcharts, ensuring an ample diversity in each chart. In Figure 3, we present a detailed breakdown of the category distribution within the flowchart, encompassing categories from daily life, various specialized filed flowcharts, coding flowcharts, mathematical flowcharts, and others.

Human-annotated To ensure an open-ended question-and-answer format, we manually constructed question-answer pairs for each flowchart. We assigned different dimensions of tasks to the same individual to annotate a particular type of question, ensuring consistency in the tasks. Additionally, to allow for greater diversity in open-ended question-and-answer scenarios, we leveraged powerful GPT-like models for auxiliary construction, aiding humans in exploring more imaginative possibilities. Please refer to the Appendix A for specific details.

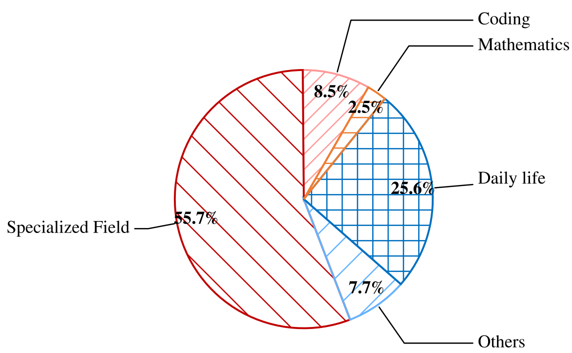


Figure 3: Distribution of Different Types of Flowcharts

3.3 Evaluation method

In this section, the evaluation of various tasks quantification methods will be introduced.

Automatic evaluation For tasks involving open-ended question answering, such as reasoning, localization recognition, and summarization, we employ GPT4 to assess the semantic similarity between standard answers and the responses generated by MLLMs. For detailed methodology of the evaluation utilizing GPT4, please refer to the Appendix B.

Models	Image Encoder	LLM	Alignment Module	ToP	Dataset Size	
					Pretraining	Finetuning
API-based (Proprietary) Models						
GPT4o (Achiam et al., 2023)						
GPT4V (Achiam et al., 2023)			/			
Qwen-VL-MAX (Bai et al., 2023b)						
3.4B~7B Models						
MiniCPM-V2 (Hu et al., 2024)	SigLip-400M	MiniCPM-2.4B	RLHF-V (Yu et al., 2024a)	3.43B	/	/
Phi-3-Vision (Abdin et al., 2024)	CLIP-ViT-L-336px	phi-3-mini-128K-instruct	SFT+DPO	4.2B	0.5T	15B
LLAVA-V1.5-7B (Liu et al., 2024a)	CLIP-ViT-L-336px	Llama2-7B	MLP	7.1B	558K	665K
ShareGPT4V-7B (Chen et al., 2023)	CLIP-ViT-L-336px	Vicuna-7B	MLP	6.7B	1.2M	665K
LLAVA-V1.6-7B (Liu et al., 2024b)	CLIP-ViT-L-336px	Vicuna-7B	Linear	7.06B	558K	760K
8B~13B Models						
LLAVA-Llama3-8B (Contributors, 2023)	CLIP-ViT-L-336px	Llama3-8B-Instruct	MLP	8.03B	558K	665K
MiniCPM-Llama3-V2.5 (Hu et al., 2024)	SigLip-400M	Llama3-8B-Instruct	RLAIF-V (Yu et al., 2024b)	8.54B	/	/
Qwen-Chat-VL (Bai et al., 2023b)	Open-CLIP-bigG	Qwen-7B	Cross-Attention	9.6B	1.4T	76.8M
LLAVA-V1.5-13B (Liu et al., 2024a)	CLIP-ViT-L-336px	Llama2-13B	MLP	13.3B	558K	665K
ShareGPT4V-13B (Chen et al., 2023)	CLIP-ViT-L-336px	Vicuna-13B	MLP	12.58B	1.2M	665K
LLAVA-V1.6-13B (Liu et al., 2024b)	CLIP-ViT-L-336px	Vicuna-13B	Linear	13.3B	558K	760K
13B~Models						
Cogvlm-Chat (Wang et al., 2023)	EVA2-CLIP-E	CogVLM-17B	Visual Expert	17.6B	1.5B	/
Cogvlm2-Llama3-Chat-19B (Wang et al., 2023)	EVA2-CLIP-E	Meta-Llama-3-8B-Instruct	Visual Expert	19.5B	/	/
LLAVA-Internlm2-Chat-20B (Contributors, 2023)	CLIP-ViT-L-336px	InternLM2-Chat-20B	deepspeed finetuning	20B	595K	150K
LLAVA-Next-Yi-34B (Liu et al., 2024b)	CLIP-ViT-L-336px	Nous-Hermes-2-Yi-34B	Linear	34.8B	558K	760K
Yi-VL-34B (Young et al., 2024)	CLIP-ViT-L-336px	Yi-34B-Chat	MLP	34B	3.1T	1.25M

Table 2: Statistics of compared API-based and open-source MLLMs, where ToP indicates Total Parameters and '/' indicates no detailed information for now.

