Ranked from Within: Ranking Large Multimodal Models Without Labels

Weijie Tu [1] **Weijian Deng** [1] **Dylan Campbell** [1] **Yu Yao** [2] **Jiyang Zheng** [2] **Tom Gedeon** [1 3 4] **Tongliang Liu** [2]

Abstract

Can the relative performance of a pre-trained large multimodal model (LMM) be predicted without access to labels? As LMMs proliferate, it becomes increasingly important to develop efficient ways to choose between them when faced with new data or tasks. The usual approach does the equivalent of giving the models an exam and marking them. We opt to avoid marking and the associated labor of determining the ground-truth answers. Instead, we explore other signals elicited and ascertain how well the models know their own limits, evaluating the effectiveness of these signals at unsupervised model ranking. We evaluate 47 state-of-the-art LMMs (*e.g.*, LLaVA) across 9 visual question answering benchmarks, analyzing how well uncertainty-based metrics can predict relative model performance. Our findings show that uncertainty scores derived from softmax distributions provide a robust and consistent basis for ranking models across various tasks. This facilitates the ranking of LMMs on unlabeled data, providing a practical approach for selecting models for diverse target domains without requiring manual annotation.

1. Introduction

Large multimodal models (LMMs), such as LLaVA (Liu et al., 2024c) and InstructBLIP (Dai et al., 2023), have demonstrated remarkable capabilities in handling a wide array of complex multimodal tasks, proving highly adaptable across diverse real-world applications (Liu et al., 2024a; Dai et al., 2023; Wang et al., 2024a; Huang et al., 2024; 2025; Zheng et al., 2025). From addressing scientific challenges (Lu et al., 2022; Hiippala et al., 2021) and performing optical character recognition (Mishra et al., 2019; Masry et al., 2022; Mathew et al., 2021) to identifying the spatial position of objects (Tong et al., 2024; x.ai, 2024), LMMs are increasingly widespread in practical settings. As LMMs proliferate, the need for rigorous evaluation metrics that accurately capture their capabilities and limitations has become more urgent. Thus, numerous benchmarks have been developed (Lu et al., 2022; Hiippala et al., 2021; Masry et al., 2022; Singh et al., 2019; Mathew et al., 2021; x.ai, 2024; Yue et al., 2024), aiming to provide reliable rankings and guide users in selecting models best suited for specific deployment scenarios.

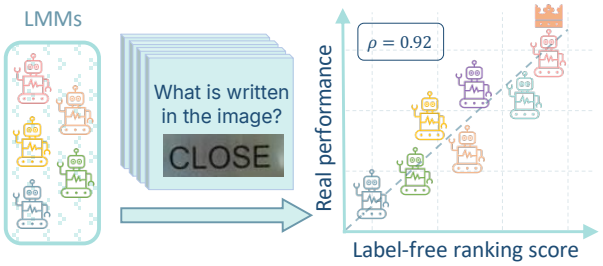


Figure 1. **Overview of unsupervised ranking for LMMs**. In scenarios where labeled data is scarce, selecting the best-performing model can be challenging. Our approach introduces a label-free proxy ranking score designed to reflect true performance, achieving a high correlation ($\rho = 0.92$) with actual metrics. This enables unsupervised comparison of LMMs, allowing users to identify the most suitable model without needing labeled data.

However, these benchmarks are based on carefully curated datasets that can require substantial resources to develop and label. For many users who may not have access to these resources, assessing model performance can be challenging. Additionally, standard evaluations are often dataset-centric, depending on fixed, human-labeled metrics that may not capture the full range of model capabilities necessary for diverse applications. As LMM tasks continue to diversify and expand, evaluating them effectively has become increasingly complex, with new applications frequently needing additional data curation and specialized capabilities.

Addressing the challenge of efficiently evaluating a set of LMMs is important for their usability and effectiveness in deployment. As illustrated in Figure 1, selecting the most suitable LMM from a range of available options is

[1]Australian National University [2]Sydney AI Centre, The University of Sydney [3]Curtin University [4]University of ÓBuda. Correspondence to: Tom Gedeon <tom.gedeon@anu.edu.au>, Tongliang Liu <tongliang.liu@sydney.edu.au>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

challenging in the absence of annotations. To address this, our work investigates label-free proxy ranking scores that closely align with the true performance ranking, enabling effective model comparison in label-scarce settings.

Our first finding is that the naïve approach of using the model's performance on an existing benchmark to rank its performance in a new target environment could be unstable. We further report that measures of prediction uncertainty are more effective ranking indicators. LMMs generate answers in an open-ended form by producing token sequences, with each token selected from the vocabulary based on the probability. This enables model uncertainty to be assessed using the logits at each token position.

To this end, we evaluate 45 different LMMs that have different training frameworks, *e.g.*, LLaVA-V1.5 (Liu et al., 2024a) and InstructBLIP (Dai et al., 2023), different visual encoders, *e.g.*, CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023), and language models, *e.g.*, Vicuna (Team, 2023) and LLaMA (Touvron et al., 2023). We investigate three categories of model uncertainty approaches: softmax probabilities, self-consistency, and labeled proxy sets. We evaluate the ranking performance on 9 widely-adopted multimodal benchmarks, which span diverse domains, including reasoning scientific questions, recognizing optical characters, and identifying objects' spatial positions. Our main findings are that

- the performance of models on one dataset may not accurately reflect the ranking of the same models on a different dataset (Section 4);
- the effectiveness of ranking methods is influenced by task characteristics (*e.g.*, closed-form or free-form generation), but probability-based variants are typically quite robust and predictive (Section 5); and
- when examining correlations in model performance across different dataset pairs, we observe that text prompt similarity better correlates with model performance across datasets than image feature similarity (Section 6).

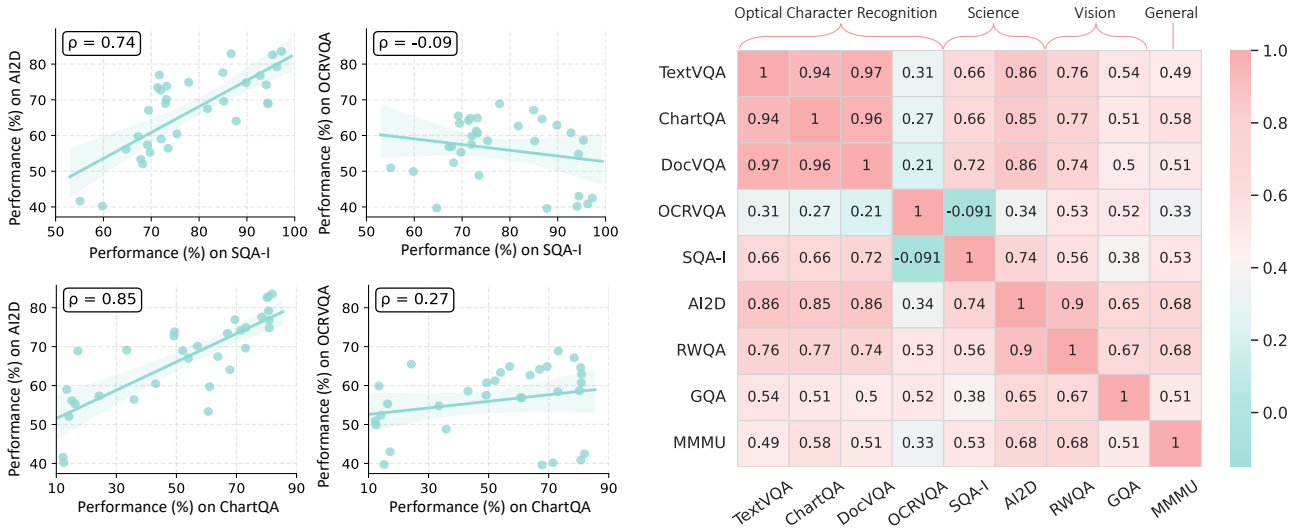
2. Related Work

Unsupervised Model Ranking. The goal of this task is to rank and select a best performant models without the access to the data annotations of target environments. The research on this task can date back to Forster *et al*. (Forster, 2000), and was further investigated in various tasks: (1) outlier detection (Zhao et al., 2022; 2021); (2) image classification (Kotary et al., 2022; Tu et al., 2024a; Zohar et al., 2024; Baek et al., 2022; Miller et al., 2021; Shi et al., 2024a); (3) time series anomaly detection (Ying et al., 2020); (4) multivariate anomaly detection in manufacturing systems (Engbers & Freitag, 2024), *etc*.

We discuss the most relevant two lines of research as follows. Miller *et al*. (2021) introduce an accuracy-on-the-line (AoL) phenomenon where strong linear correlation between probit-scaled in-distribution (ID) accuracy and out-of-distribution (OOD) accuracy across a variety of ML models. This means ID accuracy serves as a good indicator of model performance for a target domain. Shi *et al*. (2024a) revisit the established concept of lowest common ancestor (LCA) distance, which measures the hierarchical distance between labels and predictions within a predefined class hierarchy. By the observed linear correlation between ID LCA distance and OOD accuracy, it is viable to select a model based LCA distance of predictions. This work differs from prior study that we focus on ranking LMMs. This group of models makes predictions in a significantly different way from conventional classification models, which also results in different evaluation protocols.

Uncertainty Estimation for LLMs and LMMs. Uncertainty estimation seeks to quantify an ML model's confidence in its predictions (Guo et al., 2017). Recent studies have been dedicated to exploring uncertainty estimation specifically for LLMs (Kuhn et al., 2023; Malinin & Gales, 2021; Xiao & Wang, 2021; Huang et al., 2023; Lin et al., 2023; Kadavath et al., 2022; Azaria & Mitchell, 2023; Gottesman & Geva, 2024). For instance, Xiao *et al*. (2021) utilize ensemble methods to evaluate uncertainty in natural language generation models. Malinin *et al*. (2021) similarly introduce a unified approach to uncertainty estimation, leveraging ensemble methods, for autoregressive structured prediction tasks. To deal with the challenge of capturing "semantic equivalence" in natural language, Kuhn *et al*. (2023) propose semantic entropy, a method that utilizes linguistic invariances derived from shared meanings. Additionally, internal states of LMMs can be leveraged for uncertainty quantification or error detection by training a classifier (Azaria & Mitchell, 2023; Gottesman & Geva, 2024). This paper does not propose a new way to estimate uncertainty. Instead, it offers the novel insight that the existing uncertainty in LMM-generated outputs effectively reflects their relative performance across benchmarks without manual labels.

