PolyChartQA: Benchmarking Large Vision-Language Models with Multilingual Chart Question Answering

Anonymous ACL submission

Abstract

Charts are a universally adopted medium for data communication, yet existing chart understanding benchmarks are overwhelmingly English-centric, limiting their accessibility and relevance to global audiences. To address this limitation, we introduce **PolyChartQA**, the first large-scale multilingual benchmark for chart question answering, comprising 22,606 charts and 26,151 QA pairs across 10 diverse languages. PolyChartQA is constructed through a scalable pipeline that enables efficient multilingual chart generation via data translation and code reuse, supported by LLM-based translation and rigorous quality control. We systematically evaluate multilingual chart understanding with PolyChartQA on state-of-the-art LVLMs and reveal a significant performance gap between English and other languages, particularly low-resource ones. Additionally, we introduce a companion multilingual chart question answering training set, PolyChartQA-Train, on which fine-tuning LVLMs yields substantial gains in multilingual chart understanding across diverse model sizes and architectures. Together, our benchmark provides a foundation for developing globally inclusive vision-language models capable of understanding charts across diverse linguistic contexts.

1 Introduction

Charts are ubiquitous tools for visualizing quantitative data and supporting analytical reasoning across domains such as science, business, and journalism, making accurate chart interpretation essential for data-driven decision-making. Recent advances in large vision-language models (LVLMs) have enabled significant progress in perceiving and reasoning over visualizations such as plots, diagrams, and charts. These models have shown promising results on tasks including complex chart question answering (Masry et al., 2022; Xia et al., 2024; Wang et al., 2024c; Masry et al., 2025a), chart summarization (Rahman et al., 2023; Tang et al., 2023), and chart image re-generation (Moured et al., 2024; Yang et al.).

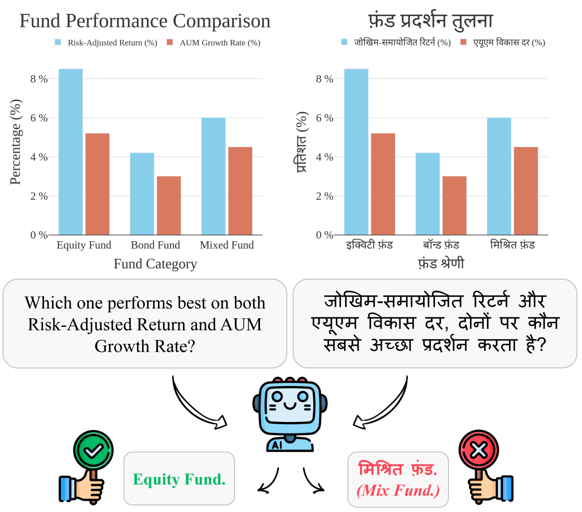


Figure 1: Example of inconsistent chart understanding by LVLMs. The model answers correctly in English but fails on the Hindi equivalent.

However, existing benchmarks for chart understanding remain overwhelmingly English-centric, overlooking the unique challenges of multilingual comprehension. As shown in Figure 1, leading LVLMs often succeed on English chart QA but struggle with their non-English versions. This English-dominant bias poses a major barrier to developing globally inclusive chart understanding models, especially for underrepresented languages. While recent works (Chen et al., 2024a; Heakl et al., 2025) have introduced bilingual chart datasets, they remain limited in scale and language coverage. To date, **no** comprehensive benchmark exists for evaluating multilingual chart understanding in LVLMs. Moreover, most multilingual multimodal benchmarks (Pfeiffer et al., 2022; Liu et al., 2021; Yu et al., 2025; Liu et al., 2024c; Xuan et al., 2025) focus on natural images rather than structured data

