

Lost in Translation: Do LVLM Judges Generalize Across Languages?

Md Tahmid Rahman Laskar^{‡,*}, Mohammed Saidul Islam[‡],
Mir Tafseer Nayeem[§], Md Amran Bhuiyan[‡], Mizanur Rahman[‡],
Shafiq Joty^{§,¶}, Enamul Hoque^{‡,*}, Jimmy Xiangji Huang^{‡,*}
[‡]York University, [§]University of Alberta,
[§]Nanyang Technological University, [¶]Salesforce AI Research

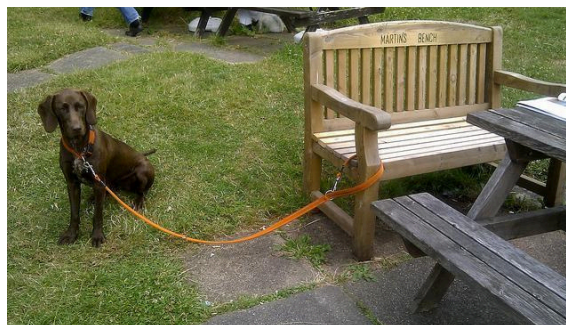
Abstract

Automatic evaluators such as reward models play a central role in the alignment and evaluation of large vision–language models (LVLMs). Despite their growing importance, these evaluators are almost exclusively assessed on English-centric benchmarks, leaving open the question of how well these evaluators generalize across languages. To answer this question, we introduce **MM-JudgeBench**, the first large-scale benchmark for multilingual and multimodal judge model evaluation, which includes over 60K pairwise preference instances spanning 25 typologically diverse languages. MM-JudgeBench integrates two complementary subsets: a general vision–language preference evaluation subset extending VL-RewardBench, and a chart-centric visual–text reasoning subset derived from OpenCQA, enabling systematic analysis of reward models (i.e., LVLM judges) across diverse settings. We additionally release a multilingual training set derived from MM-RewardBench, disjoint from our evaluation data, to support domain adaptation. By evaluating 22 LVLMs (15 open-source, 7 proprietary), we uncover substantial cross-lingual performance variance in our proposed benchmark. Our analysis further shows that model size and architecture are poor predictors of multilingual robustness, and that even state-of-the-art LVLM judges exhibit inconsistent behavior across languages. Together, these findings expose fundamental limitations of current reward modeling and underscore the necessity of multilingual, multimodal benchmarks for developing reliable automated evaluators.

1 Introduction

Recent advances in Large Language Models (LLMs) have substantially broadened their multilingual and multimodal capabilities by enabling

* Contact Emails: {tahmedge,enamulh,jhuang}@yorku.ca



Language: English

Query: What are the main objects or subjects in the image? Please describe them in detail.

Response A: The main objects in the image are a dog with an orange leash that's attached to a bench. The setting appears to be outdoors, possibly in a park-like environment, as there's grass and a wooden bench visible.

Response B: The main objects in the image are a dog, a wooden bench with the inscription "WATSON BOWL", and a table or picnic bench. The dog is seated on the grass next to the bench and seems to be tied or restrained by an orange leash that's attached to the bench.

Language: French

Translated Query: Quels sont les principaux objets ou sujets de l'image ? Veuillez les décrire en détail.

Translated Response A: Les principaux objets de l'image sont un chien avec une laisse orange attachée à un banc. Le cadre semble être en extérieur, peut-être dans un environnement semblable à un parc, car on voit de l'herbe et un banc en bois.

Translated Response B: Les principaux objets de l'image sont un chien, un banc en bois portant l'inscription « WATSON BOWL » et une table ou un banc de pique-nique. Le chien est assis sur l'herbe à côté du banc et semble attaché ou retenu par une laisse orange fixée au banc.

Example of a Correct Judge Response in English: {"winner": "A", "reasoning": "Answer A accurately describes the dog, orange leash, wooden bench, and outdoor setting. Answer B includes a hallucinated bench inscription ('WATSON BOWL')."}.

Example of an Incorrect Judge Response in French: {"winner": "B", "reasoning": "La réponse B fournit une description plus détaillée en identifiant correctement l'inscription sur le banc et le mobilier supplémentaire, ce qui la rend plus précise que la réponse A."}

Figure 1: Example from VL-RewardBench subset of MM-JudgeBench illustrating multilingual evaluation of LVLM judges for a given image. The question and candidate responses are translated from English to French. The LVLM judge (Gemini-2.5-Flash-Lite) selects the correct response A for English, while incorrectly selects B for French, highlighting the need for multilingual evaluation of LVLM judges.

reasoning over diverse languages and visual modalities (Yin et al., 2024; Shohan et al., 2024; Qin et al., 2025). As these systems are deployed at scale, their development and alignment rely critically on *automated evaluators*, most commonly implemented through reward models (Ouyang et al., 2022; Bai et al., 2022) or LLM/LVLM-as-a-judge frameworks (Zheng et al., 2023; Li et al., 2025a). In practice, automated evaluators now play a central role throughout the LVLM develop-

ment cycle, from guiding training and alignment to selecting models and benchmarking competing systems. Despite this central role, the evaluation of reward models or LVLMM judges remains overwhelmingly *English-centric*. Vision–language reward benchmarks such as VL-RewardBench (Li et al., 2025b) and Multimodal RewardBench (Yasunaga et al., 2025) focus exclusively on English, while multilingual extensions such as M-RewardBench (Gureja et al., 2025) are limited to text-only inputs (Lambert et al., 2025). As a result, no existing benchmark enables a unified study of reward models operating jointly across *languages* and *modalities*, precisely the setting in which these evaluators are expected to be increasingly applied.

To address this gap, we introduce **MM-JudgeBench**, the first large-scale benchmark for multilingual and multimodal judge model evaluation. MM-JudgeBench unifies two complementary subsets within a single framework: (i) a multilingual extension of VL-RewardBench covering vision–language preference judgments, and (ii) a chart-centric visual–text reasoning subset derived from OpenCQA (Kantharaj et al., 2022), which has been widely used to evaluate LVLMM judges on structured multimodal inputs (Laskar et al., 2025a,b). Together, these subsets span 25 typologically diverse languages and over 60K high-quality preference instances, enabling systematic analysis of LVLMM judges across multilingual and multimodal settings (see Figure 1). To support LVLMM judge improvement in multilingual settings, we also release a multilingual training set derived from Multimodal RewardBench (Yasunaga et al., 2025).

We construct MM-JudgeBench by translating all benchmark instances using *Gemini-3-Pro*¹, selected after empirical validation as a high-quality multilingual translation model. This allows us to isolate cross-lingual evaluation effects while minimizing translation noise. Using this unified benchmark, we conduct a large-scale evaluation of 22 state-of-the-art LVLMMs, including leading proprietary models, as well as open-source models spanning model sizes from 1B to 32B parameters.

Our evaluation reveals several key findings. First, while many existing LVLMMs report strong average accuracy, this overall score hides meaningful differences across languages, including clear drops in performance for certain languages.

For instance, efficiency-optimized model variants often suffer severe multilingual performance collapse despite strong English performance. Second, among open models, Qwen3-VL exhibits the most consistent multilingual behavior, outperforming many larger alternatives. Third, beyond accuracy, we uncover pronounced biases and instruction-following failures, demonstrating that correctness alone is insufficient to guarantee reliable reward-based evaluation. Finally, we show that reasoning-augmented judging and domain-adaptive fine-tuning on multilingual reward data offer noticeable performance gain. Our main contributions are listed as follows:

- **MM-JudgeBench**, the first large-scale benchmark for *multilingual and multimodal* evaluation of LVLMM judges, covering VL-RewardBench and OpenCQA across 25 typologically diverse languages and over 60K preference instances.
- A **large-scale empirical evaluation of 22 state-of-the-art LVLMMs**, spanning both proprietary and open models across a wide range of model sizes and architectures, revealing cross-lingual performance variations that are invisible under English-only evaluation. In addition, we reveal the scaling behavior of LVLMM judges in multilingual settings, alongside various biases.
- In addition, we release a *multilingual training set covering 100k preference instances* derived from MM-RewardBench to support **domain adaptive fine-tuning** as an improvement strategy for cost-efficient open models.

To support reproducibility and further research, we publicly release MM-JudgeBench along with the evaluation code at <https://github.com/tahmedge/mm-judgebench>.

2 Related Work

Reward models are a cornerstone of modern LLM development (Laskar et al., 2023; Jahan et al., 2023), particularly for aligning models with human preferences (Christiano et al., 2017; Ouyang et al., 2022). As a result, the quality of the reward model directly influences not only the downstream performance of aligned LLMs but also their safety and reliability. This dependency has elevated the evaluation of reward models as a critical research problem, motivating the development of dedicated evaluation benchmarks.

¹<https://deepmind.google/models/gemini/pro/>

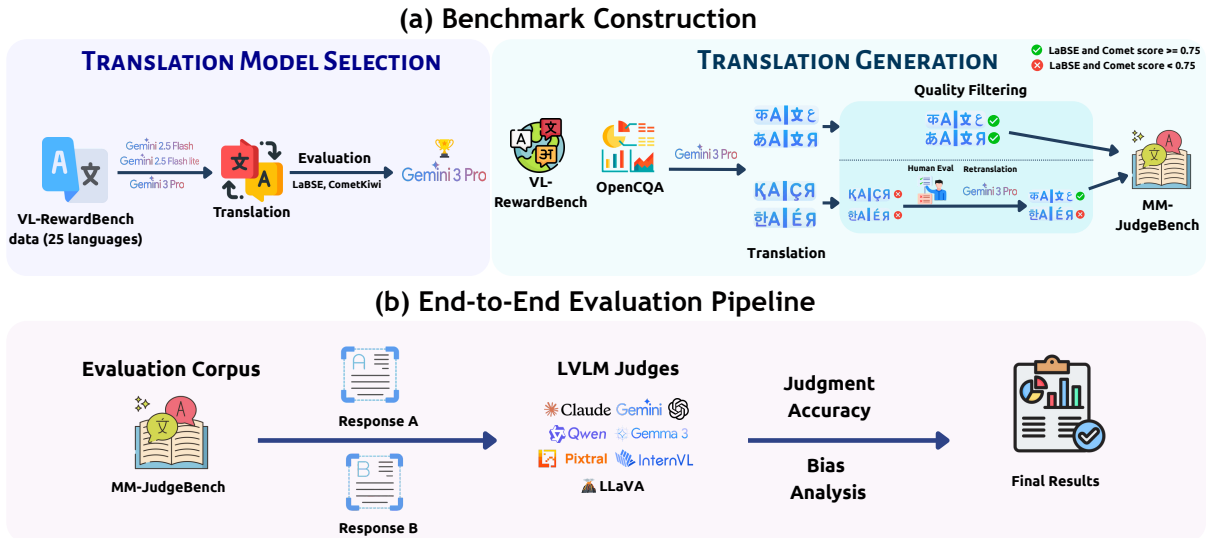


Figure 2: An overview of our methodology: (a) *Benchmark construction* step contains two stages, i.e., Translation model selection, and Translation data generation (from VL-Reward Bench and OpenCQA data); and (b) *End-to-End evaluation pipeline*.