Evaluation & Benchmarking LMMs. The rapid development of LMMs has greatly propelled advancements in multimodal models, showcasing significant improvements in their perception and reasoning abilities. This shift has rendered traditional benchmarks, which focus solely on isolated task performance (Karpathy & Fei-Fei, 2015; Antol et al., 2015). Researchers have introduced new benchmarks to evaluate LMM in a broad spectrum of multimodal tasks (Goyal et al., 2017; Lin et al., 2014; Russakovsky et al., 2015). Recent studies (Yue et al., 2024; x.ai, 2024; Fu et al., 2023) highlight the need for more comprehensive benchmarks to effectively evaluate the reasoning and



(a) Correlation study on model performance between benchmarks

(b) Correlation matrix for 9 benchmarks

Figure 2. **Correlation analysis of model performance** across benchmarks. **(a)** Scatter plots illustrating the Spearman's rank correlation coefficients (ρ) between performance on selected benchmarks, indicating how well performance on one benchmark predicts performance on another. Each point represents a model. The straight lines are fit with robust linear regression (Huber, 2011). **(b)** Heatmap of the correlation matrix for performance across eight benchmarks, with color intensity representing the strength of correlation. Higher correlations (closer to 1) appear in red, while weaker correlations approach blue. The varying correlation strength indicates that using performance on one benchmark to rank LMMs in a target deployment environment may be inconsistent or unreliable.

understanding abilities of LMMs. For instance,

Several benchmarks (x.ai, 2024; Ainslie et al., 2023; Tong et al., 2024) have been developed to assess multimodal models' real-world spatial understanding. Lu *et al*. (2024) and Zhang *et al*. (2024) introduce benchmarks specifically designed to evaluate MLLMs' mathematical reasoning, focusing on their ability to comprehend and reason about visual mathematical figures. Yue *et al*. (2024) carefully curate a diverse set of multi-discipline tasks that require college-level subject knowledge and complex reasoning. Additionally, numerous benchmarks (Mishra et al., 2019; Masry et al., 2022; Mathew et al., 2021; Liu et al., 2023) assess the performance of LMMs in optical character recognition.

3. Task Formulation

Task Definition. Let a multimodal task be represented by a dataset $T = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the input prompt and the corresponding answer for the i-th sample, respectively. We have access to M large multimodal models (LMMs), denoted as $\{f_m\}_{m=1}^{M}$. Each LMM f_m, with pre-trained weights $\boldsymbol{\theta}_m$, generates a sequence of tokens $\{z_k\}_{k=1}^{K}$ from the input prompt $\mathbf{x}$ via a decoding process: $z_k = f_m([\mathbf{x}, z_1, z_2, \ldots, z_{k-1}] \mid \boldsymbol{\theta}_m)$, where z_k represents the k-th generated token. To assess the performance of each LMM, this task employs an evaluation metric (*e.g.*, accuracy) that determines the ground-truth performance $\{g_m\}_{m=1}^{M}$ by comparing the generated sequences with the ground-truth answers $\mathbf{y}_i$.

The objective of this paper is to develop methods for computing a score s_m for each LMM without requiring access to task-specific data annotations. Ideally, these computed scores should closely correlate with ground-truth performance, allowing us to rank and select LMMs based on their performance using only these scores.

Evaluation Metric. We use Spearman's rank correlation coefficient ρ (Kendall, 1948) to evaluate the monotonic relationship between scores and model performance. Additionally, we calculate Kendall's weighted rank correlation τ_w (Shieh, 1998), which effectively highlights top-ranked items (You et al., 2021). Both coefficients range from $[-1, 1]$, with values near -1 or 1 indicating strong negative or positive correlations, and 0 indicating no correlation.

4. Uncertainty for Ranking LMMs

This section discusses the unique characteristics of ranking various LMMs compared to conventional ML models. We then introduce three distinct approaches that leverage uncertainty in model predictions for ranking.

4.1. What Makes Ranking LMMs Interesting?

Unique Challenges for Ranking LMMs. While LMMs can be considered a subset of machine learning (ML) mod-

els, ranking them introduces unique challenges not present in traditional ML models. Below, we discuss the unique characteristics of LMMs and the challenges they bring to the model-ranking process. First, LMMs often have billions of pre-trained parameters, which presents significant challenges for traditional "white-box" analysis for ranking various LMMs. Second, these models are typically trained on large datasets of instruction fine-tuning samples, which may be proprietary data. As a result, risk assessment methods that require access to training data are either unsuitable or should be adjusted. Additionally, although some LMMs may disclose their final model weights, other information, such as training loss and intermediate checkpoints, often remains undisclosed. The lack of the access to training details limits the use of techniques in unsupervised accuracy estimation (Deng & Zheng, 2021; Tu et al., 2023).

Does Accuracy-on-the-Line (AoL) Suffice? Given these challenges, one may consider leveraging the AoL phenomenon (Miller et al., 2021) as a potential method for ranking LMMs. AoL refers to the strong linear correlation between probit-scaled in-distribution (ID) accuracy and out-of-distribution (OOD) accuracy across various ML models. This suggests that ID accuracy could serve as a reliable predictor of OOD performance, making AoL suitable for selecting LMMs in target testing environments. However, there are several reasons why this is not the case. First, obtaining ID performance data is often impractical, as LMMs are typically trained on large, sometimes proprietary datasets, limiting direct access to ID metrics. Second, while performance on existing benchmarks is more readily available, using this data to rank models for new deployment environments is problematic. Figure 2 illustrates the correlation between proxy benchmark performance and target testing environments, revealing extreme variability in correlation strength. This highlights the unreliability of using benchmark performance as the sole ranking criterion. Third, relying solely on proxy benchmark performance fails to capture the unique statistical characteristics of target testing datasets. AoL tends to rank models identically across different environments, regardless of the actual deployment context.

What can we use for ranking LMMs? Our approach leverages the readily available outputs of LMMs—specifically, token prediction logits and generated tokens—without requiring intricate architectural analysis or complex extraction techniques. Inspired by recent work on uncertainty scores for classifier ranking (Hu et al., 2024; Tu et al., 2024b), we introduce a novel adaptation of these techniques tailored to the specific characteristics of LMMs. By analyzing the distribution of prediction logits and the variability in generated outputs, we aim to assess each model's self-awareness of its limitations. This enables an unsupervised model ranking method, focusing primarily on techniques that utilize these readily accessible LMM outputs.

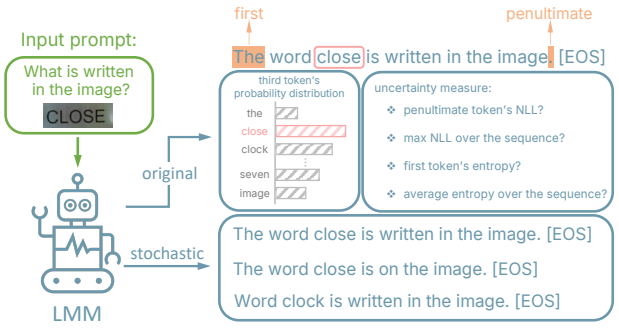


Figure 3. **An example of running one LMM for a VQA task.** We also present different token positions, methods to compute token-level uncertainty and the generation of stochastic predictions.

4.2. Assessing Uncertainty by Softmax Probabilities

LMMs generate answers in an open-form manner by producing sequences of tokens. The selection of each token can be viewed as a classification process, where the model selects the token with, *e.g.*, the highest probability over its entire vocabulary. This allows us to assess model uncertainty based on the logits at each token position.

For token-level uncertainty (Huang et al., 2023), we can focus on two specific positions: the first token and the penultimate token (*i.e.*, the token preceding the end-of-sequence token), as illustrated in Figure 3. The logit of the first token reflects the model's initial response to the input prompt, while the logit of the penultimate token captures the model's understanding of both the prompt and the generated response. The uncertainty associated with these two positions may provide insight into the model's overall confidence in answering the question. Beyond token-level uncertainty, sentence-level uncertainty can be calculated by aggregating uncertainty across all tokens in the sequence (Manakul et al., 2023; Huang et al., 2023). Specifically, sentence-level uncertainty can be quantified using the mean or the maximum negative log-likelihood (NLL) values across the entire generated sequence:

$$\mathbf{NLL}_{\text{max}} = \max_j \left(-\log p_{ij}\right) \tag{1}$$

$$\mathbf{NLL}_{\text{avg}} = -\frac{1}{J}\sum_{j=1}^{J} \log p_{ij} \tag{2}$$

where p_{ij} is the likelihood of the word generated by the LMM at the j-th token of the i-th sentence and J is the number of tokens generated in the sentence. Equation (1) quantifies a sentence's uncertainty via the least likely token, while Equation (2) uses the average per-token log-likelihood, allowing for length-independent comparisons

of uncertainty (Malinin & Gales, 2020; Murray & Chiang, 2018). Moreover, Equation (2) relates to perplexity, $\exp \mathrm{Avg}(-\log p)$ (Jelinek et al., 1977; Manning & Schütze, 1999), a standard measure of model quality.

Alternatively, entropy $\mathcal{H}$ can be used instead of the negative log-likelihood to assess uncertainty. In the context of LMMs ranking, we adopt normalized entropy to account for varying vocabulary sizes across models, thus scaling entropy to the interval $[0, 1]$ and making it comparable across different models. The normalized entropy is given by:

$$\mathcal{H}_{ij} = -\frac{1}{\log |\mathcal{W}|} \sum_{w \in \mathcal{W}} p_{ij}(w) \log p_{ij}(w) \tag{3}$$

where $p_{ij}(w)$ is the likelihood of the word w being generated at the j-th token of the i-th sentence, and W is the set of all possible words in the vocabulary.

There are eight variants of output probability-based methods, denoted as $\textbf{NLL}_{\text{F}}$, $\textbf{NLL}_{\text{P}}$, $\textbf{NLL}_{\text{max}}$, $\textbf{NLL}_{\text{avg}}$, $\textbf{Ent}_{\text{F}}$, $\textbf{Ent}_{\text{P}}$, $\textbf{Ent}_{\text{max}}$, and $\textbf{Ent}_{\text{avg}}$. "NLL" and "Ent" represent negative log-likelihood and entropy, respectively, while "F" and "P" refer to the first and penultimate tokens.

4.3. Assessing Uncertainty by Self-Consistency

Another approach involves examining the non-deterministic generations produced by models. The core intuition is that a more accurate model will produce predictions closely aligned with the original answer, while a less accurate model may yield more divergent responses with each inference. In the case of LMMs, the temperature parameter t controls the randomness of predictions: a temperature of zero forces deterministic predictions, where the model selects only the token with the highest probability. When t is larger than 0, the model generates tokens stochastically, selecting tokens based on probabilities above a threshold. As t increases, tokens are sampled from an increasingly uniform distribution (Chen et al., 2023b; Cobbe et al., 2021).