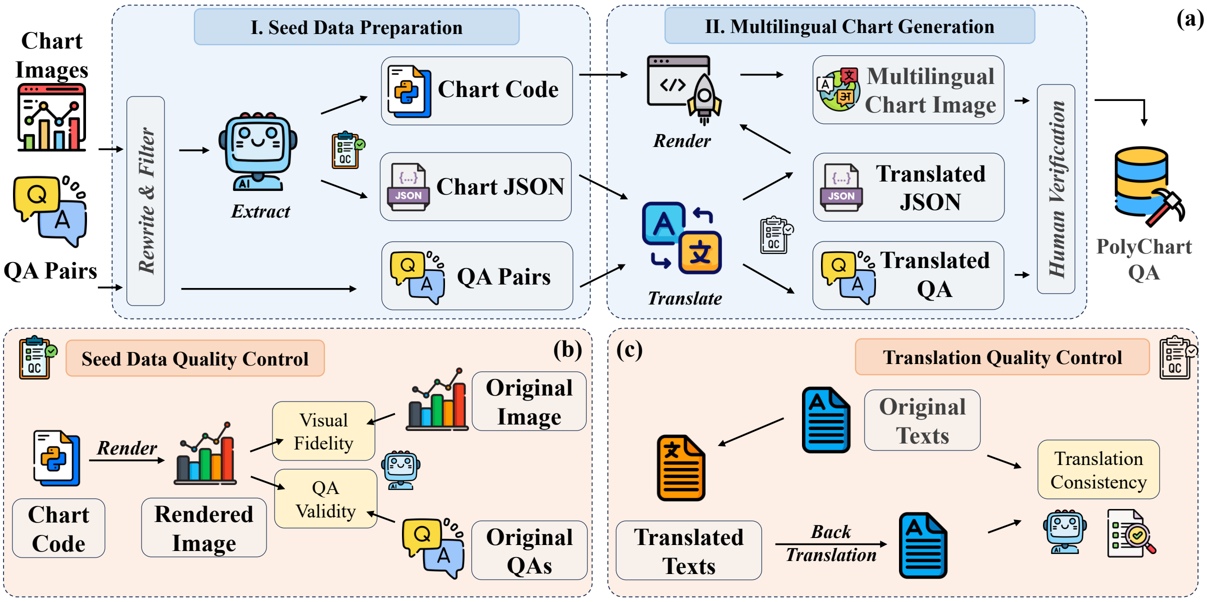


Figure 2: Overview of the PolyChartQA data pipeline. (a) The full workflow consists of two stages: **Seed Data Preparation** and **Multilingual Chart Generation**. (b) Quality control procedures applied with seed data generation. (c) Quality control procedures applied during the translation stage.

like charts, leaving multilingual chart understanding largely unexplored. A key reason for this gap is the high cost of multilingual chart annotation (Romero et al., 2024; Tang et al., 2024), which severely restricts the scalability of such benchmarks.

To overcome these challenges, we develop PolyChartQA through a scalable two-stage pipeline. In the first stage, we generate high-quality English seed data by decomposing charts into structured JSON specifications and reusable code templates. In the second stage, we employ state-of-the-art LLMs to translate chart data and QA pairs and automatically render multilingual charts. A dedicated multi-stage quality-control procedure, combining automated consistency checks with final human verification, ensures the accuracy and naturalness of the multilingual data. Using this pipeline, we construct **PolyChartQA**, the first large-scale benchmark for multilingual chart understanding, spanning 10 widely spoken languages, including English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, and Japanese, which together account for over 65% of the global population (Maaz et al., 2024). The benchmark comprises a test set of over 22K chart images with 26K QA pairs and a training set of 751K QA pairs across 131K charts, providing a diverse and rigorously curated resource for evaluating and advancing multilingual chart understanding.

Using PolyChartQA, we present the first systematic evaluation of multilingual chart question answering in LVLMs, revealing that (i) current models remain markedly weak on multilingual chart QA, especially for low-resource languages, and (ii) cross-lingual generalization is fragile, with large performance gaps across scripts and sensitivity to partial visual–textual alignment. To bridge this gap and enhance multilingual chart capabilities, we show that fine-tuning on PolyChartQA-Train across different model families yields substantial performance gains, highlighting the effectiveness of instruction tuning for multilingual chart reasoning. We further provide a detailed error analysis across languages, scripts, and question types to expose persistent failure modes. In summary, our main contributions are:

- **Unified multilingual chart construction pipeline.** We propose a reproducible automatic pipeline for constructing high-quality, large-scale multilingual chart QA datasets.
- **PolyChartQA benchmark.** We introduce PolyChartQA, the first benchmark enabling systematic evaluation of LVLMs on chart understanding in ten diverse languages.
- **Comprehensive empirical analysis.** We conduct extensive experiments and error analysis that reveal critical performance gaps and demonstrate how our datasets substantially narrow them.