Accuracy calculation Firstly, for the logical verification task, we match the output of MLLMs, either *"Yes"* or *"No,"* with the standard answers to calculate the accuracy after all questions have been answered, thereby quantifying the score of MLLMs on this task. Next, for the information extraction task, we propose a method based on effective factor to fairly compare the content generation effectiveness of different MLLMs. Then, for the information extraction task, we propose a method based on the effective factor to fairly compare the performance of different MLLMs in generating content. Suppose the label set is given by label $= [\text{text}_1, \text{text}_2, \ldots, \text{text}_n]$, where text_n represents the n-th text. The output answers are given by prediction $= [\text{pre}_1, \text{pre}_2, \ldots, \text{pre}_m]$, where pre_m is the m-th predicted text. If there is a predicted text in prediction that does not exist in label, and there are t such texts ($t \geq 1$), then the effective factor δ changes according to the following formula:

$$\delta = \delta^t,$$

At this point, if there is a predicted text in prediction that exists in label, then the initial score s changes as follows:

$$s = s \cdot \delta,$$

If $t = 0$, then for each predicted text in prediction that exists in label, the score remains the initial score s. Suppose there are i predicted texts that exist in label, the total score is $s \cdot i$. The product of the number of texts in label and the initial score is denoted as a. The quantitative score for evaluating MLLMs on this task is given by:

$$\text{Score} = \frac{s \cdot i}{a}(\%).$$

4 Experiments

4.1 Experimental setups

We conduct experiments on existing mainstream MLLMs, including both proprietary and open-source models. The parameter sizes of the open-source models range from 3.4B to 7B, 8B to 13B, and above 13B. In Table 2, we provide a detailed overview of these evaluated models in our experiments.

We employ GPT-4 as the adjudicator for LLMs to assign evaluation scores, with a focus on semantic similarity between standard answers and MLLM model outputs. Our evaluations adhere to a protocol: for reasoning and localization recognition tasks, we set the score range per question from 0 to 5. For summarization tasks, the score range per question is from 1 to 10. In the evaluation of

Models	FlowCE					Average
	LV	IE	RS	LR	SM	
API-based Models						
GPT4o	**83.81**	17.04	57.60	44.80	79.90	56.63
GPT4V	77.14	12.94	**59.40**	45.80	**82.40**	55.54
Qwen-VL-MAX	72.38	**20.32**	56.60	**48.20**	70.25	53.55
3.4B~7B Models						
MiniCPM-V2	51.43	7.00	30.00	22.00	50.20	32.13
Phi-3-Vision	**60.95**	**35.30**	**45.00**	**37.80**	**70.80**	49.97
LLAVA-V1.5-7B	53.33	4.90	14.40	18.20	35.60	25.29
ShareGPT4V-7B	50.48	3.72	12.20	16.80	33.60	23.36
LLAVA-V1.6-7B	52.38	7.20	31.20	21.40	45.90	31.62
8B~13B Models						
LLAVA-Llama3-8B	55.24	8.04	21.20	20.80	33.20	27.70
MiniCPM-Llama3-V2.5	58.10	**12.25**	**45.20**	**42.80**	17.20	35.11
Qwen-Chat-VL	50.48	3.73	38.80	23.00	41.60	31.52
LLAVA-V1.5-13B	53.33	5.36	22.60	22.20	40.50	28.80
ShareGPT4V-13B	53.33	4.46	22.20	16.60	41.50	27.62
LLAVA-V1.6-13B	**62.86**	9.47	37.40	27.80	**50.70**	37.65
13B~Models						
Cogvlm-Chat	50.48	0.34	34.80	29.60	53.20	33.68
Cogvlm2-Llama3-Chat-19B	57.14	4.70	44.60	**37.20**	**74.30**	43.59
LLAVA-Internlm2-Chat-20B	59.05	5.69	15.40	19.00	41.90	28.21
LLAVA-Next-Yi-34B	**60.95**	**12.21**	**51.20**	34.20	63.10	44.33
Yi-VL-34B	**60.95**	2.14	18.40	18.80	30.90	26.24

Table 3: Detailed evaluation results on FlowCE across different models, where "LV" stands for Logical Verification, "IE" for Information Extraction, "RS" for Reasoning, "LR" for Localization Recognition, and "SM" for Summarization. **Bold font** indicates the best performance in the same category, while underlined font indicates the second-best performance in the same category. Red indicates the highest average score among all API-based models. Blue indicates the highest average score among all open-source models.

information extraction tasks, we set the score s as 2, with an effective factor δ of 0.8.

4.2 Evaluation results

We extensively evaluate open-source MLLMs models at different parameter levels and mainstream commercial MLLMs models. All detailed evaluation results are presented in Table 3. Despite significant advancements in MLLMs in recent years, they still struggle to demonstrate understanding of flowcharts, including GPT-4o. Across five different task dimensions, only the summarization task achieves relatively high scores, peaking at 82.40 points in closed-source models. However, this is only demonstrated in closed-source models; in open-source models, the highest score reaches only 74.30 points. The highest score attained in the information extraction task is only 35.30 points, while in the reasoning task, it reaches a maximum of 59.40 points. In the localization recognition task, the highest score is 48.20 points. Even under random guessing with a score of 50.00 points in the logic validation task, the highest score reaches only 83.81 points. Among all closed-source models, GPT4o demonstrates superior overall capabilities compared to other models, but only excels in the logic validation task. Among all open-source models, Phi-3-Vision achieves the highest scores, surpassing closed-source models in the information extraction task. We will further elaborate on detailed assessments across different task dimensions. Additional cases can be found in the Appendix C.