To analyze the consistency in these stochastic predictions, we explore two common methods: BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019). BLEU is a n-gram-based metric that evaluates the similarity of generated sequences to the reference answer, while BERTScore uses a pre-trained language model to embed answers and measures similarity in embedding space. Then, we use the mean value of similarities to represent the consistency for the input sample, which can be denoted as $\frac{1}{T}\sum_{i=1}^{T} \mathrm{sim}(P_i, P_{\text{ori}})$, where T is the number of stochastic inferences, $\mathrm{sim}(\cdot)$ is the similarity function (*e.g.*, BLEU), and P_i and P_{ori} are the i-th stochastic prediction and the original answer, respectively. Following the practice outlined in (Chen et al., 2023b; Cobbe et al., 2021; Huang et al., 2023), we collect five stochastic inferences per sample and set $t = 0.7$ to maintain a relatively high degree of stochasticity in LMM generation while keeping compute overhead manageable. We adopt a unigram BLEU and denote the methods using BLEU and BERTScore as $\textbf{Sample}_{\text{BLEU}}$ and $\textbf{Sample}_{\text{BERT}}$, respectively. Furthermore, BERTScore tends to assign high similarity scores between single letters, such as "A" and "B", which should be considered very different in the contexts of MCVQ. For example, BERTScore gives a high similarity score of 0.998 for "A" and "B", making it ineffective for distinguishing model performance. To address this, we modified the response to include the full answer text (*e.g.*, "A. North America"), denoted as $\textbf{Sample}^{*}_{\text{BERT}}$.

4.4. Assessing Uncertainty With Labeled Proxy Data

Beyond the uncertainty score calculated directly on the target dataset, we also explore scores obtained using a proxy labeled dataset. Garg et al. (2022) proposed average thresholded confidence (ATC), which calculates a threshold δ on a validation set (*e.g.*, CIFAR-10 (Krizhevsky & Hinton, 2009) validation set) and considers an image correctly classified on a new dataset with the same task if its confidence score exceeds the threshold. With the derived threshold, model performance on the target domain is estimated by the proportion of samples with confidence scores higher than the threshold. For LMMs, we use an existing benchmark as a proxy set to calculate δ. The calculation of ATC is

$$\text{ATC} = \mathbb{E}_{x \in T}[\mathbb{I}[u(f(x)) > \delta]], \tag{4}$$

where $\mathbb{I}$ is a binary indicator function, and $u(\cdot)$ represents an uncertainty estimation method, for which we use $\textbf{NLL}_{\text{max}}$. ATC provides both a ranking of model performance and an estimate of the expected performance.

5. Experiments

In this section, we first introduce the experimental setup including the evaluated datasets and models we considered. Then, we show the results of ranking different large multimodal models (LMMs).

5.1. Experiment Setup

Benchmarks. We choose the task of visual question answering, which is a common way to evaluate LMMs (Liu et al., 2024a; Dai et al., 2023; Beyer et al., 2024; Lu et al., 2022; Yue et al., 2024). We evaluate on multiple choice visual question (MCVQ) and visual question answering (VQA) benchmarks. MCQ and VQA are both types of question-answering formats for evaluating LMMs. For the former, the model is provided with several answer options, out of which the correct subset is to be selected. In contrast, the latter is usually open-ended and the models may generate answers freely. We consider 8 widely-adopted MCVQ and VQA benchmarks. They are (1) the

Method	MCVQ								VQA										Average	
	SQA-I		AI2D		RWQA		MMMU		GQA		ChartQA		OCRVQA		TextVQA		DocVQA			
	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w
AoL	0.52	0.45	0.73	0.61	0.70	0.57	0.54	0.45	0.53	0.32	0.69	0.59	0.30	0.20	0.69	0.57	0.68	0.60	0.60	0.48
NLL$_{F}$	0.83	0.78	0.84	0.76	0.56	0.58	0.60	0.59	0.71	0.57	0.63	0.41	0.78	0.64	0.74	0.63	0.84	0.74	0.73	0.63
NLL$_{P}$	0.64	0.60	0.61	0.58	0.25	0.37	0.16	0.18	0.58	0.52	0.79	0.67	0.81	0.68	0.71	0.67	0.88	0.78	0.60	0.56
NLL$_{max}$	0.80	0.76	0.91	0.81	0.63	0.63	0.49	0.44	0.72	0.59	0.94	0.80	0.64	0.63	0.83	0.69	0.92	0.82	0.76	0.69
NLL$_{avg}$	0.72	0.64	0.88	0.73	0.64	0.56	0.50	0.40	0.67	0.55	0.92	0.75	0.81	0.65	0.81	0.72	0.93	0.82	0.76	0.65
Ent$_{F}$	0.59	0.51	0.82	0.59	0.65	0.56	0.64	0.52	0.54	0.20	0.64	0.45	0.69	0.57	0.71	0.56	0.80	0.70	0.68	0.52
Ent$_{P}$	0.49	0.39	0.69	0.48	0.43	0.43	0.34	0.24	0.46	0.21	0.82	0.64	0.80	0.64	0.70	0.54	0.86	0.68	0.62	0.47
Ent$_{max}$	0.58	0.39	0.82	0.59	0.67	0.60	0.66	0.54	0.54	0.21	0.88	0.66	0.53	0.52	0.76	0.60	0.87	0.73	0.70	0.54
Ent$_{avg}$	0.58	0.33	0.80	0.56	0.62	0.50	0.59	0.41	0.57	0.26	0.91	0.68	0.77	0.62	0.74	0.58	0.88	0.74	0.72	0.52
Sample$_{BLEU}$	0.65	0.68	0.76	0.59	0.48	0.36	0.37	0.12	0.44	0.41	0.89	0.61	0.47	0.58	0.81	0.60	0.90	0.63	0.64	0.51
Sample$_{BERT}$	0.46	0.52	0.70	0.55	0.52	0.28	0.45	0.44	0.51	0.47	0.90	0.61	0.75	0.72	0.77	0.59	0.89	0.64	0.66	0.54
Sample$^{*}_{BERT}$	0.67	0.56	0.78	0.62	0.67	0.48	0.59	0.39	0.51	0.47	0.90	0.61	0.75	0.72	0.77	0.59	0.89	0.64	0.73	0.56
ATC	0.73	0.74	0.80	0.72	0.51	0.40	0.41	0.35	0.39	0.20	0.95	0.85	0.28	0.21	0.69	0.64	0.89	0.77	0.63	0.54

Table 1. **Method comparison across eight multimodal tasks**. The table presents a comparison of four groups of methods: accuracy-based, output-probability-based, sample-based, and unsupervised model evaluation methods. We evaluate these methods using Spearman's rank correlation (ρ) and weighted Kendall's correlation (τ_w). Both coefficients range from -1 to 1, where values close to -1 or 1 indicate strong negative or positive correlations, respectively, and 0 indicates no correlation. The **AoL** and **ATC** performance is calculated as the average correlation when using the scores computed on the other seven domains and the model performance on the target domain. The highest correlation values for each task are highlighted in green, while the second highest values are marked in blue. All methods are ranked to **three** decimal places. Note that, we use the absolute value of correlation strength in the table for **NLL** and **Ent**. The results indicate that NLL$_{max}$ and NLL$_{avg}$ are often preferable, as they show greater stability and stronger correlation with model performance.

subset of ScienceQA (Lu et al., 2022) with images (SQA-I) and AI2D (Hiippala et al., 2021) which assess LMMs' scientific knowledge; (2) ChartQA (Masry et al., 2022), OCRVQA (Mishra et al., 2019), TextVQA (Singh et al., 2019) and DocVQA (Mathew et al., 2021) to examine their ability to recognize optical character; (3) RealWorldQA (RWQA) (x.ai, 2024) and GQA (Ainslie et al., 2023) which evaluate LMMs' vision-centric capability; (4) MMMU (Yue et al., 2024) which assays LMMs on multi-disciplinary tasks that demand college-level subject knowledge. Note that SQA-I, AI2D, RWQA and MMMU are MCVQ datasets, while the others are VQA.

Models. The goal is to choose the best LMM over all different series of LMMs. We include LLaVA-V1.5 (Liu et al., 2024a), ShareGPT4V (Chen et al., 2023a), LLaVA-NeXT (Liu et al., 2024b), InstructBLIP (Dai et al., 2023), LLaVA-NeXT-Interleave (Li et al., 2024b), LLaVA-OneVision (Li et al., 2024a), Eagle (Shi et al., 2024b), mPLUG-Owl (Li et al., 2022), InternVL (Chen et al., 2024), PaliGemma (Beyer et al., 2024), Mantis (Jiang et al., 2024), DeepSeek-VL2 (Wu et al., 2024) and Qwen2-VL (Wang et al., 2024b). In total, we collect 32 different models, which all can be accessed on Hugging Face (Wolf et al., 2020).

5.2. Key Findings

Accuracy-on-the-Line is unreliable for ranking LMMs in new domains. Table 1 summarizes the ranking capability of all methods across eight benchmarks. The AoL performance is calculated as the average correlation strength when using the other seven domains to predict model rankings in the target domain. We observe that AoL does not achieve consistently high correlation with model performance in seven out of the eight benchmarks. Although AoL shows strong results on RealWorldQA, where it performs best, uncertainty-based methods also demonstrate high correlation strength. For instance, **NLL$_{max}$** exhibits a Spearman's ρ 0.07 lower but a weighted Kendall's τ_w 0.06 higher than AoL. These findings suggest that relying on existing benchmarks alone to select models for deployment could be risky and unstable, as they may not well capture the statistical characteristics of the new target domain.

The choice of token matters for output probability-based ranking. We studied four ways of using tokens for estimating model uncertainty. For the two on specific positions, the first and penultimate tokens, their predictive performance depends on the task type (*i.e.*, MCVQ *vs.* VQA). The uncertainty associated with generating the first token is more indicative for MCVQ tasks, while the penultimate token generally proves more predictive for VQA tasks. This difference is due to the nature of MCVQ tasks, which often prompt models to generate only a single option letter (*e.g.*, "A"), making the first token crucial for evaluating response accuracy. In contrast, VQA tasks require open-form responses consisting of multiple tokens, making the penultimate token—reflecting the model's understanding of both the question and the answer—more informative.