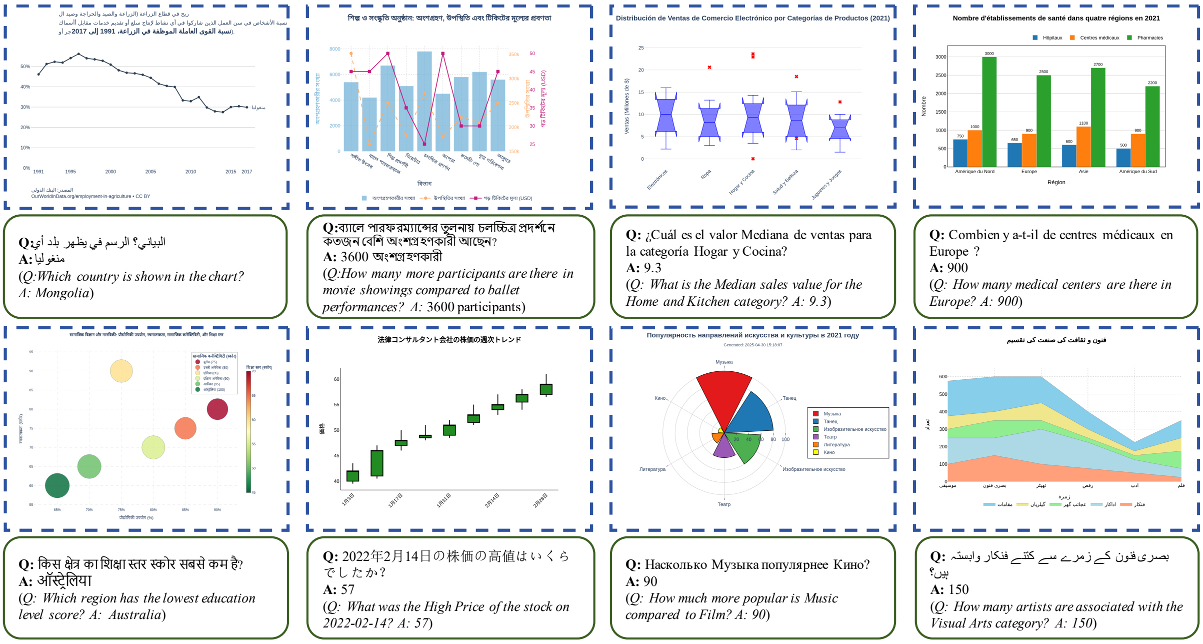


Figure 3: Multilingual chart question answering visualizations selected from PolyChartQA. First row, from left to right: Arabic, Bengali, Spanish, French. Second row, from left to right: Hindi, Japanese, Russian, Urdu.

2 Related Work

2.1 Chart Understanding Datasets

Chart understanding requires models to jointly reason over visual and textual cues under diverse instructions. Recent benchmarks evaluate LVLMs on chart question answering (Masry et al., 2022; Methani et al., 2020; Kantharaj et al., 2022a), summarization (Tang et al., 2023; Kantharaj et al., 2022b; Rahman et al., 2023), chart-to-table conversion (Xia et al., 2023, 2024; Chen et al., 2024a), and re-rendering (Moured et al., 2024; Yang et al.), with QA serving as the primary measure of fine-grained comprehension. Early datasets (Kahou et al., 2017; Kafle et al., 2018; Methani et al., 2020) mainly used synthetic charts and template-based questions, limiting diversity and realism. Later benchmarks (Masry et al., 2022; Xia et al., 2024; Liu et al., 2024a) moved toward realistic charts and human-authored questions, improving chart coverage and question complexity. However, most benchmarks remain English-only (Chen et al., 2024a; Heakl et al., 2025), limiting comprehensive evaluation and real-world deployment of LVLMs.

2.2 Multilingual LVLMs

Building on foundational monolingual models (Li et al., 2023; Team et al., 2024a,b), numerous multilingual LVLMs have emerged. Early influential works (Chen et al., 2022; Geigle et al., 2024; Beyer et al., 2024; Steiner et al., 2024) pioneered scalable multilingual vision-language alignment. More recent open-source efforts such as PALO (Maaz et al., 2024), Maya (Alam et al., 2024), Pangea (Yue et al., 2024), and Centurio (Geigle et al., 2025), together with model families including QwenVL (Bai et al., 2023, 2025; Wang et al., 2024b), InternVL (Chen et al., 2024c,d,e), and Phi-Vision (Abdin et al., 2024a,b), further broaden language coverage and improve multilingual multimodal performance. However, their ability to handle complex, text-rich visuals such as multilingual charts remains underexplored.