Results of Information Extraction In this task, models generally obtain very low scores. The highest score of 35.30 is achieved by Phi-3-Vision, with the second-place model being the proprietary model Qwen-VL-MAX, but only scoring 20.32, indicating a significant gap. In Figure 4(a), for instance, by highlighting the inherent feature "pink ellipse" in the flowchart, MLLMs are enabled to extract corresponding information, with only Phi-3-Vision producing the correct answer. In Appendix F, to demonstrate the performance variation of different models in Information Extraction tasks, we conduct visual analysis based on effective factors. For example, Phi-3-vision achieves an average ef-

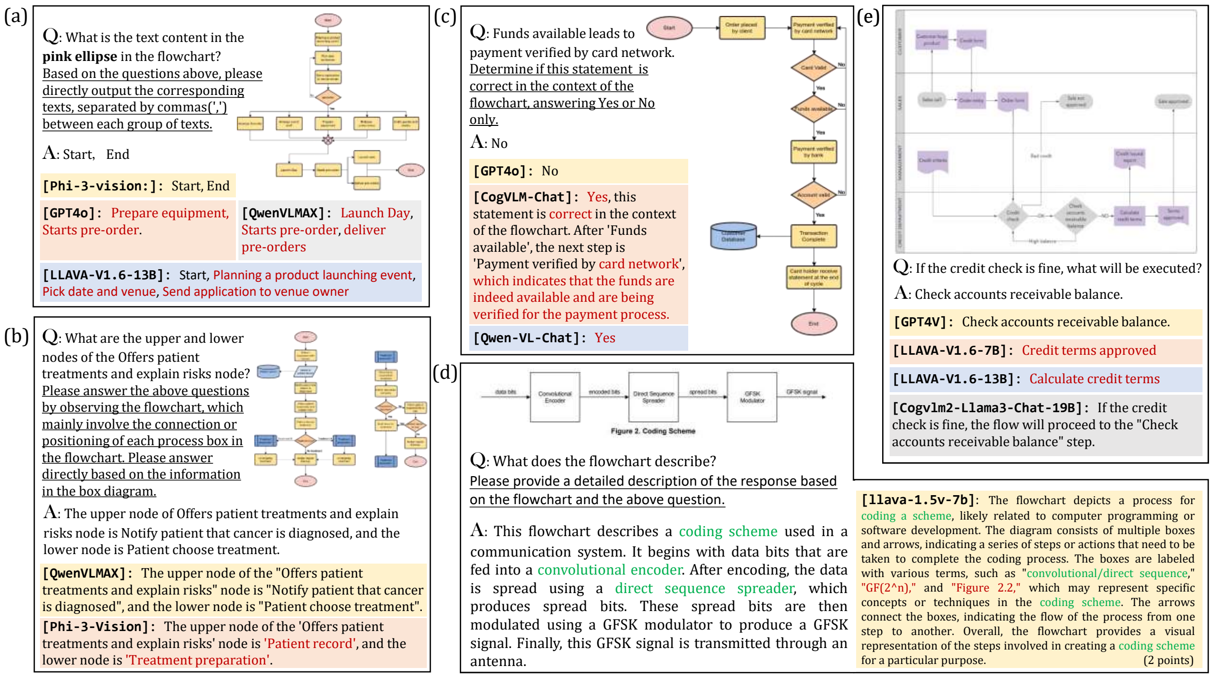


Figure 4: Some results from vary MLLMs. The words underlined indicate additional prompts. (a) showcases results on Information Extraction, (b) presents results on Localization Recognition, (c) showcases results on Logical Verification, (d) showcase results on Summarization, (e) displays results on Reasoning.

fective factor score exceeding 0.6.

Results of Localization Recognition The evaluation results of various models in this task indicate poor performance overall, with the top performer Qwen-VL-MAX scoring only 48.20 points. In Figure 4(b), detailed examples of Qwen-VL-MAX and Phi-3-Vision are presented. The response of Qwen-VL-MAX correctly identifies the upper and lower nodes of the *"Offers patient treatments and explain risks"* node as *"Notify patient that cancer is diagnosed"* and *"Patient choose treatment"* respectively. This indicates a clear understanding of the flowchart and the ability to accurately identify the relationships between different nodes. On the other hand, Phi-3-Vision incorrectly identifies the upper node as *"Patient record"* and the lower node as *"Treatment preparation."* This suggests that Phi-3-Vision struggled with accurately interpreting the connections between the nodes in the flowchart, leading to an incorrect answer.

Results of Logical Verification For this tasks, the open-source models LLAVA-V1.6-13B, Phi-3-Vision, LLAVA-Next-Yi-34b, and Yi-VL-34B have achieved the top two performances. Regarding the highest scoring model, GPT4o, as depicted in Figure 4(c), it exhibits concise and clear responses to questions with stronger instruction-following capabilities. Conversely, models such as CogVLM-Chat tend to generate more hallucinatory descriptions in their answers, leading to erroneous judgments. For instance, in the case of Qwen-Chat-VL, it outputs answers of the *"Unknown"* type, indicating a deficiency in instruction-following capability. In Figure 13 of Appendix E, we also analyze the predictive distributions of different models and visually compare them with the distribution of true labels. We find that the predictions of most models exhibit significant biases in this task. For example, ShareGPT4V-7B categorizes all results as correct. Only GPT4v, GPT4o, LLava-Next-Vicuna-13B, and Yi-VL-34B have prediction distributions that deviate from the actual results by no more than 15%. Additionally, these four models consistently rank in the top five in terms of performance.

Results of Summarization In proprietary models, the scores for this task are generally higher, with GPT4V achieving the highest score of 82.40. However, among open-source models, many still have relatively low scores. For example, MiniCPM-Llama3-V-2.5 only score 17.20, with only Phi-3-Vision, Cogvlm2-Llama3-Chat-19B, and LLAVA-Next-Yi-34B scoring above 60.00. In Figure 4(d), detailed example of LLAVA-1.5V-7B is presented. LLAVA-1.5v-7B, although detailed, provides an inaccurate and less focused response, meriting a score of 2.