For the two that consider every token in the generated output, *i.e.*, **NLL$_{max}$** and **NLL$_{avg}$**, we find that they are more stable and less task-specific compared to variants that consider individual tokens. While **NLL$_{max}$** and **Ent$_{max}$** (reflecting the least confident token) is more effective for MCVQ tasks, both methods perform similarly on VQA tasks. The higher

ranking correlation of using the first token and the least confident token as ranking indicators for MCVQ suggests that a single token (typically the option letter) holds significant meaning. Considering all potential tokens can sometimes result in an unexpectedly higher confidence level, as LMMs may generate the complete answer alongside the option letter (*e.g.*, "B. Columbia"). The tokens following the option letter often have high confidence levels, leading to an overall increase in estimated uncertainty. This tendency results in models generating entire answers with lower uncertainty scores, yielding higher ranks. These findings underscore the importance of further exploration into the optimal token positions for uncertainty estimation in LMMs. One potential approach is leveraging a language model to identify the position of the option letter and use its softmax probability as the uncertainty measure for the complete response.

The negative log-likelihood (NLL) is more stable and predictive for LMMs ranking than normalized entropy. For seven out of eight benchmarks, NLL consistently shows stronger correlation with model performance than entropy. Both NLL and entropy make use of the softmax probability distribution predicted for each token during generation. However, the NLL-based approaches only consider the maximum probability in this distribution, while the entropy considers all entries. The lower correlation strength of normalized entropy may therefore stem from the significantly different vocabulary sizes of various LMMs. For instance, LLaVA-V1.5-7B has a vocabulary of 32k tokens, while PaliGemma-3B-mix has a much larger vocabulary of 257k tokens. Despite being normalized by vocabulary size, entropy may still be more susceptible to noise than NLL. Future work could include applying dimension reduction techniques or limiting consideration to the top-k most probable tokens or a pre-defined set of tokens.

Sample-based methods are strong candidates for model ranking without intrinsic access to LMMs. The correlation strength of sample-based methods is influenced by the nature of the task. They yield higher correlation scores on VQA tasks compared to MCVQ tasks. In VQA, models typically produce more varied responses, making BLEU and BERTScore effective for capturing uncertainty. However, MCVQ tasks constrain models to select from a pre-defined set of answers. While non-zero temperature introduces some randomness, the variation between stochastic inferences is limited. Additionally, we also find that $\textbf{Sample}^{*}_{\text{BERT}}$ has higher correlation scores than $\textbf{Sample}_{\text{BERT}}$ on four MCVQ benchmarks and suggests that developing more advanced algorithms to capture uncertainty from multiple stochastic inferences could be beneficial.

Although sample-based methods show weaker overall correlations, they remain competitive without requiring access to model architecture or internal states. This highlights their potential to rank API-based or closed-source LMMs (*e.g.*, GPT-4V (Achiam et al., 2023)).

Method	MCVQ		VQA		Average
	AI2D	MMMU	TextVQA	ChartQA	
AoL	0.42	0.43	0.57	0.64	0.52
$\textbf{NLL}_{\text{max}}$	0.29	0.59	0.97	0.99	0.71
$\textbf{NLL}_{\text{avg}}$	0.26	0.28	0.98	0.98	0.63
$\textbf{Ent}_{\text{max}}$	0.37	0.62	0.97	0.96	0.73
$\textbf{Ent}_{\text{avg}}$	0.11	0.38	0.98	0.98	0.63
$\textbf{Sample}_{\text{BLEU}}$	0.63	0.73	0.76	0.64	0.69
ATC	0.12	0.59	0.85	0.84	0.60

Table 2. **Method comparison for ranking different LLaVA-prismatic** models on AI2D and TextVQA. We use Spearman's rank correlation (ρ) as the metric. We observe that uncertainty-based method still exhibit moderately high correlation strength, which indicate their effectiveness in ranking LLaVA models.

ATC can be used for LMMs ranking, but the correlation strength is influenced by the choice of proxy dataset. We compute ATC performance as the average correlation when using the other seven datasets as proxy datasets. Our analysis reveals the choice of proxy dataset is critical, as the scale of uncertainty calculated on different datasets can vary. This variation can lead to an inaccurately estimated threshold for determining instance correctness. A potential improvement involves adopting uncertainty calibration methods, such as temperature scaling (Guo et al., 2017), to calibrate model uncertainty onto a consistent scale.

6. Analysis

This section includes three distinct analyses. The first specifically investigates whether the considered methods are effective for ranking models within the same series. We use LLaVA models as a case study because they are widely adopted and representative of LMM architectures. The other two analyses consider models from different series, consistent with the broader evaluation in Section 5.

Ranking LMMs from the same series. So far, all experiments have focused on selecting LMMs from different model series. However, in some scenarios, the objective is to choose a training recipe that yields better-performing LMMs within the same model series. These models may be trained with additional fine-tuning steps, varied data augmentations, different training sources, or different language models (*e.g.*, Vicuna and Mistral (Jiang et al., 2023)) and visual encoders, such as CLIP and DINOv2 (Oquab et al., 2023). For our analysis, we use the LLaVA series due to its widespread adoption. Specifically, we employ 15 different LLaVA prismatic (LLaVA-pri) models (Karamcheti

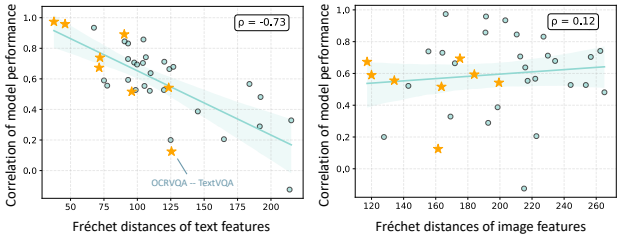


Figure 4. **Correlation Analysis of Fréchet Distances and Model Performance Correlation Across Datasets.** Orange stars indicate the dataset pairs with the highest similarity for each dataset. Observations reveal that variations in text prompt similarity are more closely aligned with changes in performance correlation than variations in image feature similarity.

et al., 2024). We report results on four datasets in Table 2: AI2D, MMMU, TextVQA, and ChartQA, with AI2D and TextVQA being used to evaluate the performance of different LLaVA variants in the original paper. The **AoL** and **ATC** performance metrics are computed in the same manner as described in Section 5, using the average correlation across other three datasets.

Our findings align with those from ranking different series of LMMs. First, using model performance on a single existing dataset may not accurately reflect LMMs rankings on a different domain, even when the models are trained with a similar pipeline and minor variations. Second, we observe that ranking different LLaVA-pri models in MCVQ presents a significant challenge, as the variance in model performance on MCVQ tasks is lower compared to VQA tasks. For instance, the performance gap on AI2D is only 4%, whereas the gap on TextVQA exceeds 20%. This indicates that ranking methods must capture subtle differences between models. We find that $\textbf{Sample}_{\text{BLEU}}$ remains effective, while NLL and Entropy may not capture these nuances accurately. Additionally, the decrease in correlation between uncertainty estimated by softmax probability suggests that although modifications to the training pipeline may have a slight effect on performance, they can lead to significant variations in the confidence levels for generation. This finding underscores the importance of assessing LMMs more comprehensively, beyond accuracy alone.

Analysis of the weak correlation of AoL. Figure 2(b) illustrates the correlation of model performance across different benchmarks. We observe that while TextVQA, ChartQA, DocVQA, and OCRVQA all aim to assess the capability of LMMs to recognize optical characters, the correlation between them varies significantly. Specifically, model performance on TextVQA, ChartQA, and DocVQA shows strong correlations, whereas performance on OCRVQA consistently exhibits low correlation with the other three datasets. To explore whether the images or texts within these datasets influence the correlation strength of model performance, we utilize CLIP-L-14-336 (Radford et al., 2021) to extract image and text embeddings. We then use the Fréchet distance (FD) (Fréchet, 1957) to measure the similarity between datasets based on these features.

Method	MCVQ		VQA	
	AI2D	MMMU	TextVQA	ChartQA
#Samples	3088	1050	5000	2500
50 samples	0.89	0.35	0.92	0.98
$\textbf{NLL}_{\text{max}}$	0.92	0.56	0.82	0.93

Table 3. **Labelling a subset of target domain** to rank models on AI2D, MMMU, TextVQA and ChartQA. We report the Spearman's rank correlation (ρ). While a small labeled set provides a reasonable ranking, it may not fully capture the overall order. In contrast, NLL_{max} offers more stable correlations, highlighting its potential for label-free model ranking.

Figure 4 presents a correlation study between the FD of dataset pairs and the correlation strength of model performance on those datasets. We observe a strong correlation for FDs computed using text input features, while FDs measured by image features show a weaker correlation. This suggests that text input similarity is likely a more influential factor for model performance correlation than image similarity. Additionally, orange stars are used to label the points representing the lowest FD for each dataset. Moreover, OCRVQA shows a low correlation strength with other datasets (Figure 2). The closest dataset in prompt feature space is TextVQA, with a FD of 124, which is considerably higher than the lowest FDs observed between other dataset pairs, typically around 70 or lower. This difference sheds light on the weak model performance correlation between OCRVQA and other OCR-focused datasets.

Ranking LMMs by a labeled subset of the target domain. Table 3 presents the correlation strength between model performance on 50 labeled instances in the target domain and the overall performance across the entire dataset. We observe that a small labeled set can provide a good ranking. However, such an approach may not fully capture the model ranking across the entire dataset, since the sampled data may not be representative (Polo et al., 2024). In contrast, $\textbf{NLL}_{\text{max}}$ gives a more stable correlation, highlighting the potential of uncertainty-based methods for effective model ranking without data annotation.

7. Conclusion and Discussion

This work studied whether the performance of large multimodal models in a new target domain can be ranked without the access to target domain labels. Our analysis identified only a weak correlation in model performance across

different domains. This motivated the investigation and evaluation of alternative approaches based on uncertainty estimates obtained from model predictions. We evaluate 45 LMMs on closed and open-ended visual question answering tasks, testing various training frameworks, visual encoders, and language models. Our experiments reveal that scores based on the negative log-likelihood of generated tokens serve as highly effective performance indicators for target domains. We also find that while stochastic sampling can be helpful, it is less effective for multiple-choice tasks, where it requires many repetitions of inference and careful temperature tuning. By establishing a baseline for uncertainty-based LMMs ranking, this study aims to motivate and inspire further research into this important area.

Potential future directions. Beyond uncertainty scores, several promising directions remain for future research. Test-time augmentation and semantic entropy present natural next steps. Additionally, the internal states of LMMs (Azaria & Mitchell, 2023; Gottesman & Geva, 2024) offer opportunities for uncertainty estimation or error detection, for example, by training a classifier on these representations. However, in unsupervised LMM ranking, this approach requires training a separate classifier per model, which can be computationally expensive. An alternative is to measure the statistical distance between a model's internal state for a given response and the distribution of internal states from multiple inference passes. A larger average distance may signal greater uncertainty and potentially indicate a lower model rank. Exploring how internal states can inform LMM ranking is an open direction. Finally, the observed asymmetry in the impact of text and image feature dissimilarity on cross-dataset correlations highlights a valuable area for improving multimodal benchmark design.