2.3 Multilingual Evaluations on LVLMs

The rapid progress of multilingual LVLMs has led to numerous benchmarks evaluating their multimodal capabilities, including general cross-lingual VQA (Pfeiffer et al., 2022; Changpinyo et al., 2022), text-centric VQA (Tang et al., 2024; Yu et al., 2025), and culturally grounded VQA (Romero et al., 2024; Liu et al., 2021; Vayani et al., 2025). Comprehensive suites such as MM-Bench (Liu et al., 2024c), MMLU-Prox (Xuan et al., 2025), and M4U (Wang et al., 2024a) further assess reasoning, dialogue, captioning, and math problem solving, while M3Exam (Zhang et al., 2023) and Exams-V (Das et al., 2024) provide large-scale multilingual evaluations. However, chart-based understanding remains largely underexplored, with limited coverage in existing benchmarks (Zhang et al., 2023; Geigle et al., 2025).

3 PolyChartQA

We present **PolyChartQA**, a large-scale multilingual chart question answering benchmark that addresses the scarcity of multilingual resources for chart understanding. As summarized in Table 1, PolyChartQA spans 10 languages (English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, and Japanese) and covers 16 diverse chart types. The dataset is built through a unified pipeline (Figure 2): we first construct high-quality English seed data comprising chart images, rendering code, structured JSON, and QA pairs, and then expand it to other languages via an LLM-assisted translation pipeline. The decoupled code-and-JSON representation further supports easy extension to related chart tasks (e.g., summarization and chart generation) without additional manual annotation. To ensure accuracy and reliability, we apply multi-stage quality control that combines automated validation with targeted human review. The remainder of this section details seed data preparation (§3.1), multilingual chart generation (§3.2), and quality control (§3.3); additional pipeline details and prompts are provided in Appendix A and Appendix F, respectively.

3.1 Seed Data Preparation

We ground multilingual generation in high-quality English chart QA data by selecting three widely used benchmarks—ChartQA (Masry et al., 2022), ChartLlama (Han et al., 2023), and ChartX (Xia et al., 2024)—for their chart diversity, question coverage, and data quality. We construct PolyChartQA-Test from the test splits of ChartQA and ChartX, and PolyChartQA-Train from the training splits of ChartQA and ChartLlama; detailed statistics are summarized in Table 6.

To ensure the quality of the seed data, we apply a two-step cleaning and validation procedure. **(i) Answer verification.** We use *Gemini-2.5-Pro* to automatically check each chart question–answer pair; if the model's prediction disagrees with the ground truth but suggests a clear correction, we manually revise the answer, otherwise we discard the sample. **(ii) Answer standardization.** We normalize verbose answers into concise canonical forms while preserving their semantics (e.g., "the highest bar value in the chart is 42.1" → "42.1"). A manual review of 10% of the cleaned data yields a pass rate above 98%, confirming the reliability of the seed datasets.

Dataset	#Lang.	Chart Types	#Charts	#QAs
ChartQA (2022)	1	3	1,612	2,500
ChartX (2024)	1	18	1,152	2,304
ChartY (2024a)	2	4	6,000	6,000
KITAB-Bench (2025)	1	16	576	576
SMPQA (2025)	11	2	1,100	4,300
ChartMind (2025)	2	7	757	757
PolyChartQA	**10**	**16**	**22,606**	**26,151**

Table 1: Comparison of different chart-related datasets and benchmarks.

Subsequently, we adopt a decoupled chart representation that separates content from visual rendering (Shinoda et al., 2024), enabling flexible multilingual generation: the same rendering code can be reused with translated JSON to produce chart images in different languages. For each cleaned chart instance, we prompt *Gemini-2.5-Pro* to generate two complementary artifacts: (i) a structured **JSON file** encoding the underlying data table, chart type, colors, and layout attributes, and (ii) an executable **Python script** that reproduces the chart using *Plotly* [1], which natively supports multilingual text rendering.

3.2 Multilingual Chart Generation

To construct multilingual chart QA datasets, we translate the English seed data into multiple target languages via a two-stage process. We first obtain multilingual textual annotations (JSONs and QA pairs), and then render the corresponding chart images in each target language by reusing the template code.