Results of Reasoning GPT4V achieve the best score of 59.40, yet still below a satisfactory level. In Figure 4(e), we present examples of responses from GPT4V, LLAVA-V1.6-7B, LLAVA-V1.6-13B, and Cogvlm2-Llama3-Chat-19B regarding reasoning tasks. Cogvlm2-Llama3-Chat-19B provided a more detailed response by repeating the conditions from the question and then indicating the correct next step, which may aid in accurate reasoning. LLAVA-V1.6-7B and LLAVA-V1.6-13B both provide incorrect answers to this question.

5 Further analysis

In this section, we explore the impact of various factors on the FlowCE benchmark.

5.1 Model parameter volume

Among all open-source models, having a larger number of parameters does not necessarily lead to better performance. For instance, the 34B parameter models Yi-VL-34B and LLAVA-Next-Yi-34B scored only 26.24 and 44.33, respectively, while Phi-3-Vision, with only 4.2B parameters, achieved the best score among the open-source models. In Table 5, we compare the average performance across three parameter scales. Although there may be a trend of improvement with increasing model parameters, this is not a definitive conclusion.

5.2 Model data volume

In Table 2, we provide detailed information on the specific pre-training and fine-tuning data volumes for each model, and further analyze how the data sources impact the performance of model on FlowCE. Despite ShareGPT4V-13B utilizing a larger dataset, its performance still lags behind LLAVA-v1.5-13B, demonstrating that the quality of the dataset is paramount. Additionally, the selection and diversity of specific datasets play a crucial role. For instance, Phi-3-Vision leverages a 0.5T image-text paired dataset that includes FLD-5B, OCR-generated synthetic data, chart comprehension datasets, and plain text data (Xiao et al., 2024; Laurençon et al., 2024). These high-quality and diverse data sources have enabled Phi-3-Vision to achieve the highest score of 35.3 in the information extraction task on FlowCE, and furthermore, it ranks first in the overall score among open-source models.

Model Parameter	Score
3.4B~7B	32.47
8B~13B	31.40
13B~	**35.21**

Table 4: Average Scores on FlowCE for Different Parameter Levels

5.3 Consensus between Humans and Evaluators

In this section, we employ manual scoring evaluations for MLLMs' responses in Reasoning, Localization Recognition, and Summarization. The aim is to investigate whether the standards set by FlowCE and the use of GPT4 as an evaluator align closely with human assessment results. We engage five human evaluators to assess the model GPT4o, which emerges as the top-performing model overall. Additionally, we select the open-source model LLAVA-V1.6-13B for manual evaluation. The criteria and detailed results of the manual assessment can be found in the Appendix D. In Table 5, we present the Pearson correlation coefficients between human ratings and GPT4 scores under our answer setting. The results demonstrate a high degree of consistency between human evaluation and our assessment methodology, indicating that our FlowCE evaluation results can be regarded as effective assessments.

	RS	LR	SM
Correlation	0.97	0.97	0.91

Table 5: The Pearson correlation coefficient between human ratings and GPT4 scores for various tasks.

6 Conclusion

To evaluate the comprehension ability of MLLMs on flowcharts, we propose the first multi-dimensional evaluation method: FlowCE. FlowCE sets up five major categories of tasks, including reasoning, information extraction, localization recognition, logical verification, and summarization, aiming to thoroughly quantify the understanding capability and performance of MLLMs on flowcharts. The FlowCE framework not only provides an effective means to evaluate the comprehension ability of MLLMs on flowcharts, but also offers guidance for model optimization and improvement, thereby promoting the development of MLLMs.

Limitations

This work has two limitations. Firstly, it establishes the FlowCE benchmark based on flowcharts derived from a diverse set of 500 real-world images. While it poses challenges for existing closed-source and open-source models, continuous expansion of both the dataset size and the number of questions is necessary going forward. Secondly, FlowCE relies entirely on manual annotation for data generation. However, as the dataset grows, dependence on manual annotation introduces inherent limitations, making it difficult to completely eliminate errors from the data.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. 2024. Egothink: Evaluating first-person perspective thinking capability of vision-language models.

XTuner Contributors. 2023. Xtuner: A toolkit for efficiently fine-tuning llm.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Dingcheng Li, Zheng Chen, Eunah Cho, Jie Hao, Xiaohu Liu, Fan Xing, Chenlei Guo, and Yang Liu. 2022. Overcoming catastrophic forgetting during domain adaptation of seq2seq language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5441–5454.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco:

Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.

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*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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.

Simon Tannert, Marcelo G Feighelstein, Jasmina Bogojeska, Joseph Shtok, Assaf Arbelle, Peter WJ Staar, Anika Schumann, Jonas Kuhn, and Leonid Karlinsky. 2023. Flowchartqa: the first large-scale benchmark for reasoning over flowcharts. In *Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing*, pages 34–46.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Bo Yang, Lijun Wu, Jinhua Zhu, Bo Shao, Xiaola Lin, and Tie-Yan Liu. 2022. Multimodal sentiment analysis with two-phase multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2015–2024.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36.

A Image collection and manual annotation

To obtain flowchart images, we first conducted image searches using the keyword "flowchart" on search engine(Baidu Search), and then saved them. However, we encountered issues such as duplicates, low resolution, incomplete images, and other unrelated photos. Therefore, we proceeded to manually select images, resulting in the creation of a real-world dataset.

To ensure the construction of question-answer pairs in open scenarios, we use manual annotation for each flowchart. Additionally, to ensure the diversity of the question-answer pairs, we employ a powerful GPT-like model to assist with the generation. The annotation process is illustrated in Figure 5. Humans can choose to use GPT to generate basic diverse question-answer pairs, which are then modified as needed.