Acknowledgments

We thank all anonymous reviewers for their constructive feedback, which helped improve the quality of this paper. Tongliang Liu is partially supported by the following Australian Research Council projects: FT220100318, DP220102121, LP220100527, LP220200949, and IC190100031.

Impact Statement

This paper aims to advance machine learning while acknowledging potential societal impacts. Our findings could be misused: adversarial researchers may train models that consistently get ranked higher than other models, despite performing poorly on data. To mitigate this, incorporating robust calibration methods would be helpful, making model confidence accurately reflects uncertainty. This would help promote safe and responsible deployment of our findings.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ainslie, J., Lee-Thorp, J., De Jong, M., Zemlyanskiy, Y., Lebrón, F., and Sanghai, S. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *EMNLP*, 2023.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *ICCV*, pp. 2425–2433, 2015.

Azaria, A. and Mitchell, T. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=y2V6YgLaW7.

Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *NeurIPS*, pp. 19274–19289, 2022.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., and Lin, D. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*, 2023a.

Chen, X., Lin, M., Schärli, N., and Zhou, D. Teaching large language models to self-debug. In *ICLR*, 2023b.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pp. 24185–24198, 2024.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

Deng, W. and Zheng, L. Are labels always necessary for classifier accuracy evaluation? In *CVPR*, pp. 15069–15078, June 2021.

Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., Lin, D., and Chen, K. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024. URL https://arxiv.org/abs/2407.11691.

Engbers, H. and Freitag, M. Automated model selection for multivariate anomaly detection in manufacturing systems. *Journal of Intelligent Manufacturing*, 2024. doi: 10.1007/s10845-024-02479-z. URL https://doi.org/10.1007/s10845-024-02479-z.

Forster, M. R. Key concepts in model selection: Performance and generalizability. *Journal of mathematical psychology*, 44(1):205–231, 2000.

Fréchet, M. Sur la distance de deux lois de probabilité. *Comptes Rendus Hebdomadaires des Seances de L Academie des Sciences*, 244(6):689–692, 1957.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. Leveraging unlabeled data to predict out-of-distribution performance. In *ICLR*, 2022.

Gottesman, D. and Geva, M. Estimating knowledge in large language models without generating a single token. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3994–4019, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.232. URL https://aclanthology.org/2024.emnlp-main.232/.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017.

Hiippala, T., Alikhani, M., Haverinen, J., Kalliokoski, T., Logacheva, E., Orekhova, S., Tuomainen, A., Stone, M., and Bateman, J. A. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021.

Hinton, G. E. and Roweis, S. Stochastic neighbor embedding. In *NeurIPS*, 2002.

Hu, D., Luo, M., Liang, J., and Foo, C.-S. Towards reliable model selection for unsupervised domain adaptation: An empirical study and a certified baseline. In *Adv. Neural Inform. Process. Syst. Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=rI7kbFTSpr.

Huang, Y., Song, J., Wang, Z., Zhao, S., Chen, H., Juefei-Xu, F., and Ma, L. Look before you leap: An exploratory study of uncertainty measurement for large language models. In *ICSE*, 2023.

Huang, Z., Liu, C., Dong, Y., Su, H., Zheng, S., and Liu, T. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In *ICML*, 2024.

Huang, Z., Niu, G., Han, B., Sugiyama, M., and Liu, T. Towards out-of-modal generalization without instance-level modal correspondence. In *ICLR*, 2025. URL https://openreview.net/forum?id=LuVulfPgZN.

Huber, P. J. Robust statistics. In *International Encyclopedia of Statistical Science*, pp. 1248–1251. Springer, 2011.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. M. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62, 1977. URL https://api.semanticscholar.org/CorpusID:121680873.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jiang, D., He, X., Zeng, H., Wei, C., Ku, M., Liu, Q., and Chen, W. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, 2024.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Karamcheti, S., Nair, S., Balakrishna, A., Liang, P., Kollar, T., and Sadigh, D. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *ICML*, 2024.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137, 2015.

Kendall, M. G. *Rank correlation methods.* Griffin, 1948.

Kotary, J., Di Vito, V., and Fioretto, F. Differentiable model selection for ensemble learning. In *IJCAI*, 2022.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. 2009.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022.

Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., and Li, C. Llava-next: Tackling multi-image, video, and 3d in large multimodal models, June 2024b. URL https://llava-vl.github.io/blog/2024-06-16-llava-next-interleave/.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Lin, Z., Trivedi, S., and Sun, J. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024a.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024b. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NeurIPS*, 2024c.

Liu, Y., Li, Z., Yang, B., Li, C., Yin, X., Liu, C.-l., Jin, L., and Bai, X. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, pp. 2507–2521, 2022.

Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. In *ICLR*, 2020.

Malinin, A. and Gales, M. Uncertainty estimation in autoregressive structured prediction. In *ICLR*, 2021.

Manakul, P., Liusie, A., and Gales, M. J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *ACL*, 2023.

Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL*, 2022.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *WACV*, pp. 2200–2209, 2021.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, pp. 7721–7735, 2021.

Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pp. 947–952, 2019. doi: 10.1109/ICDAR.2019.00156.

Murray, K. and Chiang, D. Correcting length bias in neural machine translation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://aclanthology.org/W18-6322.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pp. 311–318. Association for Computational Linguistics, 2002.

Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015.

Shi, J., Gare, G., Tian, J., Chai, S., Lin, Z., Vasudevan, A., Feng, D., Ferroni, F., and Kong, S. Lca-on-the-line: Benchmarking out-of-distribution generalization with class taxonomies. In *ICML*, 2024a.

Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Huang, D.-A., Yin, H., Sapra, K., Yacoob, Y., Shi, H., Catanzaro, B., Tao, A., Kautz, J., Yu, Z., and Liu, G. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv:2408.15998*, 2024b.

Shieh, G. S. A weighted kendall's tau statistic. *Statistics & Probability Letters*, 39(1):17–24, 1998.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *CVPR*, pp. 8317–8326, 2019.

Team, V. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/. *Based on evaluations comparing Vicuna's responses to those of ChatGPT.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pp. 9568–9578, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tu, W., Deng, W., Gedeon, T., and Zheng, L. A bag-of-prototypes representation for dataset-level applications. In *CVPR*, pp. 2881–2892, June 2023.

Tu, W., Deng, W., Zheng, L., and Gedeon, T. What does softmax probability tell us about classifiers ranking across diverse test conditions? *TMLR*, 2024a.

Tu, W., Deng, W., Zheng, L., and Gedeon, T. What does softmax probability tell us about classifiers ranking across diverse test conditions? *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=vtiDUgGjyx.

Wang, H., Huang, Z., Lin, Z., and Liu, T. NoiseGPT: Label noise detection and rectification through probability curvature. In *NeurIPS*, 2024a. URL https://openreview.net/forum?id=VRRvJnxgQe.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Transformers: State-of-the-art natural language processing. In *EMNLP*, pp. 38–45, 2020.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., Xie, Z., Wu, Y., Hu, K., Wang, J., Sun, Y., Li, Y., Piao, Y., Guan, K., Liu, A., Xie, X., You, Y., Dong, K., Yu, X., Zhang, H., Zhao, L., Wang, Y., and Ruan, C. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

x.ai. Grok 1.5v: The future of ai models, 2024. URL https://x.ai/blog/grok-1.5v.

Xiao, Y. and Wang, W. Y. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2734–2744, 2021.

Ying, Y., Duan, J., Wang, C., Wang, Y., Huang, C., and Xu, B. Automated model selection for time-series anomaly detection. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3428–3434. ACM, 2020.

You, K., Liu, Y., Wang, J., and Long, M. Logme: Practical assessment of pre-trained models for transfer learning. In *ICML*, 2021.

Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pp. 9556–9567, 2024.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *CVPR*, 2023.

Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.-W., Qiao, Y., et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *ECCV*, pp. 169–186, 2024.

Zhang, T., Das, R., Sima'an, K., and Sutskever, I. Bertscore: Evaluating text generation with bert. In *ICLR*, 2019.

Zhao, Y., Rossi, R., and Akoglu, L. Automatic unsupervised outlier model selection. In *NeurIPS*, pp. 4489–4502, 2021.

Zhao, Y., Zhang, S., and Akoglu, L. Toward unsupervised outlier model selection. In *ICDM*, pp. 773–782. IEEE, 2022.

Zheng, J., Shen, J., Yao, Y., Wang, M., Yang, Y., Wang, D., and Liu, T. Chain-of-focus prompting: Leveraging sequential visual cues to prompt large autoregressive vision models. In *ICLR*, 2025. URL https://openreview.net/forum?id=noidywkBba.

Zohar, O., Huang, S.-C., Wang, K.-C., and Yeung, S. Lovm: Language-only vision model selection. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 36, 2024.

A. Appendix

In this supplementary material, we first introduce the experimental details including the models, datasets and compute in Appendix B. Next, we visualize the image and text features using t-SNE plots. Last, we show the full results of the correlation study on all datasets.

B. Experiment Details

B.1. Datasets

All experiments are conducted on VLMEvalKit (Duan et al., 2024). We consider 8 datasets and their corresponding links to download TSV files via the toolkit:

ScienceQA (Lu et al., 2022) (https://opencompass.openxlab.space/utils/VLMEval/ScienceQA_TEST.tsv);

AI2D (Hiippala et al., 2021) (https://opencompass.openxlab.space/utils/VLMEval/AI2D_TEST.tsv);

ChartQA (Masry et al., 2022) (https://opencompass.openxlab.space/utils/VLMEval/ChartQA_TEST.tsv);

OCRVQA (Mishra et al., 2019) (https://opencompass.openxlab.space/utils/VLMEval/OCRVQA_TESTCORE.tsv);

TextVQA (Singh et al., 2019) (https://opencompass.openxlab.space/utils/VLMEval/TextVQA_VAL.tsv);

DocVQA (Mathew et al., 2021) (https://opencompass.openxlab.space/utils/VLMEval/DocVQA_VAL.tsv);

RealWorldQA (x.ai, 2024) (https://opencompass.openxlab.space/utils/VLMEval/RealWorld.tsv);

MMMU (Yue et al., 2024) (https://opencompass.openxlab.space/utils/VLMEval/MMMU_DEV_VAL.tsv)

GQA (Ainslie et al., 2023) (https://opencompass.openxlab.space/utils/VLMEval/GQA_TestDev_Balanced.tsv)

B.2. Models

We include a diverse array of large multimodal models from 12 series:

LLaVA-V1.5

```
llava_v1.5_7b
llava_v1.5_13b
```

LLaVA-NeXT

```
llava_next_mistral_7b
llava_next_vicuna_7b
llava_next_vicuna_13b
```