Text Translation. Standard machine translation systems often struggle to preserve the structure and fine-grained semantics of chart-oriented JSON files and their associated QA pairs. In contrast, recent work (Qiu et al., 2022; Chen et al., 2024b; Maaz et al., 2024) has shown that LLM-based translation achieves higher fidelity and consistency. Building on this, we adopt an LLM-based workflow with *Gemini-2.5-Pro*, which jointly translates each chart's JSON data and QA pairs to ensure semantic coherence. The model is instructed to preserve meaning while adapting to cultural and linguistic conventions to reduce translation bias. Our analyses in §3.3 indicate that the resulting multilingual corpora largely preserve the semantic content and structural properties of the original English data.

[1] https://github.com/plotly/plotly.py

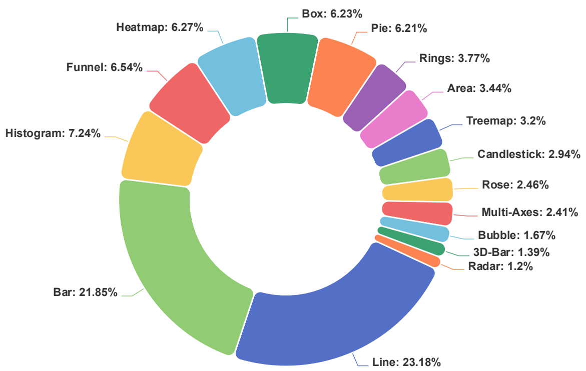


Figure 4: Distribution of chart types in POLYCHARTQA.

Chart Image Translation. Given the translated JSONs and QA pairs, we generate multilingual chart images by pairing each translated JSON with its corresponding template code and rendering the chart in the target language.

3.3 Quality Control

Our pipeline incorporates a multi-stage quality control mechanism to ensure both the accuracy and usability of the constructed dataset across all languages.

Seed Data Quality Control. To ensure the integrity of the English seed dataset, we applied a multi-stage validation process, as shown in Figure 2 (b). With both JSON files and rendering code acquired, we first executed the code to verify reproducibility and automatically removed any samples that failed to render successfully. We then examined two key aspects of data quality. **(i) Visual Fidelity**: Each regenerated chart was compared against its original version using *Gemini-2.5-Pro* to detect visual or semantic discrepancies. Charts showing notable mismatches in chart type, data values, or layout were discarded. **(ii) QA Validity**: We further verified that all questions remained answerable from the reconstructed charts, using *Gemini-2.5-Pro* and *GPT-4.1* as independent validators. Both models possess strong vision–language reasoning and code understanding capabilities, and requiring agreement between them provides a stricter and more reliable validation process. Only samples confirmed as valid by both models were retained, removing those with semantic inconsistency or linguistic errors.

Multilingual Data Quality Control. Building upon the validated seed data, we further applied a two-stage quality control procedure to ensure the reliability of the multilingual outputs. Similar to the seed stage, any samples whose chart code failed to execute during the multilingual image generation stage were automatically discarded. For the remaining data, we evaluated both text translation quality and multilingual chart image quality. **(i) Translation Quality**: As illustrated in Figure 2(c), each translated instance was back-translated into English and compared with the original. We assessed textual consistency using the METEOR (Banerjee and Lavie, 2005) metric, complemented by semantic judgements from *Gemini-2.5-Pro* to compensate for METEOR's limited sensitivity to nuanced meaning differences. Samples with back-translated content that deviated substantially from the original English semantics were filtered out. **(ii) Visual Inspection**: All remaining multilingual chart images were then manually reviewed to identify and remove those containing visual defects such as text clipping, misaligned layouts, or rendering artifacts.

Metrics	Image Quality	QA Relevance	Translation Accuracy
Avg. Score	2.87	2.93	2.89
Avg. Disag.	3.1	3.4	4.1
Avg. $\bar{\kappa}_w$	0.887	0.817	0.885

Table 2: Average human scores and inter-annotator agreement scores for each evaluation dimension. "Disag." shows the raw count of differing ratings and κ_w denotes weighted Cohen's κ.

3.4 Data Statistics

POLYCHARTQA consists of 154,121 chart images and 777,514 question answer pairs across 10 languages, split into a test set (POLYCHARTQA-Test) with 22,606 charts and 26,151 QA pairs and a training set (POLYCHARTQA-Train) with 131,515 charts and 751,363 QA pairs. It spans 16 diverse chart types (Figure 4), with representative examples shown in Figure 3. More detailed statistics of POLYCHARTQA are provided in Appendix B.