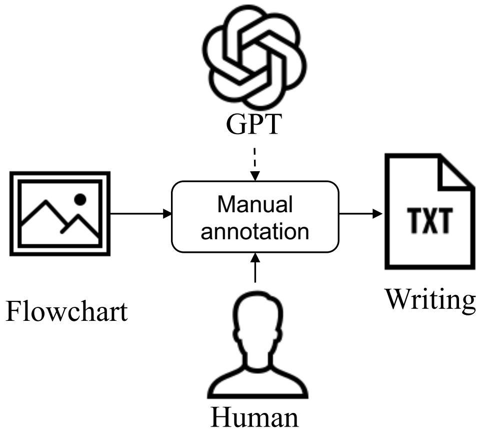


Figure 5: Manual annotation process, with optional assistance from GPT-like models for diversity construction.

B Large Language Model Judge

We use GPT-4 as an automated evaluator to score tasks in three categories: localization recognition, reasoning, and summarization. The scoring methodology is illustrated in Figure 6, where we set a score range of 0-5 for each question in the reasoning and localization recognition tasks, and a score range of 1-10 for the summarization tasks.

C Additional examples

We provide more examples in Figures 7, 8, 9, 10, 11 and 12, including the model's responses and the evaluation scores given by GPT-4.

D Manual Evaluation Protocol

We provide both the standard answers from FlowCE and the model's output answers. In Table 6, we present the detailed scores from the human evaluation.

RS and LR Tasks:

5 points: Answer is entirely correct with no factual errors.

4 points: Answer is mostly correct with minor factual errors that do not affect the main content.

3 points: Answer contains some factual errors but is still useful overall.

2 points: Answer has significant factual errors, with some content being incorrect.

1 point: Answer is severely flawed, with most content being incorrect.

0 points: Completely incorrect.

SM Task:

10 points: The answer is completely correct without any factual errors.

8 points: The answer is basically correct with only minor factual errors that do not affect the main content.

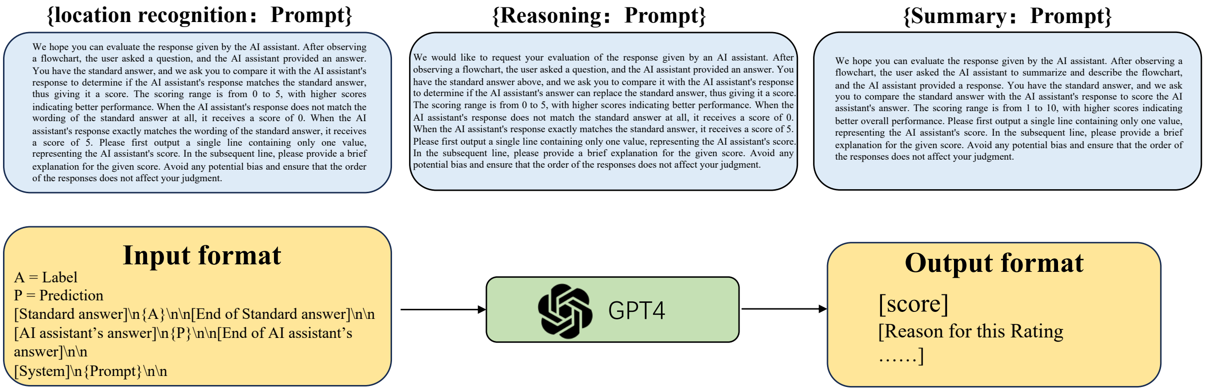


Figure 6: Using GPT-4 as an automated evaluator.

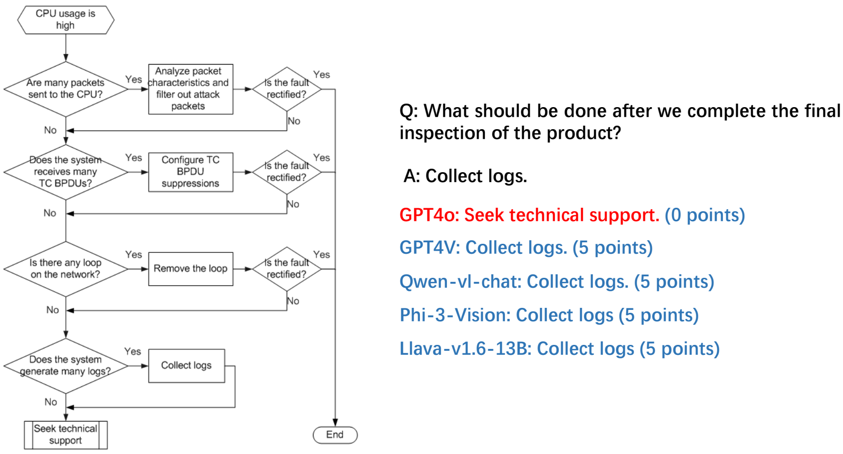


Figure 7: Model output and received scores.

6 points: The answer contains some factual errors but is still useful on the whole.
4 points: The answer has significant factual errors with incorrect parts of the content.
2 points: The answer is seriously incorrect with most of the content incorrect.
1 point: The answer is completely incorrect.

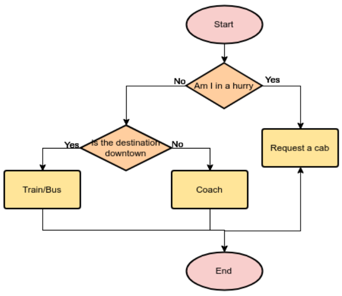
Model	GPT4-score			Human-score		
	RS	LR	SM	RS	LR	SM
GPT4o	57.6	44.8	79.9	62.6	58.4	75.9
LLAVA	37.4	27.8	50.7	45.8	36.8	47.6

Table 6: Performance Comparison between GPT4o and LLAVA.