The first systematic benchmark proposed for this purpose was RewardBench (Lambert et al., 2025). However, this benchmark was restricted to text-based modality on English-centric data. Consequently, the rapid advancement in multilingual and multimodal LLMs has underscored the need for benchmarks that can robustly assess model capabilities across diverse languages and input modalities (Yin et al., 2024; Qin et al., 2025). To this end, benchmarks like VL-RewardBench (Li et al., 2025b) and M-RewardBench (Gureja et al., 2025) attempted to mitigate this gap by utilizing multimodal and multilingual data, respectively.

However, the above-mentioned benchmarks are also limited in scope. For instance, while VL-RewardBench provides a challenging set of examples for evaluating vision-language reward models across complex reasoning tasks, it is restricted to only English-centric data. On the other hand, while M-RewardBench highlights the need for linguistically diverse benchmarks by extending RewardBench to a multilingual setting, it is restricted to text-based modality. To our best knowledge, there are no multimodal reward benchmarks that are also multilingual. Our work addresses this gap by directly extending VL-RewardBench by translating it across 25 languages, providing a comprehensive multilingual benchmark for the evaluation of multimodal reward models.

Beyond reward modeling, the closely related LVLm-as-a-judge paradigm has emerged, where capable LVLms serve as evaluators for responses

generated by other vision-language systems (Chen et al., 2024). Despite growing adoption, existing studies evaluating LVLm judges are also limited to the general multimodal task evaluation in English (Lee et al., 2024; Xiong et al., 2025). We address this gap by utilizing the OpenCQA dataset (Kantharaj et al., 2022), which is an open-domain chart question answering dataset and used by Laskar et al. (2025a,b) for the evaluation of LVLms-as-the-judge. We extend our contribution to this paradigm by also translating OpenCQA across the same 25 languages, enabling more comprehensive cross-lingual assessment of LVLm judges, as recent work underscores the need for chart-centered evaluations of LVLms (Mahbub et al., 2025a,b; Laskar et al., 2025a,b).

Beyond proprietary judges, open-source models such as LLaVA-Critic (Xiong et al., 2025) and Prometheus-Vision (Lee et al., 2024) have been trained specifically for vision-language evaluation. However, these models are developed and assessed exclusively in English. To support the development of multilingual multimodal judge models, we construct the multilingual version of MM-RewardBench (Yasunaga et al., 2025) by extending it into 24 languages.

3 Methodology

Our methodology for creating and evaluating the multilingual multimodal reward model benchmark involved *three* main stages: translation model selection, dataset construction, and model evalua-

tion. Figure 2 demonstrates an overview of our methodology, which we also describe below (furthermore, section 4.5 discusses our domain adaptive fine-tuning approach).

3.1 Translation Model Selection

To create a comprehensive multilingual benchmark, we follow the work of Gureja et al. (2025) and selected the 23 languages they used in M-RewardBench. However, since none of their selected languages are considered low-resource, we added two languages: *Bengali* and the low-resource *Kazakh* to ensure a more robust evaluation of LVLMs. The selected 25 languages cover a diverse range of language families and scripts: *Arabic (ar)*, *Bengali (bn)*, *Chinese (zh)*, *Czech (cs)*, *Dutch (nl)*, *English (en)*, *French (fr)*, *German (de)*, *Greek (el)*, *Hebrew (he)*, *Hindi (hi)*, *Indonesian (id)*, *Italian (it)*, *Japanese (ja)*, *Kazakh (kk)*, *Korean (ko)*, *Persian (fa)*, *Polish (pl)*, *Portuguese (pt)*, *Romanian (ro)*, *Russian (ru)*, *Spanish (es)*, *Turkish (tr)*, *Ukrainian (uk)*, and *Vietnamese (vi)*.

For the translation process, we utilized Gemini-3-Pro. We select this model for translation due to the impressive performance of the Pro version of Gemini in WMT25 (Kocmi et al., 2025).

To validate our choice of Gemini-3-Pro as the translation model, we conducted a comparative analysis of translation quality across 25 languages in VL-RewardBench against Gemini-2.5-Pro alongside two other cheaper Gemini variants: Gemini-2.5-Flash and Gemini-2.5-Flash-Lite. First, the query and the pairwise responses were translated. Then, the quality of the translations was assessed using two standard translation quality estimation metrics: Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022) and CometKiwi (Rei et al., 2022). As shown in Table 1, the Pro versions (Gemini-3-Pro and Gemini-2.5-Pro) achieved the highest scores on both metrics, confirming their suitability for translation (see Appendix A.7 for the translation quality of OpenCQA). While both Gemini-3-Pro and Gemini-2.5-Pro perform identically, we selected Gemini-3-Pro since it is the latest version of the Gemini-Pro series. This helps ensure long-term reproducibility, as older closed models are often deprecated.

3.2 Data Construction

Translation process. For data translation, in each dataset, we create a prompt for Gemini-3-Pro to translate the query (if available) and the

Model	LaBSE	CometKiwi
Gemini-3-Pro	0.91	0.85
Gemini-2.5-Pro	0.91	0.85
Gemini-2.5-Flash	0.88	0.83
Gemini-2.5-Flash-Lite	0.79	0.74

Table 1: Translation quality comparison of Gemini models in VL-RewardBench.

corresponding answers generated by two different models for each image from English to 24 other languages. For VL-RewardBench and MM-RewardBench, we translate the answers that are publicly available in the dataset. For OpenCQA, we use the answers generated using Gemini-1-Pro and Claude-3-Haiku by Islam et al. (2024) and then translate them. For all benchmarks, we translate the queries provided in each dataset. We denote the multilingual (M) versions as M-VL-RewardBench, M-OpenCQA, and M-MM-RewardBench.

Prompt design and API efficiency. Given the per-day rate limit of API calls² for powerful closed models, we ask the translator LLM (e.g., Gemini-3-Pro) to translate the query (where available) and the answers from English to 24 other languages within a single prompt (see Appendix A.1 for details). The translator LLM was required to generate the response in an Array of JSON objects format with the following keys: **(i)** Translated Query, **(ii)** Translated Answer A, **(iii)** Translated Answer B, **(iv)** Language. Then, we write a parsing script to extract the translated data for each language from the JSON-formatted output. This enables the reduction of API calls by 24 times.

Quality filtering and final datasets. To ensure high quality of the translated datasets, we have also added a quality filtering step. Inspired by prior threshold-based filtering work (Schwenk, 2018; Batheja and Bhattacharyya, 2023), samples having LaBSE and Comet scores below 0.75 are first manually inspected by a human via back translation. If the human reviewer finds that the translation is bad, we re-translate that sample by tuning the decoding parameters. After re-translation, if the score is still below 0.75, we remove that sample. Our final datasets contain 31K samples from VL-RewardBench, 30K samples from OpenCQA, and over 100K samples from MM-RewardBench.

²<https://ai.google.dev/gemini-api/docs/rate-limits>

3.3 LVLM Evaluation

We select a wide range of both proprietary and open-source vision-language models to evaluate in our proposed MM-JudgeBench benchmark. The models include: GPT-5³ and its variants (Mini and Nano), Gemini-2.5-Flash and Gemini-2.5-Flash-Lite (Comanici et al., 2025), Claude-4.5-Haiku⁴, Grok-4.1-Fast⁵, Gemma-3 (4B, 12B, 27B) (Team et al., 2025), InternVL-3.5 (Wang et al., 2025) (1B, 2B, 4B, 8B, 14B), Qwen3-VL (Bai et al., 2025) (2B, 30B-A3B, 4B, 8B, 32B), and Pixtral-12B (Agrawal et al., 2024). We also select the LLaVA-Critic-7B (Xiong et al., 2025) model, which is an LVLM trained for reward modeling.

For each model and language, we provide the query and two candidate answers to the LVLM judge and prompt it to select the better answer. In addition, the LVLM is also asked to provide a rationale behind their selection since adding rationale helps improve the performance of reward models or judges (Xiong et al., 2025).

LVLM Judge Prompt

You are a strict and fair judge for vision-language tasks. You will be shown an image and a user question, plus two candidate answers A and B. Decide which answer is better based on the following criteria.

- Correctness with respect to the image and question.
- Completeness and level of detail.
- Relevance and clarity (no unnecessary verbosity).
- Avoiding hallucinations or unsupported claims.

Return ONLY a JSON object with this schema:
{"winner": "A | B", "reasoning": "brief reason"}

For evaluation, we measure the pairwise accuracy, which is the percentage of times the LVLM correctly identified the preferred response in a given pair. Additionally, we investigate the presence of biases like position and length bias. *Position bias* refers to the tendency of a model to prefer answers based on their position in the input (e.g., always choosing the first or second response). To measure this, we present each pair of answers to the models in both the original and a reversed order. The position bias is then calculated as the overall difference in judgment accuracy between the two orderings. *Length bias* is measured by evaluating whether the model prefers an answer that is longer than the other answer but incorrect.

