

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

Yan Yang¹ Dongxu Li¹ ✉

Haoning Wu² Bei Chen Liu Liu³ Liyuan Pan⁴ ✉ Junnan Li⁵

¹ANU ²NTU ³KooMap, Huawei

⁴BITSZ & School of CSAT, BIT ⁵Salesforce AI Research

dongxuli1005@gmail.com liyuan.pan@bit.edu.cn

Project Page: <https://yan98.github.io/ProBench/>


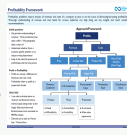
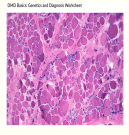
Coding; Screenshots and UI Elements	Knowledge; Document and Text-based Images	Science; Medical Images
<p>Query: i want you to write a Rshiny code in rstudio to generate above visualization. Can you do that?</p> 	<p>Query: Explain this framework to me in detail and in chronological order. I am an aspiring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

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

Abstract

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

1 Introduction

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

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

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

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

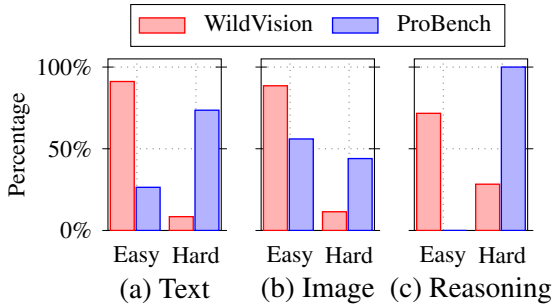


Figure 2: Comparison with WildVision (Lu et al., 2024) on challenge levels of (a) text, (b) image, and (c) reasoning for user instruction queries. To ensure a fair comparison, we follow WildVision by selecting the top 500 highest-quality queries from the single-round conversations.

mostly set in general chat environments and require much less domain knowledge to solve.

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

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

- we introduce ProBench, an open-ended multimodal benchmark tailored for professional work scenarios requiring expert-level knowledge, featuring 4,000 samples across 10 task fields over 56 sub-fields. The benchmark also features multi-round conversations up to 13 turns and multi-linguistic tracks in 17 lan-

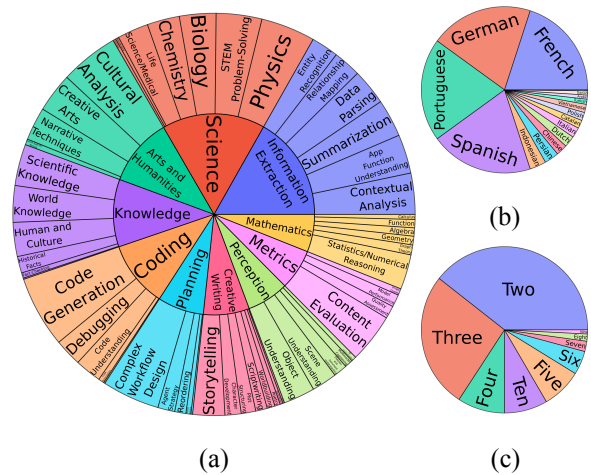


Figure 3: ProBench overview. Distributions of (a) task fields on the single-round track, (b) languages on the multi-linguistic track, and (c) conversation rounds on the multi-round tracks.

guages;