E Detailed Comparison of Logical Verification Task

In this section, we present a detailed comparison of the predicted results for a subset of models: GPT4o, Phi-3-Vision, LLaVA-Next-Vicuna-13B, Qwen-Chat-VL, LLaVA-Next-Vicuna-7B, LLaVA-V1.5-13B, Cogvlm2-Llama3-Chat-19B, and Cogvlm-Chat. Each subplot in Figure 13 compares the predicted results (in blue) with the actual answer labels (in red) for each model. The score below each subplot indicates the overall performance of the model based on its accuracy in predicting the correct category.

GPT4o stands out with the highest accuracy, achieving a score of 83.81, indicating robust performance in aligning predictions with actual labels. Phi-3-Vision, while scoring 60.95, demonstrates a noticeable discrepancy in the *"No"* category with lower prediction accuracy. LLaVA-Next-Vicuna-13B, with a score of 62.86, shows moderate alignment but also exhibits substantial errors in the *"No"* category. Qwen-Chat-VL and Cogvlm-Chat, both scoring 50.48, indicate significant prediction errors and lower overall accuracy, particularly evident in the *"No"* and *"Unknown"* categories. LLaVA-Next-Vicuna-7B and LLaVA-V1.5-13B, scoring 52.38 and 53.55 respectively, also reflect moderate performance but with specific inaccuracies in the



Q: I'm in a hurry, how should I choose at this moment?
A: Request a cab.

GPT4o: Request a cab. (5 points)

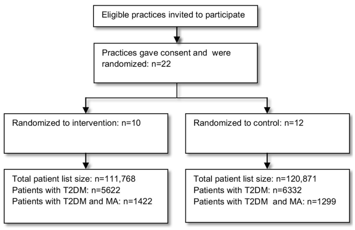
GPT4V: Request a cab. (5 points)

Qwen-vl-chat: Request a cab. (5 points)

Phi-3-Vision: Train/Bus (0 points)

Llava-v1.6-13B: Start (0 points)

Figure 8: Model output and received scores.



Q: What is the common previous step between Randomized to intervention: n=10 and Randomized to control: n=12?
A: Practices gave consent and were randomized: n=22

GPT4o: Practices gave consent and were randomized: n=22 (5 points)

Qwen-vl-chat: Eligible practices invited to participate. (0 points)

Llava-v1.6-13B: The common previous step between "Randomized to intervention: n=10" and "Randomized to control: n=12" is "Practices gave consent and were randomized: n=22". (2 points)

Figure 9: Model output and received scores.

"No" category. Cogvlm2-Llama3-Chat-19B, with a score of 57.14, shows better performance than some other models but still falls short in accurately predicting the *"No"* responses. These results suggest that while certain models like GPT4o exhibit strong performance, others require significant improvements in understanding and predicting both *"Yes"* and *"No"* categories accurately. The varying scores underscore the necessity for further refinement in training methodologies and model architectures to enhance predictive accuracy across all categories.

The phenomenon where some models exhibit a near 100% probability in answering *"Yes"* can be attributed to several factors:

- **Training Data Bias**: The training datasets may have an imbalance where affirmative answers (*"Yes"*) are disproportionately represented compared to negative ones (*"No"*). This bias in the training data can lead the models to favor *"Yes"* responses, as they learn to associate the affirmative answer with higher probabilities during the training process.

- **Model Overfitting**: Certain models might be overfitted to specific patterns in the training data, especially if those patterns predominantly involve affirmative responses. Overfitting can cause the model to generalize poorly to new, unseen data, resulting in a high likelihood of predicting *"Yes"* regardless of the actual context or question.

- **Algorithmic Tuning**: The hyperparameters and algorithmic settings of some models might be tuned in a way that inadvertently biases the model towards affirmative responses. This could include settings related to decision thresholds, loss functions, or other optimization parameters that skew the model's predictions towards *"Yes"*.

- **Lack of Contextual Understanding**: Some models may lack the nuanced understanding required to accurately discern between *"Yes"* and *"No"* in complex scenarios. This deficiency can lead them to default to a *"Yes"* answer, especially if they are not effectively capturing and processing the context of the queries.

- **Evaluation Metrics**: The evaluation metrics used during the training and validation phases might inadvertently prioritize accuracy in affirmative answers due to the distribution of the training data. This focus on affirmative accuracy can lead the models to perform better on *"Yes"* predictions, inflating the probability of such responses in practical applications.

These factors collectively contribute to the observed high probability of *"Yes"* responses in some models, highlighting the need for balanced training data, careful tuning, and improved contextual understanding in model development.

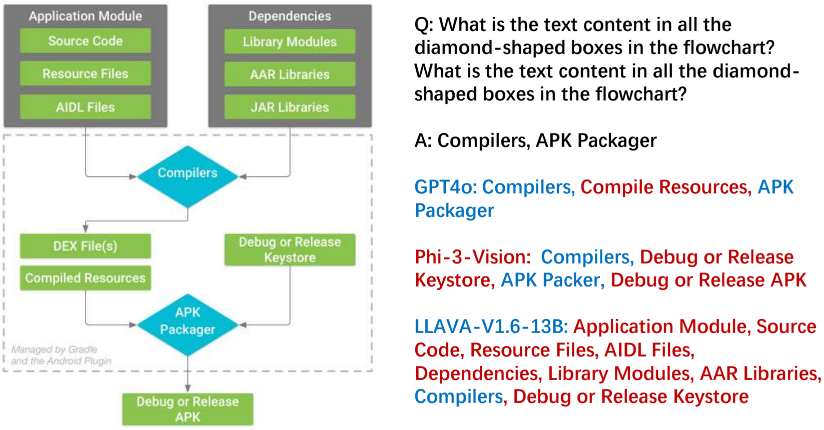


Figure 10: Model output and received scores.