LLaVA-NeXT-Interleave

```
llava_next_interleave_7b
llava_next_interleave_7b_dpo
```

LLaVA-OneVision

```
llava_onevision_qwen2_0.5b_ov
llava_onevision_qwen2_7b_ov
llava_onevision_qwen2_0.5b_si
llava_onevision_qwen2_7b_si
```

ShareGPT4V

```
sharegpt4v_7b
sharegpt4v_13b
```

InstructBLIP

```
InstructBLIP_7b
InstructBLIP_13b
```

Eagle

```
Eagle-X5-7B
Eagle-X5-13B
Eagle-X5-13B-Chat
```

InternVL

```
Mini-InternVL-Chat-2B-V1-5
Mini-InternVL-Chat-4B-V1-5
InternVL2-1B
InternVL2-2B
InternVL2-4B
InternVL2-8B
```

PaliGemma

```
paligemma-3b-mix-224
paligemma-3b-mix-448
```

Mantis

```
Mantis-8B-Idefics2
Mantis-8B-clip-llama3
Mantis-8B-siglip-llama3
```

mPLUG-Owl2

```
mPLUG-Owl2
```

Qwen2-VL

```
Qwen2-VL-2B-Instruct
```

Qwen-VL

```
deepseek_vl2_tiny
```

LLaVA Prismatic

```
reproduction-llava-v15+7b
one-stage+7b
full-ft-multi-stage+7b
full-ft-one-stage+7b
in1k-224px+7b
dinov2-224px+7b
clip-224px+7b
siglip-224px+7b
clip-336px-resize-crop+7b
clip-336px-resize-naive+7b
siglip-384px-letterbox+7b
llama2-no-cotraining+7b
llava-lvis4v+7b
llava-lrv+7b
llava-lvis4v-lrv+7b
```

B.3. Compute and Library

PyTorch version is 2.01.0+cu117. All experiment is run on four A6000 GPUs. All 45 models can be downloaded via Huggingface with different versions of transformer library:

`transformers==4.33.0` for mPLUG-Owl2 (Li et al., 2022) and InstructBLIP (Dai et al., 2023);

`transformers==4.37.0` for LLaVA-V1.5 (Liu et al., 2024a), ShareGPT4V (Chen et al., 2023a), InternVL (Chen et al., 2024) series;

`transformers==latest` for LLaVA-NeXT (Liu et al., 2024b), LLaVA-OneVision (Li et al., 2024a), LLaVA-NeXT-Interleave (Li et al., 2024b) PaliGemma-3B (Beyer et al., 2024), Mantis (Jiang et al., 2024), Eagle (Shi et al., 2024b) and LLaVA Prismatic (Karamcheti et al., 2024) series.

C. Visualization of Image and Text Features

In the main paper, we demonstrate that distances in text features contribute more significantly to the weak correlation of model performance across datasets than image features. To visualize this finding, we utilize t-distributed stochastic neighbor embedding (t-SNE) (Hinton & Roweis, 2002). Our results show that CLIP-L-14-336 (Radford et al., 2021) effectively captures distinctions between images from different datasets, with images from the same dataset forming distinct clusters. Additionally, the text features of visual question answering (VQA) and multiple-choice visual questioning (MCVQ) tasks are separated by the text encoder of CLIP. Notably, the text features of OCRVQA are distant from those of ChartQA, TextVQA, and DocVQA, despite all being VQA tasks. This finding supports our observation in the main paper that the effectiveness of ranking methods is influenced by dataset characteristics (*i.e.*, VQA *vs.* MCVQ).

D. Full Results of Correlation Study

In the following, we present the full results of correlation study for ranking different series of large multimodal models. We only show $\textbf{NLL}_{\text{F}}$, $\textbf{NLL}_{\text{P}}$, $\textbf{NLL}_{\text{min}}$, $\textbf{NLL}_{\text{mean}}$, $\textbf{Ent}_{\text{F}}$, $\textbf{Ent}_{\text{P}}$, $\textbf{Ent}_{\text{max}}$, $\textbf{Ent}_{\text{mean}}$, $\textbf{Sample}_{\text{BLEU}}$, $\textbf{Sample}_{\text{BERT}}$, $\textbf{Sample}^{*}_{\text{BERT}}$. ATC and Accuracy on the line are not included because their performance are computed by the average correlation strength across eight datasets.

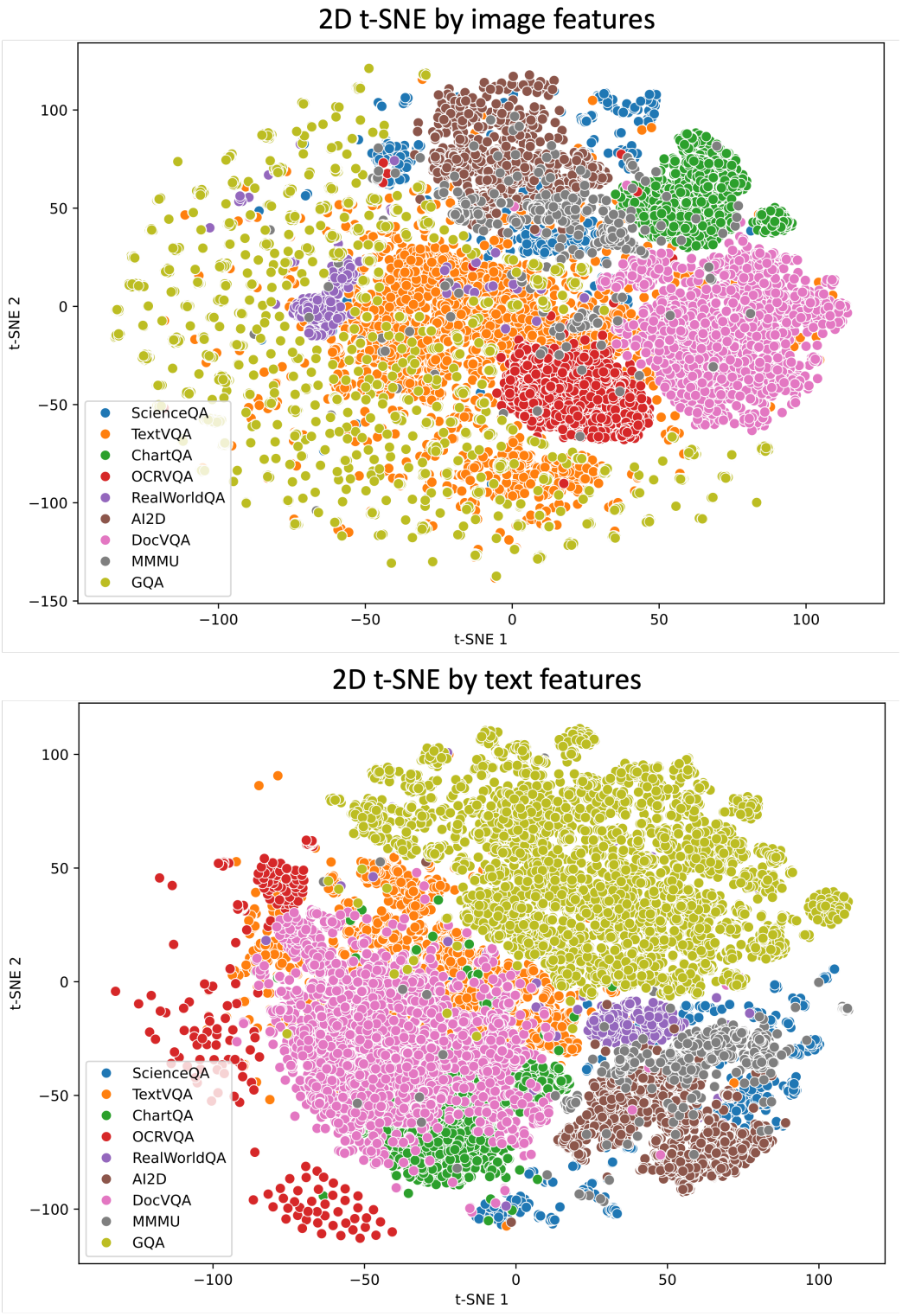


Figure 5. **Two t-SNE plots are presented: one using image features (Top) and the other using text features (Bottom) of the datasets.** We observe that the text features of OCRVQA are more scattered and significantly distant from those of other datasets.

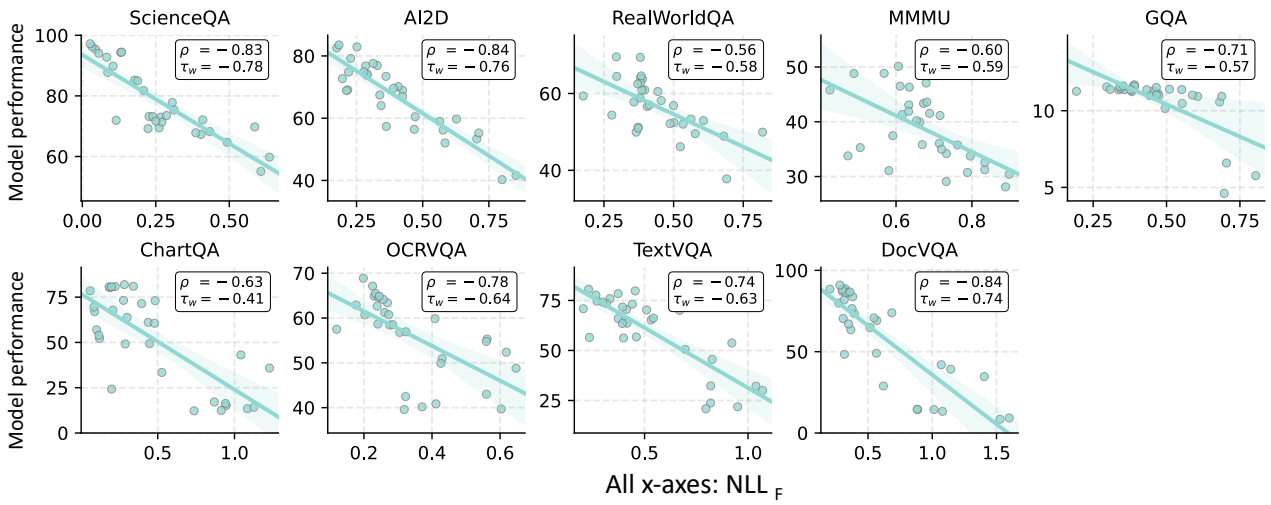


Figure 6. **Correlation study between NLL_F and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