³<https://openai.com/gpt-5/>

⁴<https://www.anthropic.com/claude/haiku>

⁵<https://x.ai/news/grok-4-1-fast>

4 Results and Discussion

We conduct a series of experiments to evaluate various LVLM judges in MM-JudgeBench.

Evaluation Settings: The evaluation considers assessing the pairwise accuracy, focusing on factual correctness, completeness, and relevance. To ensure robustness, we generate the judgment by providing the candidate answers in both the original and the reversed order independently, and then report the average accuracy. In addition, we evaluate biases like positional bias and length bias. As the LVLMs generate the winner alongside the reasoning, we parse the LVLM-judge predicted winner from their corresponding JSON-formatted responses using a parsing script (Laskar et al., 2024). If the parsing script cannot properly parse the judgment from the response, we consider the LVLM-generated answer wrong. All the open models were run using vLLM (Kwon et al., 2023) on an A100 GPU, with the decoding parameters being primarily set to the default values in HuggingFace (Wolf et al., 2020) for the respective open-source models. For the closed models, decoding parameters are also set to the default values as provided by the respective model providers. For all supporting models, the temperature value is set to 0. Below, we demonstrate our findings.

4.1 Main Results on M-VL-RewardBench

Table 2 reports the accuracy of vision-language models evaluated on the M-VL-RewardBench subset of MM-JudgeBench across 25 languages. In addition, we report the per-model average (Avg) and variance (Var) across languages. Below, we summarize our key findings.

Closed models dominate, but efficient variants are uneven. GPT-5 achieves the highest overall accuracy (81.3%) with the lowest variance (0.2), remaining stable from the high-resource English (81.7%) to the low-resource Kazakh (80.6%). Cost-effective closed models such as GPT-5-Mini, GPT-5-Nano, Gemini-2.5-Flash, and Grok-4.1-Fast retain reasonable accuracy (71–78%) with variance mostly below 1 (except GPT-5-Nano). However, some cost-efficient variants perform quite poorly: Gemini-2.5-Flash-Lite collapses to 40.8% (variance 2.6), and Claude-4.5-Haiku drops to 56.4% (variance 2.1).

Qwen3-VL is the strongest open family and scales consistently. Qwen3-VL-32B achieves the best performance among open models (68.8% on M-VL-RewardBench), with accuracy improving

Model	Avg	Var	ar	bn	zh	cs	nl	en	fr	de	el	he	hi	id	it	ja	kk	ko	fa	pl	pt	ro	ru	es	tr	uk	vi
Closed-source																											
GPT-5	81.3	0.2	81.4	80.6	81.6	81.6	82.5	81.7	81.6	81.4	81.4	81.7	81.0	81.0	81.6	81.5	80.6	81.1	81.3	81.9	82.0	81.6	80.6	80.7	80.9	81.4	80.8
GPT-5-Mini	78.1	0.4	78.4	76.8	77.8	78.1	78.6	78.6	78.1	78.4	78.5	78.2	77.6	77.4	78.7	78.6	76.6	78.2	77.5	78.6	79.2	78.6	78.3	78.6	77.7	78.2	77.6
GPT-5-Nano	73.2	1.2	73.5	72.1	73.9	73.5	74.3	71.2	74.3	74.7	70.5	72.5	72.9	72.9	74.5	73.1	71.2	73.9	73.6	73.4	74.2	72.7	74.7	73.8	73.4	73.3	72.6
Gemini-2.5-Flash	76.7	0.5	77.6	75.4	77.0	77.2	76.7	78.6	76.0	76.4	76.4	77.3	76.7	77.3	77.6	76.0	75.7	77.0	75.9	77.1	76.7	76.7	77.1	76.8	76.3	76.0	76.9
Gemini-2.5-Flash-Lite	40.8	2.6	35.8	39.5	40.3	42.0	41.8	40.2	41.7	40.3	39.7	40.2	41.7	42.8	40.4	42.4	37.9	42.0	41.5	41.3	42.1	42.5	42.2	41.9	38.9	40.6	39.7
Claude-4.5-Haiku	56.4	2.1	57.5	55.7	55.3	55.9	55.9	54.6	58.3	57.1	58.7	57.3	56.7	56.3	57.1	58.6	51.9	57.5	55.3	55.2	57.1	55.8	58.0	57.1	55.6	55.8	56.8
Grok-4.1-Fast	71.3	0.7	72.4	70.4	70.2	71.7	72.7	72.0	71.9	71.5	70.8	71.3	71.3	71.0	73.0	70.6	69.2	70.8	71.4	72.0	71.5	71.5	71.4	71.3	70.7	69.8	71.0
Open-source																											
Qwen3-VL-32B	68.8	3.3	68.1	66.4	69.4	69.4	70.1	69.6	69.8	70.2	68.8	67.0	68.2	69.9	70.5	69.1	62.0	68.8	68.1	69.8	70.1	69.6	70.2	69.9	66.9	69.8	69.2
Qwen3-VL-8B	62.2	1.8	61.8	59.8	62.6	62.8	63.1	62.9	62.9	62.9	60.3	61.2	61.8	61.9	63.1	63.4	58.5	61.6	62.2	62.1	64.8	61.6	63.1	64.2	63.2	62.1	61.9
Qwen3-VL-4B	61.9	2.5	62.3	61.0	63.4	60.6	63.3	62.5	64.4	62.8	60.9	59.9	61.2	61.2	63.1	63.6	57.3	63.4	60.6	60.5	64.3	61.2	62.3	63.0	61.7	60.9	62.7
Qwen3-VL-30B-A3B	63.4	2.2	63.5	61.3	63.7	62.6	64.4	65.3	64.7	63.9	62.2	63.3	62.2	63.8	65.1	63.8	58.3	63.7	62.6	63.8	64.9	62.8	64.0	65.4	62.6	63.7	64.4
Qwen3-VL-2B	54.3	1.6	53.8	55.7	53.1	53.7	54.8	55.5	55.2	55.3	53.6	52.2	55.4	55.4	54.9	54.5	50.5	53.4	52.7	53.8	55.7	55.0	55.4	55.0	54.8	54.0	54.7
InternVL-3.5-14B	58.8	3.1	58.3	57.7	58.7	59.3	58.9	62.0	59.9	59.9	57.5	59.0	57.8	57.7	59.3	57.8	53.3	58.5	59.6	58.5	61.4	59.6	60.2	60.9	55.7	58.7	59.1
InternVL-3.5-8B	52.9	0.9	53.7	53.8	51.8	51.6	52.6	53.8	52.1	51.9	54.6	53.3	55.0	52.1	52.1	53.6	52.6	53.0	54.6	52.9	52.3	52.4	53.2	53.0	52.2	52.1	53.1
InternVL-3.5-4B	54.8	1.0	55.4	53.6	55.3	54.2	54.2	57.1	54.8	55.2	56.9	54.2	55.5	54.0	54.4	53.0	54.2	54.7	54.8	53.7	56.0	55.0	55.2	55.9	53.7	54.2	53.8
InternVL-3.5-2B	53.6	1.2	53.5	52.6	52.8	54.1	53.9	57.5	54.4	53.8	52.4	52.6	52.8	53.2	53.8	53.0	54.7	53.5	51.4	53.6	54.3	53.2	54.3	54.4	53.0	52.8	53.8
InternVL-3.5-1B	50.0	3.8	50.1	49.8	50.2	49.6	49.6	54.3	47.7	49.0	54.9	50.1	49.8	47.7	49.1	49.6	54.3	50.2	51.3	48.8	50.6	48.4	49.3	49.5	50.4	48.8	47.0
Gemma-3-27B	45.0	4.0	43.7	42.4	45.0	47.4	46.0	50.2	46.8	45.0	44.8	44.3	42.5	45.6	46.2	44.6	40.1	43.7	42.6	46.2	46.2	45.5	46.2	46.7	43.3	45.6	44.6
Gemma-3-12B	42.0	0.8	41.1	41.4	41.1	41.4	41.8	44.0	41.3	42.6	44.4	42.3	40.7	41.7	41.6	41.4	41.5	42.2	42.9	41.8	42.5	43.1	42.3	41.9	41.4	42.5	41.3
Gemma-3-4B	34.9	1.7	35.9	31.7	33.9	35.4	34.6	36.1	35.3	35.7	36.3	33.8	31.8	35.9	34.6	34.8	34.0	33.5	35.5	35.3	35.5	35.9	33.6	36.1	35.6	33.8	36.7
LLaVA-Critic-7B	49.5	3.4	49.1	49.8	50.3	51.6	48.6	50.1	47.4	49.2	51.6	49.2	51.8	48.5	48.8	48.8	52.5	48.2	51.0	49.0	47.7	52.2	48.2	44.1	51.4	49.2	49.2
Pixtral-12B	41.7	1.5	39.7	41.6	43.3	42.6	41.7	43.2	43.1	42.7	42.4	40.7	41.6	41.4	44.3	43.1	40.7	39.9	41.3	41.0	41.5	40.6	41.4	41.5	42.5	39.4	42.3

Table 2: Overall Performance on the M-VL-RewardBench subset. Here, ‘Avg’ and ‘Var’ are computed across languages, and lower is better for ‘Var’. Green denotes the best and Red denotes the worst in each row.

monotonically from 54.3% at 2B to 68.8% at 32B. Even the smaller Qwen3-VL-4B (61.9%) surpasses all InternVL-3.5 and Gemma-3 variants as well as Pixtral-12B, and several Qwen3 variants outperform the weaker closed models (Claude-4.5-Haiku, Gemini-2.5-Flash-Lite). Surprisingly, the reward-specialized LLaVA-Critic-7B scores below 50%, suggesting poor generalization.

Closed-source versus open-source gap. Our evaluation reveals a clear performance gap between the strongest closed and open models. For instance, the most optimized GPT-5 variant, GPT-5-Nano, outperforms the best open model, Qwen3-VL-32B, by approximately 4.5 points on average. However, many open models (e.g., Qwen3) substantially outperform some closed models (e.g., Claude-4.5-Haiku and Gemini-2.5-Flash-Lite). This indicates that architectural design and training strategies could play a critical role (see Appendix A.10 for further analysis). More broadly, upstream tokenization likely also contributes, since compression-oriented metrics alone miss systematic cross-lingual fragmentation (Alqahtani et al., 2026; Nayeem et al., 2025).