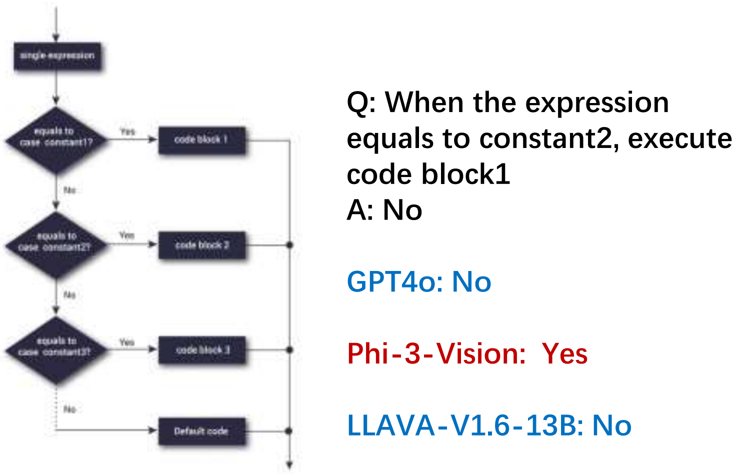


Figure 11: Model output and received scores.

F Detailed Comparison of Informaton Extraction Task

As shown in Figure 14, there are notable differences in the distribution of effective factor values among various models. Phi-3-Vision and GPT-4 have a broad distribution of effective factor values, with more data points in the higher effective factor region (above 0.6). Other models have a more scattered distribution of effective factor values, with most concentrated in the low effective factor region, particularly Qwen-Chat-VL, where the majority of data points are below 0.2. Despite the Cogvlm family having relatively high effective factors and fewer incorrect answers, the overall number of labels is also small. After subtracting the incorrect labels, the number of correct labels is minimal, resulting in a very low total score.

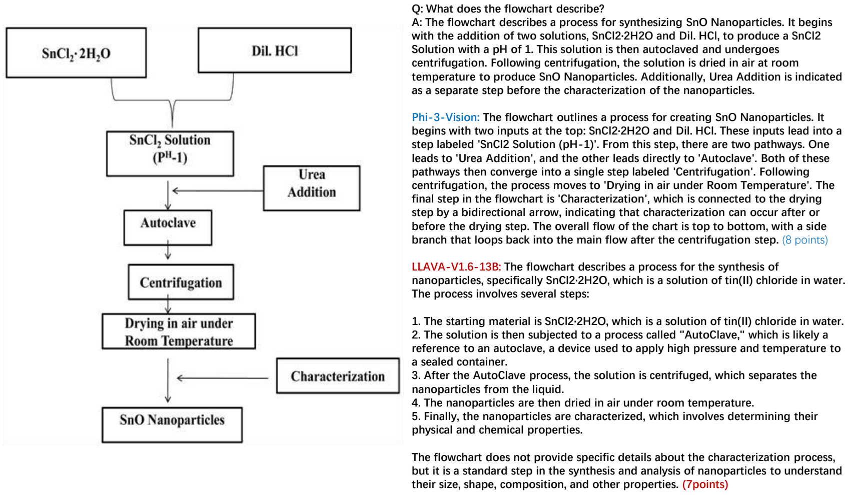


Figure 12: Model output and received scores.