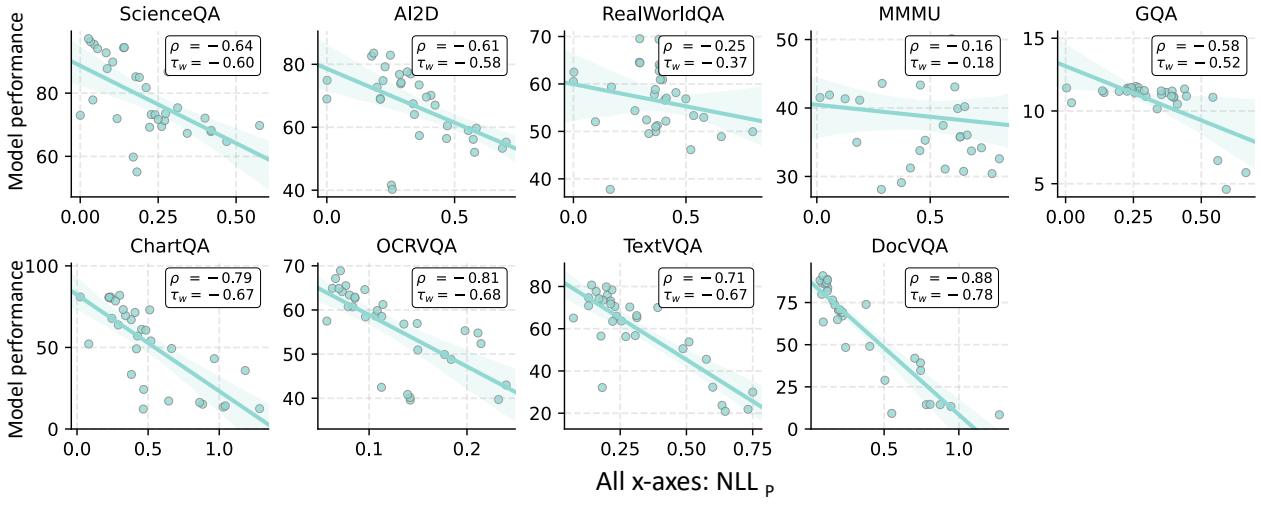


Figure 7. **Correlation study between NLL_P and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

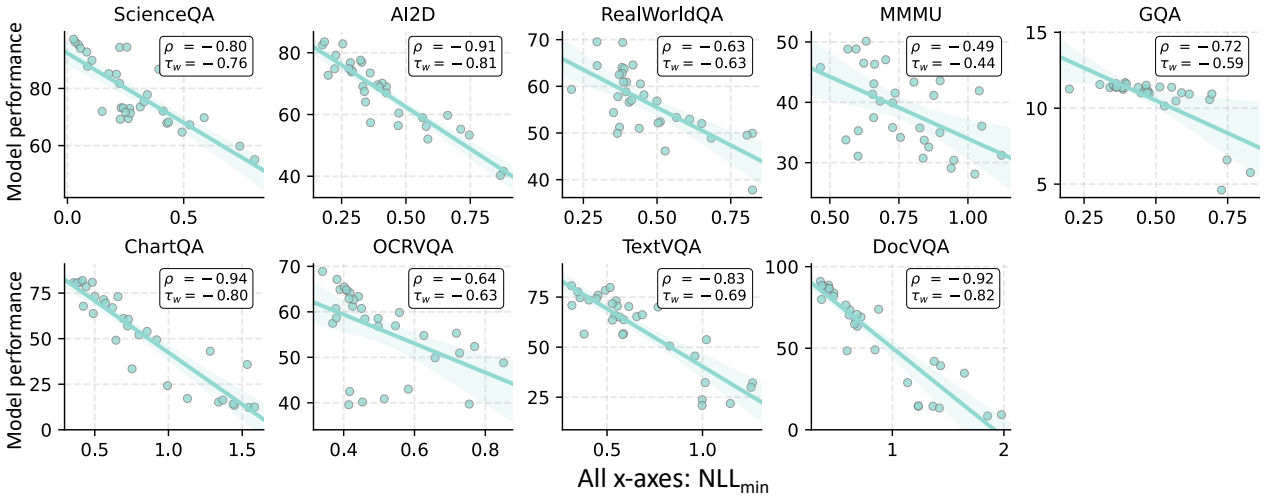


Figure 8. **Correlation study between NLL_{min} and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

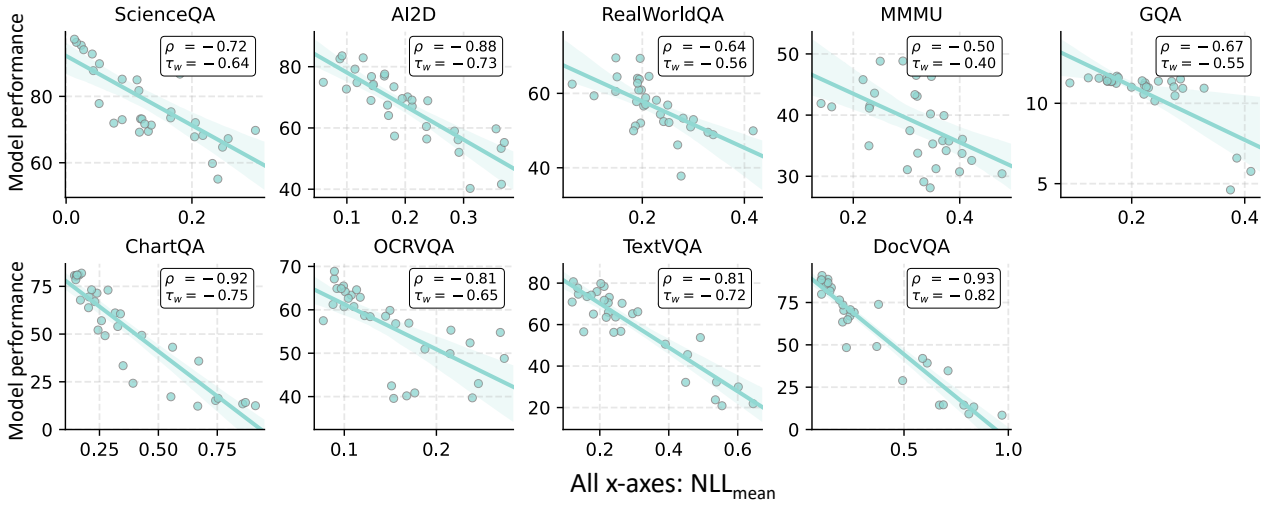


Figure 9. **Correlation study between NLL_{mean} and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

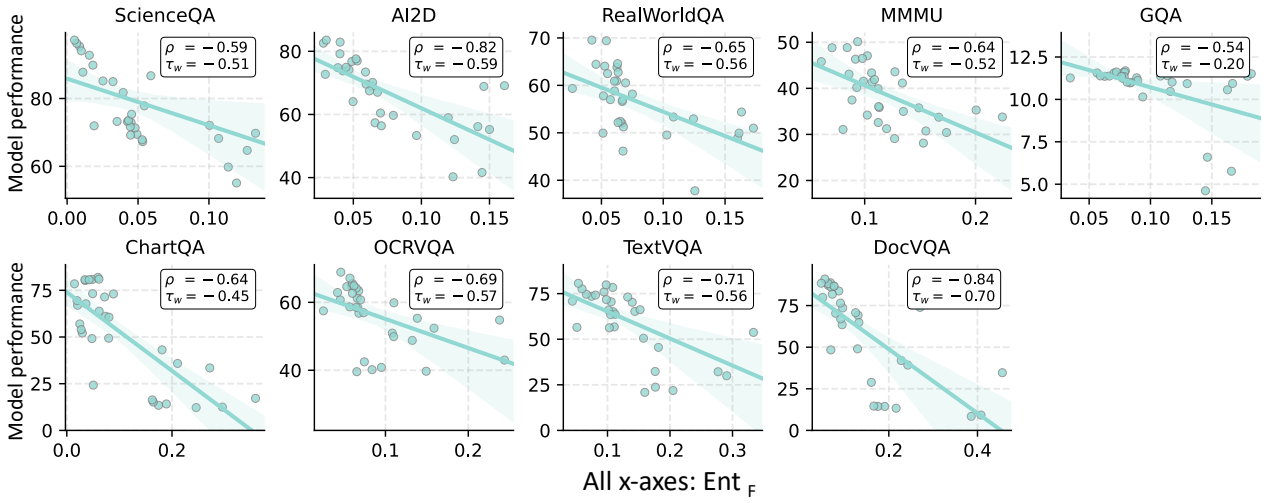


Figure 10. **Correlation study between Ent_F and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

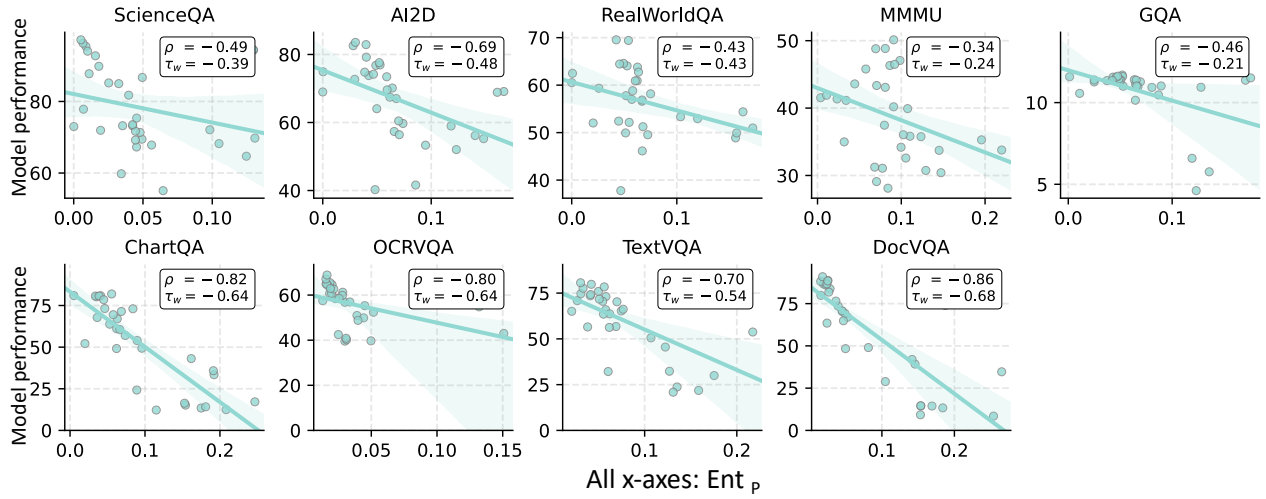


Figure 11. **Correlation study between Ent_P and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