Performance variation across languages. A key finding from our evaluation on the VL-RewardBench subset is the inclusion of the low-resource Kazakh language, in which most LVLMS achieve the poorest performance, as demonstrated in color red in Table 2. This suggests that current training strategies may not adequately capture the linguistic diversity required for robust multilingual reward model performance. We also observe that

English tends to achieve the best accuracy in comparison to other languages for most models.

4.2 Main Results on M-OpenCQA

We next evaluate LVLMS on the M-OpenCQA subset of MM-JudgeBench that focuses on chart-centric visual-text reasoning. Unlike M-VL-RewardBench, M-OpenCQA does not provide gold preference labels. To enable systematic evaluation, we select GPT-5, which is the strongest model on M-VL-RewardBench, as a high-quality reference judge to annotate pairwise preferences⁶. Consequently, we restrict this analysis to open-source LVLMS, examining how well they can replicate the judgments of a powerful proprietary model. Results are reported in Table 3.

Overall performance trends. Across model families, performance on M-OpenCQA in terms of both accuracy and variance is usually better than on M-VL-RewardBench. This may reflect the increased difficulty in M-VL-RewardBench, which consists of diverse multimodal tasks.

Qwen-3-VL again leads among open models. Among open models, Qwen3-VL-32B again achieves the best result (accuracy 67.4%, variance 1.4), with all Qwen3 variants above 2B exceeding 60% while also scaling consistently from 2B to 32B. InternVL-14B is competitive on average (66.3%) but less stable (variance 2.1). Other open families lag behind, with all Gemma-3 models, Pixtral-12B, as well as the specialized LLaVA-

⁶We evaluate GPT-5 judgments on M-OpenCQA using 3 human evaluators having expertise in NLP, Data Science, and Computer Vision. We randomly collect 300 samples from OpenCQA and find that 93.5% of the time, at least 2 annotators agree with GPT-5 judgment.

Model	Avg	Var	ar	bn	zh	cs	nl	en	fr	de	el	he	hi	id	it	ja	kk	ko	fa	pl	pt	ro	ru	es	tr	uk	vi
Qwen3-VL-32B	67.4	1.4	67.3	65.7	69.0	67.4	68.5	70.9	67.6	67.1	66.9	67.0	67.0	67.8	68.4	66.7	64.7	66.1	66.7	67.2	68.0	67.8	67.7	68.3	66.2	67.0	67.8
Qwen3-VL-8B	64.5	0.5	65.2	65.2	64.6	64.8	63.3	66.3	64.0	64.6	65.3	64.0	64.0	63.5	64.2	64.4	63.9	64.2	64.4	64.8	64.2	63.9	64.4	64.0	65.2	65.4	64.2
Qwen3-VL-4B	62.7	0.8	64.1	62.4	63.7	61.2	63.2	64.2	62.9	62.3	62.4	63.0	61.8	63.6	62.6	63.3	60.3	62.9	61.8	62.4	63.3	63.1	63.0	63.0	62.0	62.6	62.7
Qwen3-VL-30B-A3B	63.7	0.9	64.1	62.4	64.7	64.6	63.8	66.6	64.0	63.4	62.6	63.3	62.6	64.2	64.4	63.6	62.0	63.6	62.9	64.2	64.1	63.3	63.5	64.2	63.8	62.9	63.5
Qwen3-VL-2B	56.2	0.4	55.7	57.7	56.0	55.7	56.0	57.0	56.1	56.5	55.7	55.2	57.4	56.5	56.6	56.5	55.2	56.7	55.7	56.2	55.6	56.4	55.6	56.9	55.2	56.2	56.3
InternVL-3.5-14B	66.3	2.1	65.9	64.1	67.8	66.7	65.8	71.1	66.8	67.8	65.0	65.5	66.0	67.5	67.3	66.4	63.4	66.6	65.6	66.3	67.3	67.0	66.2	66.7	64.8	65.7	65.4
InternVL-3.5-8B	63.1	0.4	63.3	63.5	63.2	63.0	61.9	64.6	63.7	63.1	61.7	62.3	63.6	62.8	64.2	63.3	62.2	63.5	62.3	63.0	63.2	62.6	63.0	63.7	63.2	63.4	63.3
InternVL-3.5-4B	62.7	2.0	63.1	62.2	63.3	61.1	64.7	65.7	63.9	63.0	61.0	62.0	61.7	63.6	64.0	62.7	59.4	62.4	60.9	61.8	64.0	62.0	63.7	64.7	62.0	62.4	62.8
InternVL-3.5-2B	57.7	1.9	58.2	55.1	59.6	57.0	58.4	60.6	59.6	58.0	56.4	57.3	56.0	57.2	58.2	59.3	56.0	58.4	55.5	56.5	58.6	58.9	58.0	58.0	57.7	56.8	57.4
InternVL-3.5-1B	55.3	2.4	55.0	53.2	56.7	54.3	55.4	57.3	56.5	56.1	54.2	56.3	53.9	56.0	56.6	55.2	50.4	53.8	54.5	56.3	56.7	56.0	55.1	57.3	54.0	56.5	55.4
Gemma-3-27B	57.0	3.5	58.3	55.3	57.8	57.4	56.5	64.1	57.0	56.7	55.6	56.0	56.4	56.1	59.1	55.2	55.2	55.3	56.0	57.0	57.5	58.7	57.2	56.1	55.5	56.8	58.1
Gemma-3-12B	59.1	0.2	58.7	59.6	58.8	59.6	59.2	60.4	59.3	59.3	59.0	59.7	58.8	59.0	58.7	58.5	58.8	58.5	58.9	59.7	58.7	59.0	58.8	59.8	58.4	58.6	58.7
Gemma-3-4B	57.0	0.6	57.1	55.7	57.4	57.7	57.4	58.0	57.8	57.7	56.8	56.2	57.2	56.8	57.3	57.0	55.6	56.6	56.2	57.2	57.2	57.7	56.2	57.6	57.0	54.9	56.6
LLaVA-Critic-7B	57.0	2.3	57.6	56.1	56.9	56.6	57.2	56.0	58.8	57.6	55.4	56.4	56.4	57.9	58.2	54.2	53.9	59.1	56.6	57.6	57.2	58.2	59.2	59.2	54.3	57.2	58.4
Pixtral-12B	56.8	0.7	56.6	56.8	56.7	57.5	56.5	60.0	57.0	57.3	57.4	56.4	56.1	56.4	56.8	57.2	55.9	56.4	56.3	57.0	57.7	57.2	56.4	56.9	56.0	56.0	56.5

Table 3: Overall performance on the M-OpenCQA subset. ‘Avg’ and ‘Var’ are computed across languages, and lower is better for Var. **Green** denotes the best and **Red** denotes the worst per row.

Model	M-OpenCQA	M-VL-RewardBench	Average
Qwen3-VL-32B	1.28 0.43	5.58 3.91	3.43 2.18
Qwen3-VL-8B	1.98 0.84	2.49 1.40	2.24 1.12
Qwen3-VL-4B	0.91 0.59	36.36 43.97	18.64 22.28
Qwen3-VL-30B-A3B	3.42 0.43	9.00 7.00	6.21 3.72
Qwen3-VL-2B	2.90 1.57	18.12 52.94	10.51 27.26
InternVL-3.5-14B	2.34 1.02	5.35 14.36	3.84 7.69
InternVL-3.5-8B	1.76 1.43	22.92 21.10	12.34 11.26
InternVL-3.5-4B	2.17 1.40	38.45 27.96	20.31 14.68
InternVL-3.5-2B	5.87 3.26	55.90 25.95	30.88 14.60
InternVL-3.5-1B	6.20 2.44	30.60 136.10	18.40 69.27
Gemma-3-27B	1.47 0.47	8.24 6.53	4.86 3.50
Gemma-3-12B	0.92 0.57	8.07 8.79	4.50 4.68
Gemma-3-4B	1.19 0.44	12.94 18.13	7.06 19.28
LLaVA-Critic-7B	1.98 2.68	15.82 180.05	8.90 91.36
Pixtral-12B	3.50 1.71	40.56 11.93	22.03 16.82

Table 4: Avg. Position Bias and Variance (separated by ‘|’, *left* part denotes position bias and *right* part denotes position bias variance) for the open-source LVLMs in the M-VL-RewardBench. Lower values indicate better with **Green** indicates the best and **Red** indicates the worst, per column.

Critic-7B, achieving below 60%. In summary, these trends mirror M-VL-RewardBench, confirming Qwen3’s robustness across datasets.

Language-specific patterns persist. Similar to M-VL-RewardBench, English consistently achieves the highest accuracy across most models (13 out of 15 times). Meanwhile, low-resource languages like Kazakh continue to present challenges, with most models achieving their lowest scores on this language (10 out of 15 times). However, the variance in per-language performance is generally lower on M-OpenCQA compared to M-VL-RewardBench, suggesting that chart reasoning is less sensitive to language-specific factors.

4.3 Bias Analysis

Table 4 summarizes the average positional bias for open-source LVLMs on both M-VL-RewardBench and M-OpenCQA. In terms of the average across both datasets, we find that the Qwen3-VL-8B is the least prone to positional bias (only 2.24%). Moreover, Qwen3-VL-32B shows the lowest bias variance (2.18%), demonstrating judgment capability with less bias across languages. In summary, we observe that M-

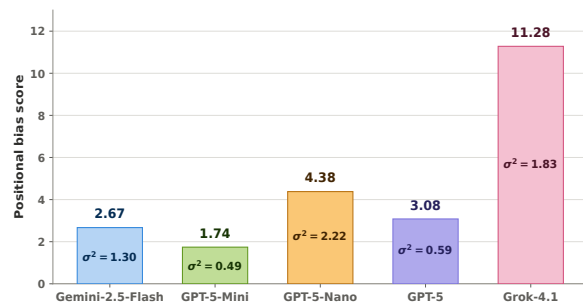


Figure 3: Position Bias in multilingual M-VL-RewardBench for closed models. Lower values indicate better.