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

2 ProBench

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

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

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

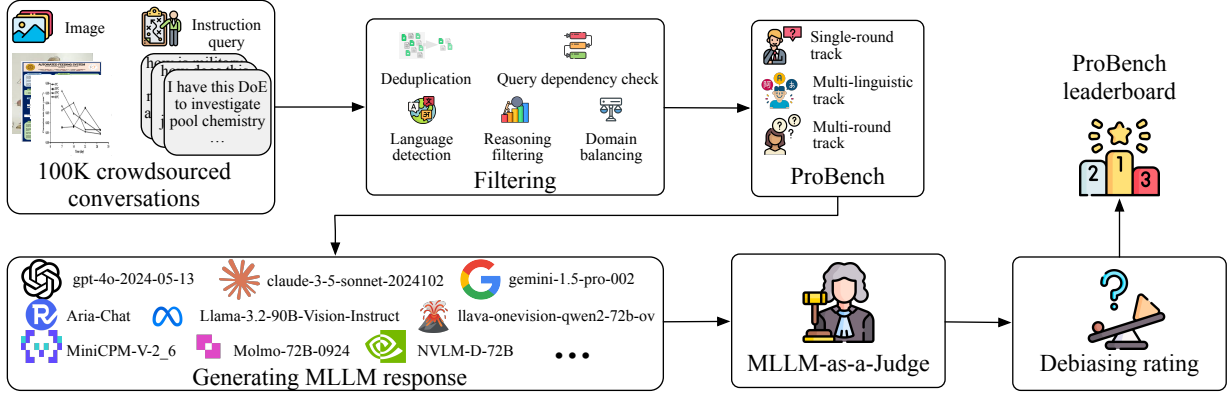


Figure 4: Framework of ProBench. Starting with 100K crowdsourced conversations, we identify high-quality user queries to curate single-round, multi-linguistic, and multi-round tracks. Using MLLM-as-a-Judge, we benchmark and rank 24 state-of-the-art MLLMs with ELO ratings. To ensure fairness, the ELO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

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

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

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

However, when using MLLM-as-a-Judge, the comparison results can be sensitive to model presentation order and confounded by response style variations (Li et al., 2024c). To address these challenges, the ProBench incorporates the Bradley-Terry model (Hunter, 2004) as an additional layer atop the ELO system. For N MLLMs and M pairwise comparisons, each round $1 \leq m \leq M$ compares model i and model j . We have $\mathbf{X}_m^{\text{win}} \in \mathbb{R}^N$ to indicate which model is presented first¹, while $\mathbf{X}_m^{\text{sty}} \in \mathbb{R}^S$ captures S stylistic differences between the outputs of models i and j (*e.g.*, word counts, and use of markdown). The Bradley-Terry model then refines the rating of model i as

$$r_i^{\text{ref}} = C + K \times \hat{\beta}_i, \\ \hat{\beta}, \hat{\gamma} = \arg \min_{\beta, \gamma} \sum_{m,i,j} \ell_{\text{bce}}(\beta^T \mathbf{X}_m^{\text{win}} + \gamma^T \mathbf{X}_m^{\text{sty}}, s_{i,j}),$$

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

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

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

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

2.1 Benchmark establishment

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

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

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

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

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

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

2.2 MLLM-as-a-Judge and ranking

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

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

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

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

3 Experiment

3.1 Experimental setup

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

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

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

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

Aria-Chat (Li et al., 2024b), InternVL2_5-8B (Wang et al., 2024b), InternVL2_5-26B (Wang et al., 2024b), InternVL2_5-38B (Wang et al., 2024b), InternVL2_5-78B (Wang et al., 2024b), Pixtral-12B-2409 (Agrawal et al., 2024), Pixtral-Large-Instruct-2411 (Agrawal et al., 2024), Qwen2-VL-7B-Instruct (Wang et al., 2024a), Qwen2-VL-72B-Instruct (Wang et al., 2024a), MiniCPM-V-2_6 (Yao et al., 2024), Llama-3.2-11B-Vision-Instruct (Dubey et al., 2024), Llama-3.2-90B-Vision-Instruct (Dubey et al., 2024), Molmo-7B-O-0924 (Deitke et al., 2024), Molmo-7B-D-0924 (Deitke et al., 2024), Molmo-72B-0924 (Deitke et al., 2024), NVLM-D-72B (Dai et al., 2024), llava-onevision-qwen2-7b-ov (Li et al., 2024a), and llava-onevision-qwen2-72b-ov (Li et al., 2024a).