To assess the quality of POLYCHARTQA-Test, we conduct a **human evaluation** on a randomly sampled 20% subset for each language. Bilingual annotators rate each instance along three dimensions: **(i) Translation Quality**, assessing semantic accuracy, fluency, and naturalness while avoiding bias or misinformation; **(ii) Chart Image Quality**, evaluating visual clarity, text legibility, and overall presentation; and **(iii) QA Correctness**, verifying question relevance and factual consistency with

Model	#Params	EN	ZH	FR	ES	RU	JA	AR	UR	HI	BN	Avg. (w/ EN)	Avg. (w/o EN)
Proprietary Models													
GPT-4o	-	55.9	46.0	53.4	54.4	52.4	45.4	50.5	48.7	51.3	48.2	50.9	50.2
Gemini-2.5-Pro	-	**70.6**	**67.7**	**69.0**	**69.3**	**67.6**	**68.6**	**69.1**	**67.5**	**68.6**	**66.0**	**68.5**	**68.2**
Open Source Models													
InternVL-2.5 (Chen et al., 2024c)	2B	27.8	3.3	14.7	9.2	9.5	2.0	4.3	0.3	1.2	0.1	7.8	5.1
InternVL-3 (Zhu et al., 2025)	2B	43.7	35.3	30.8	33.5	25.6	26.9	17.1	14.6	15.7	11.9	25.6	23.1
Qwen2-VL (Wang et al., 2024b)	2B	42.3	33.6	37.6	37.7	35.9	22.2	28.8	19.1	24.4	23.0	30.7	29.1
Qwen2.5-VL (Bai et al., 2025)	3B	**67.4**	**59.6**	**61.8**	**62.5**	**58.0**	48.8	51.4	37.2	45.7	43.0	53.7	51.8
PaliGemma2 (Steiner et al., 2024)	3B	26.6	14.7	19.7	21.5	13.9	10.7	15.7	15.9	12.2	14.3	10.2	16.3	14.9
Phi-3.5-Vision (Abdin et al., 2024a)	4.2B	45.1	17.5	37.2	36.9	26.9	15.7	9.3	4.7	10.6	10.6	23.2	20.2
DeepSeek-VL2 (Wu et al., 2024)	4.5B	40.1	38.8	26.4	34.1	19.9	0.0	14.2	13.8	19.1	16.3	24.8	22.5
Phi-4 Vision (Abdin et al., 2024b)	5.6B	62.3	46.0	55.9	44.6	48.7	41.6	29.7	23.4	33.4	18.3	40.6	37.7
LLaVA-OneVision (Li et al., 2024)	7B	18.7	10.1	13.1	14.2	9.4	8.3	7.5	5.2	7.1	5.7	10.1	9.0
LLaVA-v1.6 (Liu et al., 2024b)	7B	24.8	12.9	18.9	18.2	13.5	11.5	12.0	7.7	10.0	6.7	13.9	12.4
Qwen2-VL (Wang et al., 2024b)	7B	56.4	54.3	53.4	52.7	52.2	47.3	40.5	32.0	43.9	40.3	47.3	46.1
Qwen2.5-VL (Bai et al., 2025)	7B	60.5	58.3	57.2	59.0	56.8	**55.6**	**52.0**	**43.7**	**49.4**	**46.4**	**53.8**	**53.0**
InternVL-2.5 (Chen et al., 2024c)	8B	39.2	26.3	32.4	33.5	29.5	22.6	10.9	11.2	14.0	13.4	23.5	21.4
InternVL-3 (Zhu et al., 2025)	8B	54.1	39.4	43.4	45.8	38.1	39.7	21.4	17.2	20.2	17.5	33.8	31.0
Llama-3.2-Vision (Grattafiori et al., 2024)	11B	15.5	16.9	14.1	12.9	15.4	9.6	13.1	14.4	21.3	17.5	15.2	15.2
Chart Specific Models													
TinyChart (Zhang et al., 2024)	3B	**45.6**	15.1	23.5	26.7	12.3	11.1	10.7	9.3	10.6	7.9	**17.9**	14.2
ChartGemma (Masry et al., 2025b)	3B	14.4	7.2	17.2	**30.2**	15.2	9.0	9.5	6.0	**13.5**	6.2	11.1	10.6
ChartInstruct (Masry et al., 2024)	7B	23.8	15.2	21.2	21.7	16.6	12.6	6.7	0.1	3.9	0.0	12.3	10.7
ChartLlama (Han et al., 2023)	13B	11.7	7.9	**26.7**	21.9	**21.4**	12.0	11.8	**15.6**	10.6	**13.1**	15.6	14.1
ChartAssistant (Meng et al., 2024)	13B	25.8	**15.8**	25.1	24.4	18.5	**14.2**	**11.9**	11.7	11.5	9.3	17.1	**15.9**
Multilingual Models													
Centurio (Geigle et al., 2025)	-	7.9	4.0	3.6	3.0	1.5	2.5	2.0	1.5	1.5	1.0	2.9	2.2
Pangea (Maaz et al., 2024)	7B	**24.7**	**13.6**	**19.8**	**21.3**	**15.8**	**11.5**	**13.1**	**12.1**	**13.1**	**13.1**	**16.1**	**14.9**
PALO (Maaz et al., 2024)	7B	11.5	6.0	10.5	9.9	7.0	5.9	7.0	5.0	5.2	3.6	7.3	6.7
Maya (Alam et al., 2024)	8B	8.7	6.4	7.6	7.2	6.8	6.0	7.1	5.7	6.9	5.6	6.8	6.6