OpenCQA usually has lower position bias and variance in comparison to M-VL-RewardBench. Figure 3 further shows positional bias trends for top-performing closed models (above 70% accuracy) on M-VL-RewardBench. We observe that the GPT-5-Mini exhibits the lowest positional bias, whereas Grok-4.1-Fast shows very high positional bias. Other models (GPT-5 and Gemini-2.5-Flash) also demonstrate low position bias and variance. Taken together, these results reveal that robustness to positional coherence is highly model- and dataset-dependent, with only a small subset of LVLMs demonstrating bias-resilient judging behavior across multilingual and multimodal settings. These highlight that only accuracy is insufficient to guarantee reliable reward modeling. In Appendix A.2, we further analyze the cross-lingual amplification of the positional bias. In Appendix A.3, we also report the length bias.

4.4 Impact of Reasoning

For each LVLm judge, we follow prior work (Laskar et al., 2025a) to generate the preferred answer alongside the reasoning behind choosing the particular answer in JSON format (Appendix A.4 also evaluates the capabilities of different models in generating the output in the required format). We further investigate the performance by

Model Family	4B	8B	32B/14B
Qwen3-VL-Instruct	70.70	71.64	71.66
Intern-VL-Instruct	47.49	64.59	64.95

Table 5: Reasoning quality in M-OpenCQA in terms of LaBSE similarity scores for the top-performing open-source models (Qwen3 and InternVL-3.5) across different sizes. Here, 32B denotes Qwen3 while 14B denotes InternVL.

comparing the direct generation of the judgment against the generation of the judgment alongside the reasoning. We evaluate the Qwen3-VL-2B and Qwen3-VL-8B models with and without reasoning to find that the overall accuracy is degraded by 2.2% and 4%, respectively, if no reasoning is used when LVLMs generate the judgment. This validates the utilization of reasoning in judgment.

In Table 5, we further evaluate the reasoning quality via LaBSE scores on M-OpenCQA and find that Qwen3-VL consistently outperforms comparable InternVL models (71.64 vs. 64.59 at 8B) and shows consistent scaling (70.70 at 4B to 71.66 at 32B) with diminishing returns beyond 8B. InternVL exhibits noticeable gain from 4B (47.49) to 8B (64.59), but a narrow gain to 14B.

4.5 Domain Adaptation via Fine-Tuning

We fine-tune the Qwen3-VL-8B-Instruct model in our training subset: M-MM-RewardBench. Since M-MM-RewardBench does not have the reference reasoning for the corresponding judgment, we apply supervised fine-tuning only on the judgment labels and train the model to directly predict the judgment without any reasoning. We ran a total of 2 epochs, with the learning rate being set to $2e - 5$. On M-VL-RewardBench, the fine-tuned model improves by +14% over direct prompting and +10% over rationale-augmented prompting (both in zero-shot). These substantial improvements demonstrate the effectiveness of our training subset in improving the judging performance across other datasets, opening up the possibility of using weak supervision techniques to further improve the performance (Laskar et al., 2020, 2022).

4.6 Task-Level Performance Analysis

The M-VL-RewardBench subset contains three task categories from the original VL-RewardBench: (i) General Multimodal Instructions, (ii) Hallucination-Oriented Queries, and (iii) Mathematical Reasoning. Below, we conduct a task-level performance analysis.

Based on the average across all models, we find that hallucination-oriented queries have an aver-

Model	Hallucination	Mathematical	General
Qwen3-VL-32B	73.42	68.51	49.71
Qwen3-VL-8B	65.14	64.39	47.51
Qwen3-VL-4B	64.66	63.06	48.75
Qwen3-VL-2B	55.59	58.01	44.15
GPT-5	83.55	87.68	60.94
GPT-5-Mini	81.26	81.83	58.31
GPT-5-Nano	76.61	74.45	56.56

Table 6: Per Task Accuracy by Model in M-VL-RewardBench.

age accuracy of 73.42, followed by mathematical reasoning (68.51) and general multimodal instructions (49.71). These results indicate that structured evaluation scenarios (hallucination detection, mathematical verification) are currently better suited to LVLm-as-a-judge evaluation, while open-ended multimodal instruction tasks remain challenging. Table 6 further reports task-level accuracy for the best-performing closed (GPT-5) and open (Qwen3-VL) model series. We observe that even the best-performing GPT-5 model achieves only about 60% accuracy in the general multimodal instruction following task. We further demonstrate the English vs. non-English performance gap across tasks in Appendix A.9.

4.7 Translation Sensitivity Analysis

To verify that our findings are not artifacts of a specific translation system, we repeat the Qwen3-VL evaluation on M-VL-RewardBench using two weaker translators (Gemini-2.5-Flash and Gemini-2.5-Flash-Lite) in addition to our default Gemini-3-Pro. Figure 4 shows that weaker translations monotonically reduce average accuracy (e.g., Qwen3-VL-32B: $68.8 \rightarrow 64.4 \rightarrow 58.1$) and inflate cross-lingual variance sharply, with Qwen3-VL-2B variance jumping from 1.6 to 13.2 under the weakest translator. Crucially, model rankings remain preserved across all three translation systems, indicating that the cross-lingual trends reported in Tables 2 and 3 reflect genuine judge behavior rather than translation artifacts.

4.8 Additional Results

Beyond using M-MM-RewardBench as a training resource, we benchmark LVLm judges on it to verify that the training corpus is substantive. Given its scale (100K samples), we exclude some proprietary models (GPT-5, Gemini-2.5-Flash, Claude-4.5-Haiku, Grok-4.1-Fast) due to high API cost, and report results for the remaining LVLms in Table 7. We observe that model rankings largely

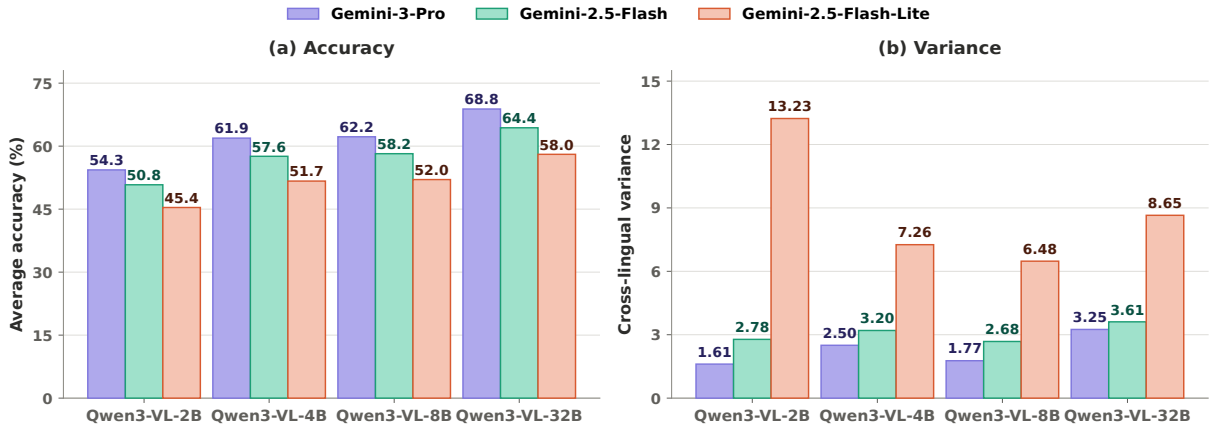


Figure 4: Translation sensitivity analysis on M-VL-RewardBench with Gemini models as the translator and Qwen3-VL as the judge. Higher is better for Accuracy, while Lower is better for Variance.