3.2 Experimental result

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

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

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

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

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

3.3 Ablation and discussion

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

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

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

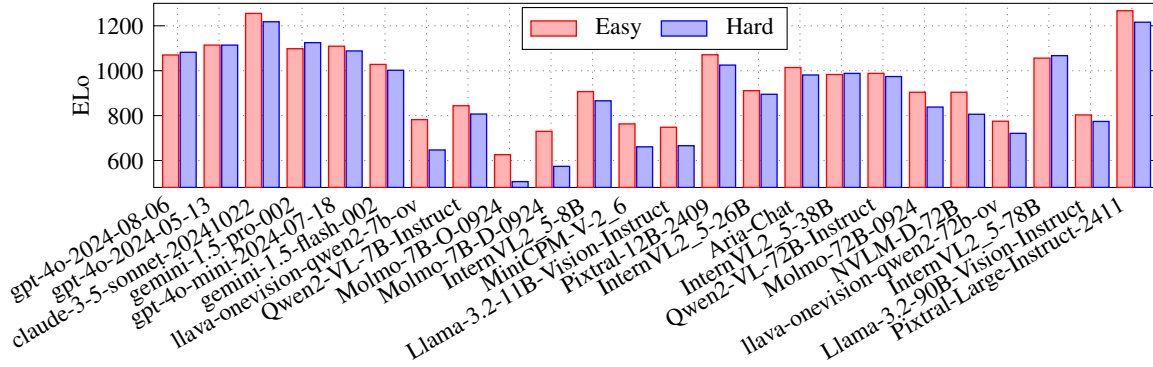


Figure 5: Ablation study of reasoning challenge. We show the ELO ratings of MLLMs on two levels: easy and hard.

	Compl.	Concis.	Corre.	Helpf.	Relve.	Text.	Perc.	Reas.	Know.	Reje.
🌟 claude-3-5-sonnet-20241022	53.06%	7.73%	24.01%	12.40%	2.80%	21.25%	33.97%	32.95%	3.82%	8.02%
🌈 gemini-1.5-pro-002	37.91%	8.28%	30.28%	19.36%	4.17%	24.89%	36.50%	30.80%	3.91%	3.91%
🌀 gpt-4o-2024-08-06	52.44%	8.88%	19.60%	17.44%	1.64%	20.95%	40.85%	29.71%	6.23%	2.25%
🌀 gpt-4o-mini-2024-07-18	29.88%	11.82%	39.78%	12.18%	6.34%	21.12%	36.64%	38.51%	3.54%	0.20%
🌈 gemini-1.5-flash-002	38.03%	6.38%	27.64%	23.42%	4.53%	28.55%	25.16%	29.06%	6.92%	10.31%
📦 Pixtral-Large-Instruct-2411	21.73%	8.94%	55.77%	9.15%	4.40%	20.28%	39.59%	36.48%	3.65%	0.00%
🌀 InternVL2_5-78B	46.58%	5.79%	28.37%	16.95%	2.32%	23.39%	37.51%	34.51%	4.06%	0.53%
🌀 Qwen2-VL-72B-Instruct	35.35%	4.75%	43.24%	13.84%	2.82%	21.85%	34.33%	37.21%	5.88%	0.72%
📦 Molmo-72B-0924	47.77%	3.05%	37.09%	10.35%	1.74%	21.95%	38.24%	32.93%	4.58%	2.29%
🌀 NVLM-D-72B	42.97%	5.19%	34.16%	14.11%	3.57%	29.86%	31.35%	30.98%	7.19%	0.62%
🌀 Llama-3.2-90B-Vision-Instruct	35.71%	3.42%	41.82%	13.46%	5.59%	21.46%	33.97%	26.48%	5.21%	12.88%
📦 llava-onevision-qwen2-72b-ov	48.43%	2.30%	34.38%	13.11%	1.78%	24.67%	38.80%	28.02%	4.19%	4.31%
📦 Pixtral-12B-2409	25.85%	7.16%	51.44%	10.90%	4.65%	23.39%	37.51%	34.51%	4.06%	0.53%
🌀 InternVL2_5-38B	49.97%	4.49%	28.93%	14.81%	1.80%	24.28%	39.48%	29.14%	6.10%	1.00%
🌀 Aria-Chat	32.22%	6.21%	48.15%	9.77%	3.66%	21.08%	40.46%	32.57%	5.60%	0.28%
🌀 InternVL2_5-26B	42.54%	3.49%	38.94%	12.70%	2.33%	22.86%	32.91%	34.92%	8.15%	1.16%
🌀 Llama-3.2-11B-Vision-Instruct	31.17%	3.85%	49.05%	11.71%	4.22%	21.33%	31.95%	34.69%	5.66%	6.37%
🌀 InternVL2_5-8B	34.01%	6.33%	42.65%	13.04%	3.97%	20.86%	38.12%	33.53%	6.99%	0.50%
🌀 Qwen2-VL-7B-Instruct	47.23%	5.49%	31.93%	12.98%	2.37%	26.16%	32.70%	32.80%	7.95%	0.40%
🌀 MiniCPM-V-2_6	40.48%	3.32%	41.28%	12.60%	2.32%	24.20%	33.11%	32.34%	7.15%	3.19%
📦 llava-onevision-qwen2-7b-ov	28.76%	5.34%	44.58%	16.61%	4.71%	30.56%	31.45%	27.89%	5.74%	4.35%
📦 Molmo-7B-D-0924	33.44%	2.27%	51.73%	8.77%	3.79%	27.06%	32.03%	29.16%	8.89%	2.87%
📦 Molmo-7B-O-0924	36.37%	1.54%	46.92%	12.30%	2.86%	29.76%	30.85%	29.86%	7.94%	1.59%