Table 3: Overall performance on POLYCHARTQA. Bold values in each model category denote the best performance and underlined values denote the second best.

the chart. Each instance was annotated by one annotator and independently reviewed by another to ensure reliability. As summarized in Table 2, all three dimensions achieve near-ceiling performance, with average scores above 2.8 (out of 3) and strong inter-annotator agreement ($\bar{\kappa}_w > 0.8$), confirming the overall reliability of POLYCHARTQA-Test. Additional details are provided in Appendix C.

4 Experiments

4.1 Experimental Setup

To thoroughly assess the multilingual perception and reasoning abilities of modern LVLMs on our multilingual chart benchmark, we select 22 representative state-of-the-art models from four categories: open-source general MLLMs, open-source multilingual LVLMs, chart-specific LVLMs, and closed-source LVLMs.

All baseline models are evaluated under their official configurations. During inference, we set the decoding temperature to 0.01 and top_p to 0.7. We use a unified multilingual prompt: "Answer the question using a word or phrase in <target_language> or a number in digits. <Question>" All results are averaged over 8 independent runs. Experiments are conducted on 8 NVIDIA A100 GPUs.

4.2 Evaluation Results

Metrics. Following prior work (Masry et al., 2022), we adopt a type-aware relaxed accuracy metric: numerical predictions are considered correct if within 5% relative error of the ground truth; non-numerical answers require exact string match.

Zero-shot Evaluation. Table 3 reports the zero-shot performance of various models on POLYCHARTQA. A substantial gap is observed between closed-source and open-source models: *Gemini-2.5-Pro* achieves the best overall performance across all languages (Avg. 68.5), while *GPT-4o* is notably lower (Avg. 50.9).

Among open-source models, *Qwen2.5-VL* is the strongest, performing well across both high- and low-resource languages and even surpassing *GPT-4o* on average. By comparison, *InternVL-3* and *DeepSeek-VL2* show larger drops on non-English inputs, indicating limited robustness for multilingual chart understanding.

Chart-specific models also struggle in multilingual settings, as prior chart-focused models that perform well in English fail to generalize effectively to other languages. Multilingual LVLMs such as *Pangea*, *PALO*, *Maya*, and *Centurio* exhibit weak overall accuracy on POLYCHARTQA, suggesting that broad multilingual pretraining alone is insufficient for text-rich chart reasoning and grounding.

Across model families, accuracy is relatively stable for high-resource languages such as English, Chinese, and French, but degrades sharply for low-resource languages, particularly Urdu and Hindi, consistent with prior findings (Maaz et al., 2024). This trend indicates that current multilingual training pipelines provide insufficient chart-specific grounding in low-resource settings, likely due to data scarcity and imbalanced language representation.

Cross-lingual Performance Varies by Model Families. We evaluate four representative model families, Qwen2.5-VL, InternVL3, PaliGemma2, and LLaVA-v1.6, under cross-lingual input settings where either the chart image or the QA pair is replaced with its English counterpart, as shown in Table 4. We observe clear family-level differences as the linguistic alignment between modalities varies. Qwen2.5-VL achieves its best performance under fully aligned multilingual inputs, while introducing English into either modality slightly degrades accuracy, consistent with its strong zero-shot performance on non-English data and reliance on language-consistent visual–text alignment. In contrast, InternVL3, PaliGemma2, and LLaVA-v1.6 show improved accuracy when English is introduced, reflecting a heavier dependence on English as a pivot language to compensate for weaker non-English grounding. These results indicate that robust multilingual chart understanding requires exposure to diverse cross-lingual alignment patterns beyond English-centric supervision.