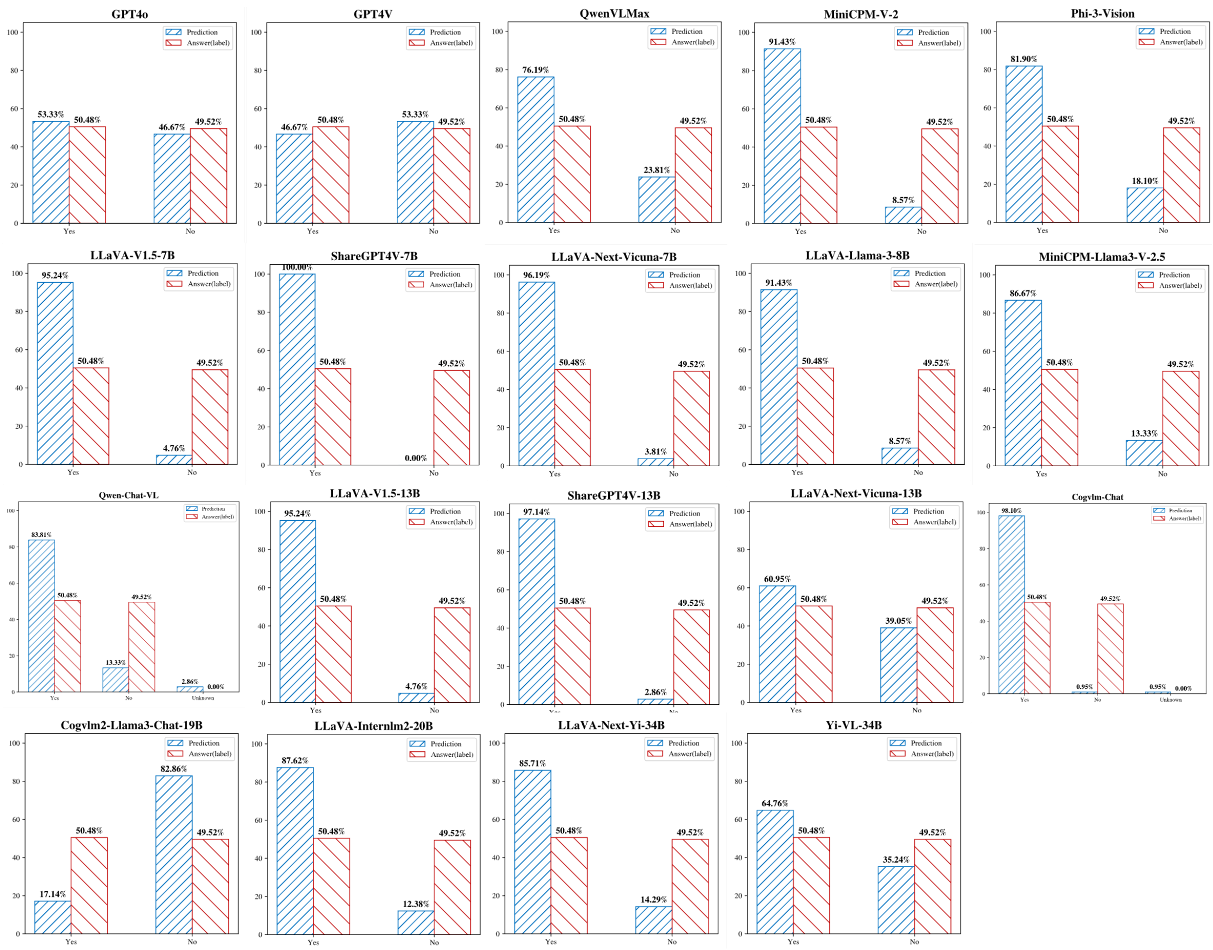


Figure 13: The prediction distributions of different models on the logical verification task.

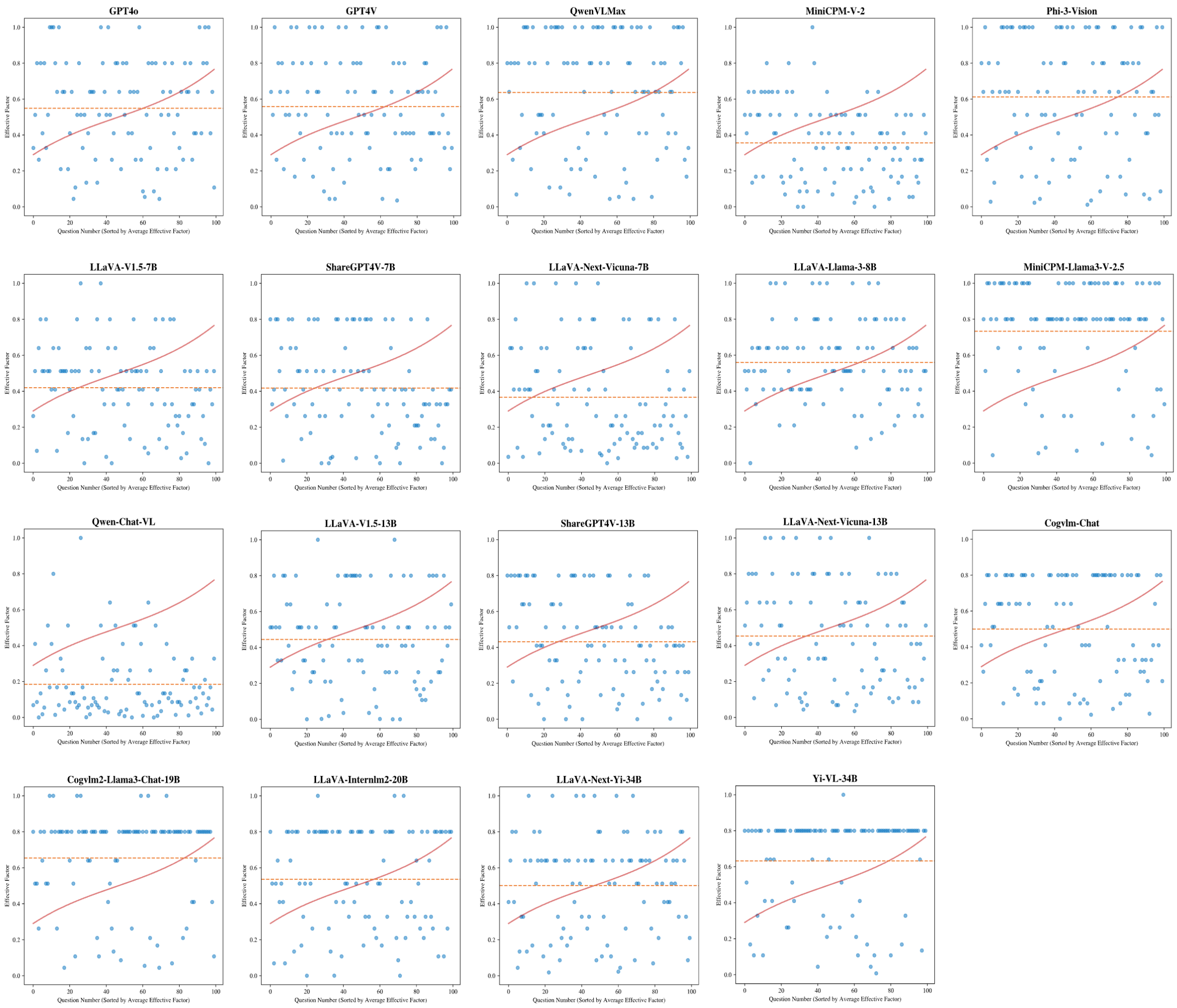


Figure 14: Comparison of MLLMs Performance on Information Extraction tasks based on effective factor distribution. The red line represents the smoothed ascending order of the average effective factor across all models for each specific question. The orange line indicates the average effective factor for each model across all question.