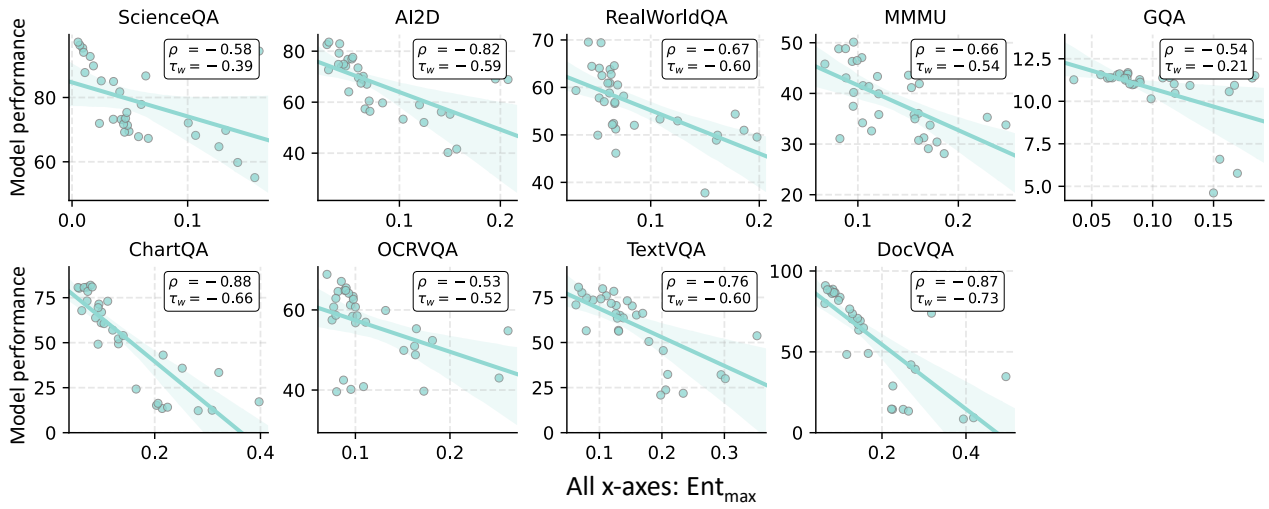


Figure 12. **Correlation study between Ent_{max} and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

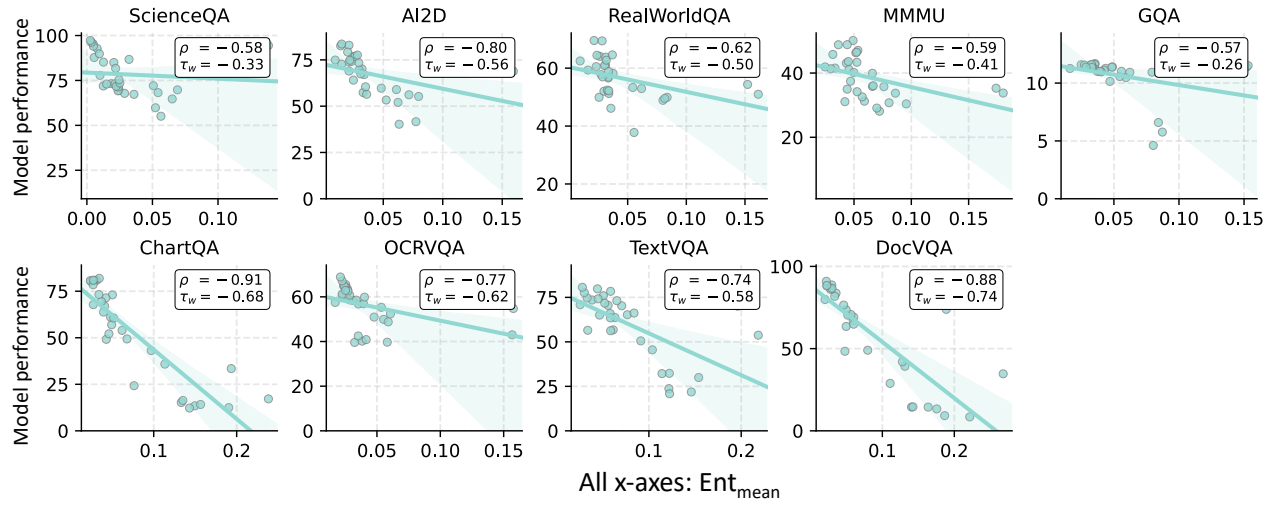


Figure 13. **Correlation study between Ent_{mean} and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

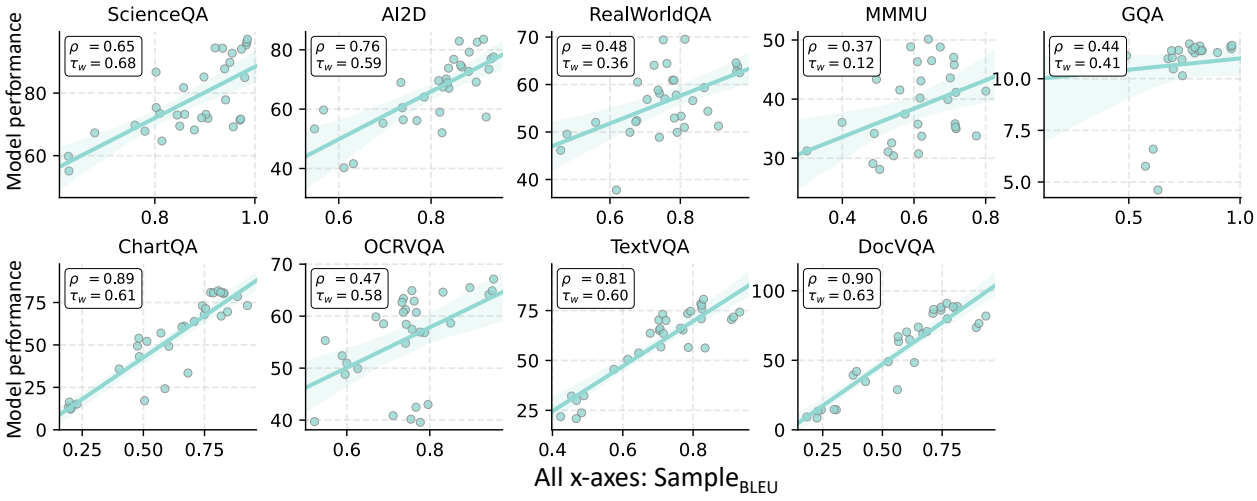


Figure 14. **Correlation study between Sample$_{BLEU}$ and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

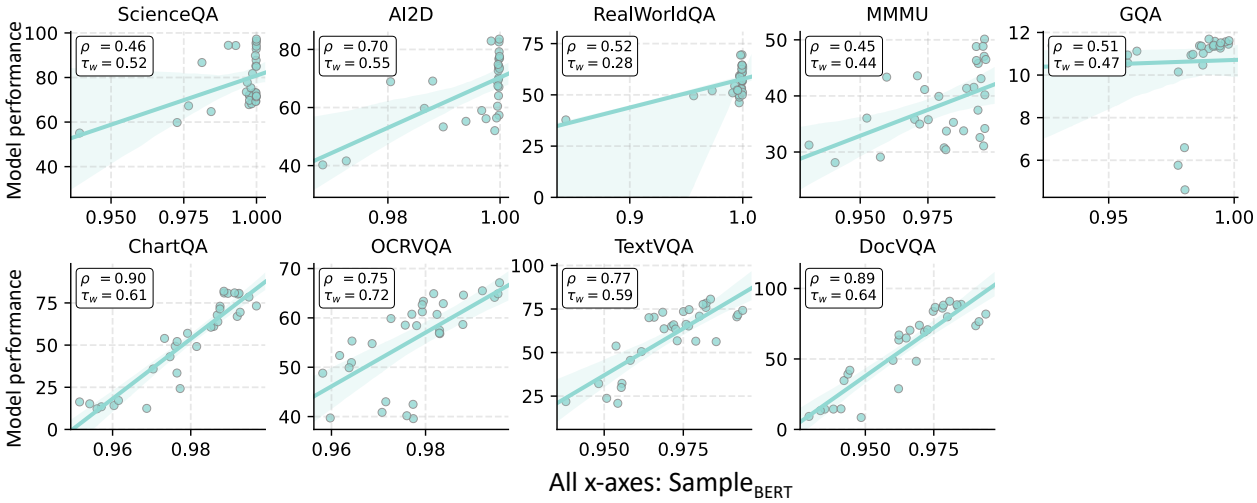


Figure 15. **Correlation study between Sample$_{BERT}$ and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).

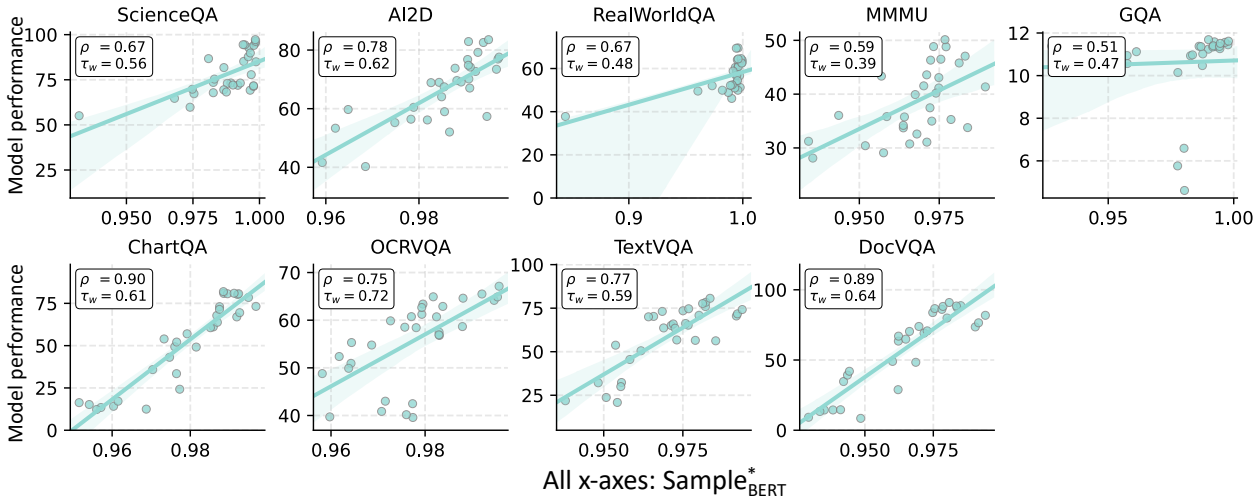


Figure 16. **Correlation study between Sample$^*_{\text{BERT}}$ and model performance.** Spearman's correlation (ρ) and weighted Kendall's correlation (τ_w) are metrics. Each point denotes a model. Straight lines are fit with robust linear regression (Huber, 2011).