Model	Avg	Var	ar	bn	zh	cs	nl	en	fr	de	el	he	hi	id	it	ja	kk	ko	fa	pl	pt	ro	ru	es	tr	uk	vi
Closed-source																											
GPT-5-Mini	70.9	0.1	70.6	71.1	71.1	70.5	70.8	71.3	70.2	70.9	70.7	71.0	70.7	70.8	71.6	70.4	70.7	71.3	71.1	70.7	71.2	71.2	71.5	71.2	70.7	70.3	71.3
GPT-5-Nano	65.4	0.4	65.0	65.1	65.6	65.2	66.6	67.0	66.0	65.3	64.6	64.9	64.2	64.9	66.1	64.9	65.3	65.5	65.1	65.7	66.1	65.1	65.7	65.0	65.4	65.4	65.8
Gemini-2.5-Flash-Lite	60.5	0.6	59.4	60.1	60.8	61.0	61.1	60.5	61.6	60.3	59.6	60.5	59.6	61.6	61.5	60.8	59.7	59.4	59.8	60.4	61.0	61.7	60.5	60.9	60.8	58.6	60.5
Open-source																											
Qwen3-VL-32B	65.4	0.9	65.7	64.8	65.1	65.4	66.0	65.9	66.1	65.5	65.7	65.3	65.7	66.2	66.2	66.3	61.2	65.2	64.8	65.5	66.0	65.6	65.3	65.8	65.1	65.7	65.3
Qwen3-VL-8B	60.9	6.3	61.8	56.4	62.6	59.0	63.3	63.2	60.7	61.9	58.5	61.0	57.5	63.4	61.8	62.5	52.1	61.9	61.1	62.0	62.0	62.2	61.9	62.2	61.3	59.5	62.7
Qwen3-VL-4B	57.4	9.3	60.7	55.0	61.0	53.4	58.7	61.2	54.5	58.9	53.1	58.6	55.7	61.8	59.4	60.9	48.5	56.5	57.1	56.0	56.0	56.3	58.5	57.2	57.3	57.3	61.0
Qwen3-VL-30B-A3B	61.1	5.7	62.2	56.5	60.2	60.6	62.6	63.2	63.0	62.1	61.3	61.3	57.7	61.9	62.4	62.2	52.0	61.2	62.3	61.2	61.5	62.6	62.5	62.1	61.0	61.9	62.8
Qwen3-VL-2B	51.4	3.9	51.3	48.6	54.3	50.5	51.8	53.8	53.3	53.1	47.4	49.8	49.6	53.1	52.5	51.2	46.3	51.9	51.0	51.5	53.2	49.5	51.5	54.1	51.6	50.9	52.4
InternVL-3.5-14B	58.4	5.9	58.9	53.6	62.4	56.6	57.9	61.7	60.7	59.9	55.8	58.2	58.0	59.6	59.9	61.1	51.0	59.5	58.4	57.2	58.0	57.0	60.5	58.1	57.7	58.3	60.4
InternVL-3.5-8B	58.2	12.7	60.2	53.4	62.3	56.0	57.3	64.3	56.6	59.5	49.7	57.4	56.7	61.5	61.1	61.0	49.5	60.0	56.3	60.5	58.5	57.9	61.7	57.1	58.1	55.3	61.9
InternVL-3.5-4B	54.4	12.7	57.2	47.8	60.1	50.9	55.2	59.4	55.3	56.2	47.3	54.6	53.5	58.4	55.1	56.8	46.4	55.6	53.2	51.4	55.7	52.7	56.5	56.0	57.0	50.7	58.0
InternVL-3.5-2B	51.2	5.4	51.6	49.9	57.0	49.4	53.8	54.9	51.1	52.7	44.9	52.1	51.3	53.1	52.1	50.5	48.2	48.7	49.0	51.0	51.4	50.4	53.5	51.5	50.5	49.5	51.3
InternVL-3.5-1B	46.4	6.6	48.4	43.4	51.1	42.0	43.3	47.4	46.6	44.2	43.8	49.5	48.3	48.5	49.5	47.7	45.0	48.2	46.9	42.6	47.1	45.4	43.4	47.2	48.2	43.0	50.2
Gemma-3-27B	61.9	0.9	62.5	61.3	62.2	62.2	63.0	63.3	61.4	61.0	60.7	62.6	62.4	63.0	62.0	62.9	59.3	62.7	62.2	62.0	61.7	60.2	61.7	61.3	61.6	61.8	62.8
Gemma-3-12B	57.6	3.6	58.4	56.5	59.6	57.7	57.7	58.8	56.7	57.5	55.0	58.5	57.4	60.2	59.0	58.6	50.1	58.6	57.7	59.1	57.1	57.0	58.9	58.4	56.4	56.1	57.9
Gemma-3-4B	48.3	2.3	50.0	47.2	48.1	46.4	47.2	48.0	48.9	48.3	46.5	49.9	52.1	51.0	47.0	49.3	45.0	49.4	48.5	48.8	47.1	46.9	49.5	48.2	47.6	48.2	49.5
Pixtral-12B	48.1	2.7	48.5	48.5	50.1	47.9	45.4	48.7	48.2	49.4	47.0	48.6	46.0	50.1	48.4	48.1	45.5	49.8	48.7	47.6	50.3	45.1	49.8	51.3	45.7	46.6	48.3

Table 7: Overall performance on the M-MM-RewardBench subset. ‘Avg’ and ‘Var’ are computed across languages; lower is better for ‘Var’. Green denotes the best and Red denotes the worst in each row.

mirror M-VL-RewardBench and M-OpenCQA: GPT-5-Mini leads overall (70.9%), Qwen3-VL-32B tops among the open models (65.4%) with consistent scaling, and Kazakh again yields the worst per-model accuracy for most LVLMs. One notable shift is that Gemma-3-27B becomes competitive (61.9%), suggesting its weakness on M-VL-RewardBench is task-specific. This analysis provides additional evidence that our cross-lingual findings reflect genuine judge behavior.

5 Conclusion and Future Work

We introduce **MM-JudgeBench**, a benchmark for evaluating LVLm-as-a-judge models in multilingual and multimodal settings. It comprises two evaluation subsets, M-VL-RewardBench (general multimodal reasoning) and M-OpenCQA (chart-specific visual-text reasoning), spanning 25 languages with over 60K samples, complemented by a 100K-sample multilingual training set derived from MM-RewardBench. Our evaluation of a broad range of closed and open models yields

three key insights: (i) flagship closed models achieve the strongest overall accuracy and cross-lingual stability, though some of their optimized variants often degrade substantially; (ii) among open models, the Qwen3-VL family shows the most consistent scaling behavior, at times rivaling closed-source alternatives; and (iii) performance varies markedly across languages, with low-resource languages such as Kazakh posing persistent challenges. Our robustness analyses further reveal biases that are highly model- and dataset-dependent, underscoring that accuracy alone is insufficient to ensure reliable judging. Overall, MM-JudgeBench exposes cross-lingual limitations of LVLm judges that remain invisible under English-only evaluation. While supervised fine-tuning on our training set improves open-model performance, future work will explore additional strategies such as reinforcement learning (Kaelbling et al., 1996), and extend MM-JudgeBench to more languages and evaluation tasks (Laskar et al., 2025c).

Limitations

MM-JudgeBench is derived by translating English-centric datasets into 24 additional languages. Despite rigorous translation model selection and quality filtering, some linguistic or cultural nuances may be lost. However, by controlling translation quality with human intervention and keeping visual content fixed, MM-JudgeBench isolates cross-lingual judge behavior, enabling a focused study of multilingual generalization. Although the benchmark spans 25 typologically diverse languages, it does not cover many other low-resource languages. However, we address a critical gap in prior work by incorporating low-resource languages like Kazakh while demonstrating the limitations of most models in this language. Furthermore, our benchmark is primarily constructed with LLM-generated translations. Nonetheless, we conduct human⁷ evaluation of the translated data (see Appendix A.6) and find that humans also rate our multilingual datasets with high ratings. While OpenCQA preference labels are generated using GPT-5 and our human evaluation confirms agreement with it as a high-quality reference judge, there are still some risks of having reference-model bias. Nonetheless, we further validated the choice of GPT-5 by comparing cross-judge consistency with another stronger LVLM, Gemini-2.5-Pro (Spearman $\rho = 0.93$, Pearson $r = 0.94$).

Acknowledgments

We thank all the anonymous reviewers and the area chair of ACL 2026 for their excellent review comments. This research is supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada, the York Research Chairs (YRC) program, Canada Foundation for Innovation (CFI), Google’s Gemini Academic Program for the API Credits, CUPE 3903 Research Grant, and Digital Research Alliance of Canada for the computing resources.

Ethics Statement

This work evaluates automated vision–language judges rather than deploying them. Our findings reveal substantial cross-lingual variance and bias,

⁷The human annotators possess strong English proficiency and hold graduate-level degrees from institutions where English is a primary academic language.

highlighting risks in using LVLM judges as substitutes for human evaluation in multilingual settings. MM-JudgeBench is designed to expose, not obscure, these risks by making multilingual failure modes measurable and transparent. Practitioners should avoid deploying judge models in high-stakes or user-facing scenarios without multilingual validation and human oversight. All data are derived from publicly available benchmarks, contain no personally identifiable information, and are processed using automated translation. The licensing requirements are maintained accordingly while using different tools. Additional human compensation is not required since it was done by two authors of this paper. We also provide necessary details related to our human evaluation studies (see Appendix A.5 and A.6). Finally, we used AI-based writing assistants only to improve the presentation of the paper.

References

- Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.
- Sawsan Alqahtani, Mir Tafseer Nayeem, Md Tahmid Rahman Laskar, Tasnim Mohiuddin, and M Saiful Bari. 2026. [Stop taking tokenizers for granted: They are core design decisions in large language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8432, Rabat, Morocco. Association for Computational Linguistics.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Akshay Batheja and Pushpak Bhattacharyya. 2023. “a little is enough”: Few-shot quality estimation based corpus filtering improves machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14175–14185.

- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-rewardbench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58.
- Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368.
- Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. [Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers](#). In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.
- Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.
- Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022. Opencqa: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, and 1 others. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413.
- 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.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797.
- Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. [A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816, Miami, Florida, USA. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. [A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2020. [WSL-DS: Weakly supervised learning with distant supervision for query focused multi-document abstractive summarization](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5647–5654, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.
- Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Ridwan Mahbub, Ahmed Masry, Mizanur Rahman, Amran Bhuiyan, Mir Tafseer Nayeem, Shafiq Joty,

- Enamul Hoque, and Jimmy Huang. 2025a. [Judging the judges: Can large vision-language models fairly evaluate chart comprehension and reasoning?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1203–1216, Vienna, Austria. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Ridwan Mahbub, Mizanur Rahman, Amran Bhuiyan, Israt Jahan, Mir Tafseer Nayeem, Shafiq Joty, Enamul Hoque, and Jimmy Huang. 2025b. [Deploying tiny LVLM judges for real-world evaluation of chart models: Lessons learned and best practices.](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1906–1918, Suzhou (China). Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Israt Jahan, Elham Dolatabadi, Chun Peng, Enamul Hoque, and Jimmy Huang. 2025c. [Improving automatic evaluation of large language models \(LLMs\) in biomedical relation extraction via LLMs-as-the-judge.](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25483–25497.
- Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. [Prometheus-vision: Vision-language model as a judge for fine-grained evaluation.](#) In *Findings of the association for computational linguistics ACL 2024*, pages 11286–11315.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025a. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge.](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.
- Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, and 1 others. 2025b. [V1-rewardbench: A challenging benchmark for vision-language generative reward models.](#) In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24657–24668.
- Ridwan Mahbub, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, and Enamul Hoque. 2025a. [The perils of chart deception: How misleading visualizations affect vision-language models.](#) In *2025 IEEE Visualization and Visual Analytics (VIS)*, pages 6–10.
- Ridwan Mahbub, Mohammed Saidul Islam, Mir Tafseer Nayeem, Md Tahmid Rahman Laskar, Mizanur Rahman, Shafiq Joty, and Enamul Hoque. 2025b. [From charts to fair narratives: Uncovering and mitigating geo-economic biases in chart-to-text.](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28929–28947, Suzhou, China. Association for Computational Linguistics.
- Mir Tafseer Nayeem, Sawsan Alqahtani, Md Tahmid Rahman Laskar, Tasnim Mohiuddin, and M Saiful Bari. 2025. [Beyond fertility: Analyzing strr as a metric for multilingual tokenization evaluation.](#) *Preprint*, arXiv:2510.09947.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. [Training language models to follow instructions with human feedback.](#) *Advances in neural information processing systems*, 35:27730–27744.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. 2025. [A survey of multilingual large language models.](#) *Patterns*, 6(1).
- Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and 1 others. 2022. [Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task.](#) In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.
- Holger Schwenk. 2018. [Filtering and mining parallel data in a joint multilingual space.](#) In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 228–234.
- Faisal Tareque Shohan, Mir Tafseer Nayeem, Sam-sul Islam, Abu Ubaida Akash, and Shafiq Joty. 2024. [XL-HeadTags: Leveraging multimodal retrieval augmentation for the multilingual generation of news headlines and tags.](#) In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12991–13024, Bangkok, Thailand. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. [Gemma 3 technical report.](#) *arXiv preprint arXiv:2503.19786*.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. [InternV3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency.](#) *arXiv preprint arXiv:2508.18265*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020.

Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2025. Llava-critic: Learning to evaluate multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13618–13628.

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *arXiv preprint arXiv:2502.14191*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging LLM-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Appendix

A.1 Translation Related Details

We demonstrate the translation prompt below.

Translation Prompt
<p>Translate the following texts (Query, Response_1, Response_2) into the following languages:</p> <p>French, Portuguese, Greek, Chinese, Hindi, Arabic, Czech, German, Hebrew, Indonesian, Italian, Japanese, Korean, Dutch, Persian, Polish, Romanian, Russian, Spanish, Turkish, Ukrainian, Vietnamese, Bengali, and Kazakh.</p> <p>Provide your response in an array of JSON format with the following keys for each JSON object: (i) language, (ii) Translated_Query, (iii) Translated_Response_1, (iv) Translated_Response_2</p>

An example of the output format of Gemini-3-Pro translation is demonstrated below.

JSON Formatted Output

```
[
{
  "Language": "French",
  "Translated_Query": "Quels sont les principaux objets ou sujets de l'image ? Veuillez les décrire en détail.",
  "Translated_Response_1": "Les principaux objets de l'image sont un chien avec une laisse orange attachée à un banc. Le cadre semble être en extérieur, peut-être dans un environnement semblable à un parc, car on voit de l'herbe et un banc en bois.",
  "Translated_Response_2": "Les principaux objets de l'image sont un chien, un banc en bois portant l'inscription « WATSON BOWL » et une table ou un banc de pique-nique. Le chien est assis sur l'herbe à côté du banc et semble attaché ou retenu par une laisse orange fixée au banc."
},
. . . . .
. . . . .
. . . . .
{
  "Language": "German",
  "Translated_Query": "Die Hauptobjekte im Bild sind ein Hund mit einer orangefarbenen Leine, die an einer Bank befestigt ist. Die Umgebung scheint im Freien zu sein, möglicherweise in einer parkähnlichen Umgebung, da man Gras und eine Holzbank sehen kann.",
  "Translated_Response_1": "Die Hauptobjekte im Bild sind ein Hund mit einer orangefarbenen Leine, die an einer Bank befestigt ist. Die Umgebung scheint im Freien zu sein, möglicherweise in einer parkähnlichen Umgebung, da man Gras und eine Holzbank sehen kann."
  "Translated_Response_2": "Die Hauptobjekte im Bild sind ein Hund, eine Holzbank mit der Aufschrift „WATSON BOWL“ sowie ein Tisch oder eine Picknickbank. Der Hund sitzt auf dem Gras neben der Bank und scheint mit einer orangefarbenen Leine, die an der Bank befestigt ist, angebunden oder zurückgehalten zu werden."
},
]
```

A.2 Cross-Lingual Amplification of Positional Bias

We additionally compare positional bias in English versus non-English languages across model families in Table 8 and observe that positional bias consistently increases in non-English settings across all three model families, with InternVL-3.5 and Gemma-3 showing nearly 2x amplification.

Model Family	Avg. English Pos. Bias	Avg. Non-English Pos. Bias
Qwen3-VL	11.03	14.44
InternVL-3.5	15.79	31.18
Gemma-3	5.11	9.95

Table 8: Comparison of average positional bias in English versus non-English languages across three model families.

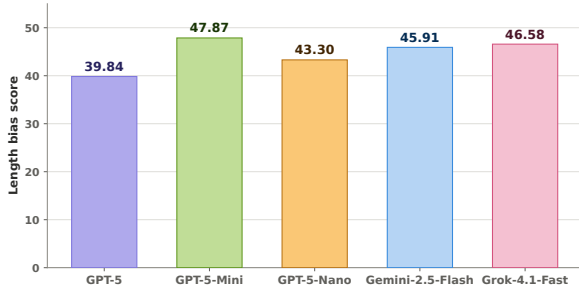


Figure 5: Length Bias in M-VL-RewardBench for closed-source models. Lower values indicate better.

Model	M-OpenCQA	M-VL-RewardBench	Average
Qwen3-VL-32B	65.38	39.25	52.32
Qwen3-VL-8B	64.55	36.36	50.46
Qwen3-VL-4B	63.82	40.15	51.99
Qwen3-VL-30B-A3B	69.63	39.37	54.50
Qwen3-VL-2B	51.75	34.11	42.93
InternVL-3.5-14B	66.16	37.39	51.78
InternVL-3.5-8B	67.31	39.43	53.37
InternVL-3.5-4B	61.58	36.66	49.12
InternVL-3.5-2B	51.82	37.30	44.56
InternVL-3.5-1B	50.34	37.15	43.25
Gemma-3-27B	70.77	37.14	53.96
Gemma-3-12B	69.83	37.61	53.72
Gemma-3-4B	63.72	39.70	51.71
LLaVA-Critic-7B	58.56	35.08	46.82
Pixtral-12B	65.13	42.99	54.06

Table 9: Average Length Bias for open-source LVLMS. Lower values indicate better with Green indicates the best and Red indicates the worst, per column.

A.3 Length Bias Analysis

We show the length bias for the open models in Table 9 and for the closed models in Figure 5. For the open models, we find that the LVLMS tend to show higher length bias in MM-OpenCQA (above 50%) than M-VL-RewardBench (below 50%). Overall, we observe Qwen3-VL-2B demonstrates the lowest length bias. In terms of closed models, we find that the GPT-5 model shows the lowest length bias, while GPT-5-Mini shows the highest.

A.4 Output Format Instruction Following Behavior Analysis.

We evaluate instruction-following fidelity by measuring compliance with the required JSON output format in multilingual settings (Figure 6). We find most closed models exhibit 100% format following adherence, with Claude-4.5-Haiku and GPT-5-Nano achieving slightly below 100% and the

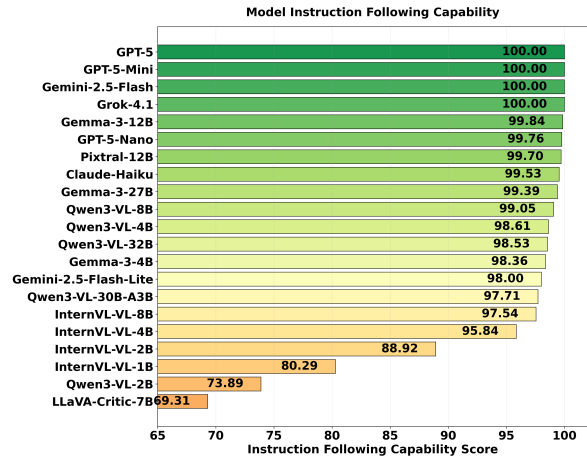


Figure 6: Output Instruction Following Accuracy in M-VL-RewardBench.

Gemini-2.5-Flash-Lite showing modest degradation (98%). Among open models, while none of them could reach 100% accuracy, almost all of them with more than 2B parameters achieve above 95% accuracy (with Gemma-3-12-B achieving the best 99.84% accuracy). Meanwhile, models below 2B parameters show poorer output format instruction following accuracy (below 90% accuracy). Surprisingly, LLaVA-Critic-7B, despite being specifically trained for reward modeling, achieves only 69.3% accuracy. This highlights that specialized training can make certain reward models weaker at following instructions.

A.5 Human Evaluation Instructions in OpenCQA

Below, we show the instructions for our human evaluation in the OpenCQA dataset.

Instructions for Humans to Evaluate GPT-5 Judgments in M-OpenCQA

You are given an image and a question related to the image. You are also provided with two candidate answers, A and B, generated by two models.

Your objective is to decide which answer is better based on the following criteria.

- Correctness with respect to the image and question.
- Completeness and level of detail.
- Relevance and clarity (no unnecessary verbosity).
- Avoiding hallucinations or unsupported claims.

A.6 Human Evaluation of the Translation Quality

We conduct human evaluation on our MM-JudgeBench benchmark (on both subsets) to as-

sess the quality of the translated data. However, given the difficulty in finding high-quality human evaluators having expertise across these 25 languages, we translated both the M-VL-RewardBench and M-OpenCQA datasets from the 24 non-English languages back to English. After back-translation, we check the BERTScore (Zhang et al.) by comparing the original English data with the back-translated English data and observe 90% similarity score.

Afterwards, we sample about 200 examples covering all languages from our benchmark and instruct the human annotator who has expertise in the English language to identify the quality of each example after back translation. We specifically ask the human annotator to evaluate based on the following criteria:

- **Perfect:** The back-translated data is perfect without any translation errors.
- **Good:** The back-translated data does not have any major issues, but has some minor issues in terms of wording/synonyms, etc. However, this should not change the context of the original data.
- **Bad:** The back-translated data has some major issues (e.g., factual errors) that affect the content of the original data.

Based on our human evaluation, we find that the human annotator rates **55.7%** of the data as perfect, **43.8%** as good, and only **0.5%** of the samples as bad. This demonstrates the high quality of Gemini-3-Pro translation.

To justify the reliability of the human evaluation via back translation, we have further conducted a native-speaker evaluation for Bengali across 300 samples. We find that the native speaker rated 73% of the samples as Perfect and 27% as Good, with no samples labeled as Bad.