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

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

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

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

3.4 Distilled local evaluator

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

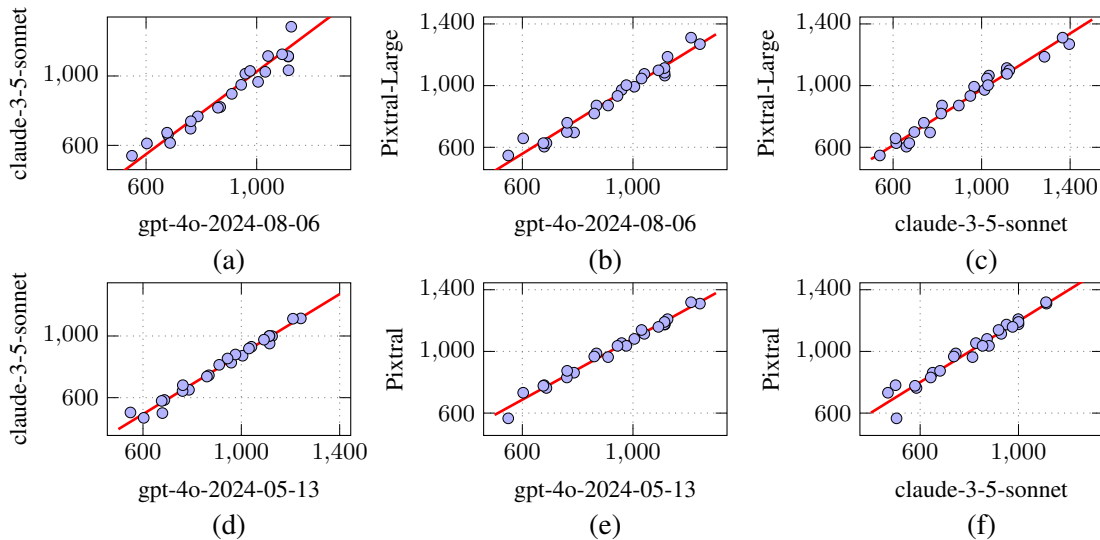


Figure 7: Ablation study of MLLM-as-the-Judge. (a-c) Pairwise comparisons of Elo scores for MLLMs evaluated using different MLLM judges. They are gpt-4o-2024-08-06, claude-3-5-sonnet-20241022 (claude-3-5-sonnet), and Pixtral-Large-Instruct-2411 (Pixtral-Large), respectively. (d-f) Comparison of using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

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

4 Related work

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

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

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

5 Conclusion

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

References

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

- Ziyang Luo, Haoning Wu, Dongxu Li, Jing Ma, Mohan Kankanhalli, and Junnan Li. 2024. Videoautoarena: An automated arena for evaluating large multimodal models in video analysis through user simulation. *arXiv preprint arXiv:2411.13281*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.
- Shanghaoran Quan, Jiayi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, et al. 2025. Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings. *arXiv preprint arXiv:2501.01257*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.
- T Wolf. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.