A.7 Translation Quality of M-OpenCQA Based on LaBSE and CometKiwi

We compute the LaBSE score and the CometKiwi score in the M-OpenCQA dataset and find that it has 0.86 LaBSE and 0.84 CometKiwi scores, which are well above our quality filtering threshold of 0.75. We also observe that the LaBSE and CometKiwi scores are above 0.8 across all languages in our M-OpenCQA dataset, without having any significant differences. Note that we also observe similar trend in M-VL-RewardBench.

Model Family	Hallucination	Mathematical	General
InternVL-3.5	2.95	1.19	1.39
Qwen3-VL	0.34	1.99	0.07
GPT Series	-1.18	-0.24	3.53

Table 10: Cross-lingual performance degradation per task in M-VL-RewardBench

A.8 Example from M-OpenCQA

An example from M-OpenCQA in our MM-JudgeBench benchmark is shown in Figure 7.

A.9 English vs Non-English Performance Gap Across Tasks

To analyze multilingual robustness, we report the average accuracy gap between English and non-English languages, computed across comparable model sizes (2B/4B/8B for Qwen3-VL and InternVL-3.5; Mini/Nano for GPT) in Table 10. Positive values indicate higher performance in English, while negative values indicate better performance in non-English languages.

InternVL-3.5, which shows the largest gap on Hallucination Queries, while Qwen3-VL exhibits its largest gap on Mathematical Reasoning. Qwen3-VL maintains near-parity on General Multimodal tasks (0.07), whereas InternVL and GPT show larger degradation. Interestingly, GPT models slightly improve in non-English settings for Hallucination and Math, suggesting potential benefits from broader multilingual pretraining. Overall, multilingual consistency depends strongly on both task taxonomy and model family.

A.10 Performance Comparison Based on Script Groups and Resources

This section provides a performance comparison based on the seven language script groups⁸ and the resourcedness⁹. We select the best open-source series model (Qwen3) and the best closed-source series model (GPT-5 series) and compare the performance gap between models in M-VL-RewardBench based on the corresponding efficiency tiers: GPT-5 vs Qwen3-32B, GPT-5-Mini vs Qwen3-8B, and GPT-5-Nano vs Qwen3-4B. Below, we first describe these script groups.

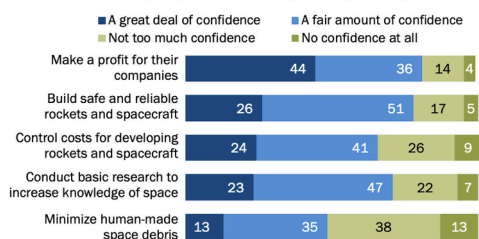
Latin Script Latin script comprises 13 languages: English, French, German, Polish, Por-

⁸Each script group represents a distinct writing system with unique linguistic characteristics that could influence model performance.

⁹High, Medium, and Low.

Many Americans are confident that private space companies will be profitable but skeptical they will keep space clean of debris

% of U.S. adults who say they have _____ that private space companies will ...



Language: English

Query: What Americans think about debris on space?

Response A: According to the chart, many Americans are skeptical about whether private space companies will keep space clean of debris. While a majority have a "great deal" or "fair amount" of confidence that private companies will make a profit, only 13% have that level of confidence in them minimizing human-made space debris. In fact, 38% have "not too much confidence" and 13% have "no confidence at all" in private companies addressing the space debris issue. This suggests that while Americans are optimistic about the profitability of private space ventures, they are more doubtful about their commitment to mitigating the growing problem of space debris.

Response B: 47% of US adults have a fair amount of confidence that private space companies will conduct basic research to increase knowledge of space.

Language: French

Translated Query: Que pensent les Américains des débris spatiaux ?

Translated Response A: Selon le graphique, de nombreux Américains sont sceptiques quant à la capacité des entreprises spatiales privées à garder l'espace exempt de débris. Alors qu'une majorité a « beaucoup » ou « pas mal » de confiance dans le fait que les entreprises privées feront des bénéfices, seuls 13 % ont ce niveau de confiance dans le fait qu'elles minimiseront les débris spatiaux d'origine humaine. En fait, 38 % ont « peu de confiance » et 13 % n'ont « aucune confiance » dans le fait que les entreprises privées s'attaquent au problème des débris spatiaux. Cela suggère que, bien que les Américains soient optimistes quant à la rentabilité des entreprises spatiales privées, ils doutent davantage de leur engagement à atténuer le problème croissant des débris spatiaux.

Translated Response B: 47 % des adultes américains ont une confiance raisonnable dans le fait que les entreprises spatiales privées mèneront des recherches fondamentales pour accroître la connaissance de l'espace.

Example of a Correct Judge Response: {"winner": "A", "reasoning": "Response A directly addresses Americans' views on space debris and accurately summarizes the chart's data showing skepticism about private companies' ability to minimize debris. Response B discusses confidence in basic research, which is not the focus of the question about space debris."}

Example of an Incorrect Judge Response: {"winner": "B", "reasoning": "La plupart des Américains n'ont aucune confiance dans le fait que les entreprises spatiales privées réduiront les débris spatiaux, ce que la réponse B résume avec exactitude."}

Figure 7: Example from MM-JudgeBench (OpenCQA) illustrating multilingual reward evaluation. A chart image-question pair with candidate responses is translated from English into French, where Response A accurately reflects the chart's content. The LVLMM judge selects the correct response in English but incorrect in French, highlighting the need for language-invariant, visually grounded judgment.

tuguese, Romanian, Spanish, Czech, Dutch, Turkish, Vietnamese, Indonesian, and Italian. It uses the Roman alphabet (A-Z) with diacritical marks and is the most widely used writing system globally. Latin script is left-to-right and phonetic, making it straightforward for tokenization.

Cyrillic Script Cyrillic script includes 3 languages: Russian, Ukrainian, and Kazakh. Derived from the Greek alphabet, it is used primarily in Eastern Europe and Central Asia. This group exhibits linguistic diversity, ranging from high-resource (Russian) to low-resource (Kazakh) languages. The script requires separate tokenization strategies from Latin.

Greek Script Greek script represents a single language: Greek. It is an ancient alphabet with unique character sets and diacritical marks. Greek has been extensively studied in linguistics and holds significant cultural importance.

Arabic Script Arabic script includes 2 languages: Arabic and Persian. It is a right-to-left writing system used primarily in the Middle East and parts of Asia. Arabic script includes diacriti-

cal marks and represents some of the most widely spoken languages in the world.

Hebrew Script Hebrew script represents a single language: Hebrew. It is a right-to-left bidirectional writing system with unique properties and historical significance. Hebrew includes diacritical marks (vowel points) and presents specific challenges for tokenization due to its bidirectional nature.

Devanagari Script Devanagari script includes 2 languages: Hindi and Bengali. These languages have significant speaker populations but relatively lower resource availability compared to Latin script languages.

CJK Scripts CJK scripts include 3 languages: Chinese, Japanese, and Korean. Chinese uses logographic characters, Japanese combines logographic kanji with syllabic scripts, and Korean uses the alphabetic hangul script. These languages present unique challenges for tokenization due to large character sets and complex morphology.

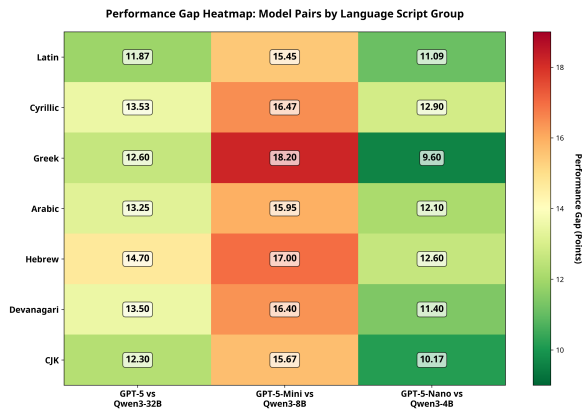


Figure 8: Performance Comparison based on Scripts Groups.

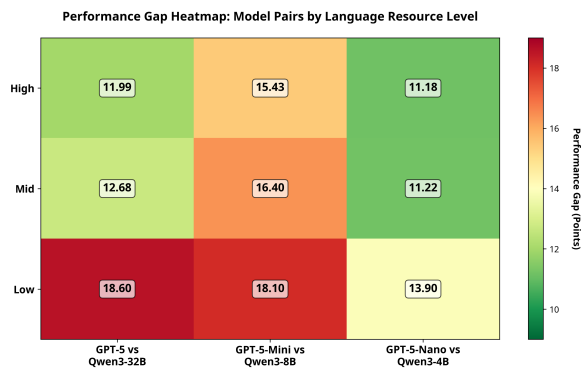


Figure 9: Performance Comparison based on Resource Level.

Figure 8 shows a consistent closed–open gap across all script groups, but the gap is largest for the mid-tier pairing (GPT-5-Mini vs Qwen3-8B) across every script (between 15.45 to 18.20 points), suggesting the relative advantage of the closed model is most pronounced in that tier rather than monotonically increasing with “model size.”

In terms of resource availability, we categorize the languages following Aya-101¹⁰, as demonstrated below:

- **High-Resource (15 languages):** English, Chinese, Japanese, Spanish, French, German, Italian, Portuguese, Russian, Dutch, Arabic, Polish, Hindi, Persian, Czech
- **Mid-Resource (9 languages):** Korean, Turkish, Vietnamese, Indonesian, Greek, Hebrew, Romanian, Ukrainian, Bengali
- **Low-Resource (1 language):** Kazakh

Figure 9 shows that performance gaps between GPT-5 and Qwen3 models grow as language resources decrease, with the largest gaps consistently appearing for low-resource languages. The mid-tier pairing (GPT-5-Mini vs Qwen3-8B) exhibits the strongest gap across all resource levels, while the nano vs Qwen3-4B tier shows the smallest and most stable differences, indicating closer competitiveness at smaller scales.

¹⁰<https://huggingface.co/CohereLabs/aya-101>