Model Size	Multi. Img.	Multi. QA	Avg. (w/ EN)	Avg. (w/o EN)
Qwen2.5-VL-3B	✗	✓	49.6	47.3
	✓	✗	52.1	49.9
	✓	✓	**53.7**	**51.8**
Qwen2.5-VL-7B	✗	✓	48.3	46.6
	✓	✗	51.0	49.5
	✓	✓	**53.8**	**53.0**
InternVL3-2B	✗	✓	**27.9**	**25.8**
	✓	✗	27.6	25.2
	✓	✓	25.6	23.1
InternVL3-8B	✗	✓	**42.0**	**40.2**
	✓	✗	37.6	34.8
	✓	✓	33.8	31.0
PaliGemma2-3B	✗	✓	**29.0**	**28.4**
	✓	✗	18.6	17.1
	✓	✓	16.3	14.9
LLaVA-v1.6-7B	✗	✓	**18.0**	**16.3**
	✓	✗	17.3	16.2
	✓	✓	13.9	12.4

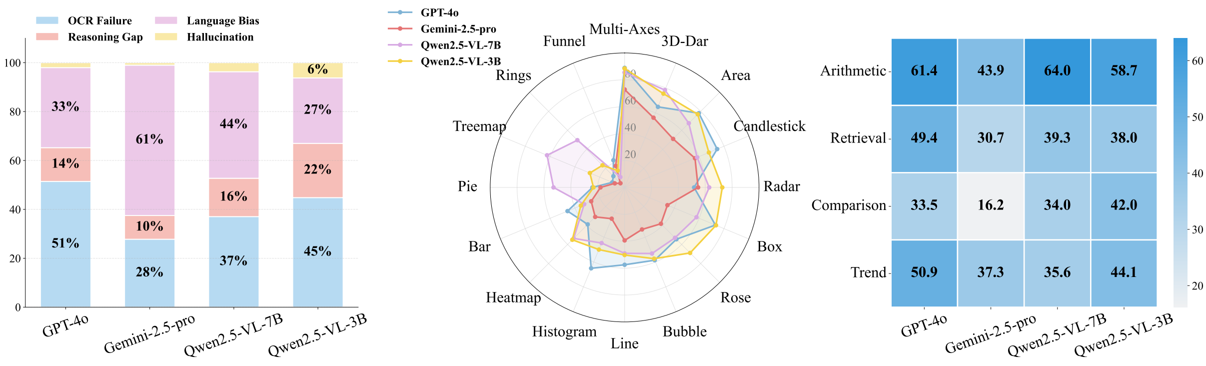
Table 4: Cross-lingual performance of different LVLMs. *Multi. Img.* and *Multi. QA* indicate whether the chart image or QA pair is multilingual. Bold numbers denote the best results for each model.

Model	Avg. (w/ EN)	Avg. (w/o EN)
Qwen2.5-VL-3B	53.7	51.8
+ *fine-tuning*	**61.1** (+13.8%)	**60.2** (+16.2%)
Qwen2.5-VL-7B	53.8	53.0
+ *fine-tuning*	**66.9** (+24.3%)	**66.1** (+24.7%)
InternVL-3-2B	25.6	23.1
+ *fine-tuning*	**33.3** (+30.1%)	**31.2** (+35.1%)
InternVL-3-8B	33.8	31.0
+ *fine-tuning*	**44.0** (+30.2%)	**41.4** (+33.5%)
PaliGemma2-3B	16.3	14.9
+ *fine-tuning*	**29.0** (+77.9%)	**28.4** (+90.6%)
LLaVA-v1.6-7B	13.9	12.4
+ *fine-tuning*	**25.5** (+83.5%)	**24.0** (+93.5%)

Table 5: Fine-tuning results using PolyChartQA-Train across different model families and sizes. Performance gains are highlighted in green.

Fine-tuning Significantly Boosts Multilingual Chart Understanding. Multilingual chart comprehension poses a significant challenge for LVLMs. To address this limitation, we investigate a straightforward yet highly effective strategy: fine-tuning these models on dedicated multilingual chart instruction data using PolyChartQA-test. For a comprehensive evaluation, we selected 6 representative LVLMs spanning various architectures and sizes: Qwen2.5-VL-3B, Qwen2.5-VL-7B, InternVL3-2B, InternVL3- 8B, PaliGemma2-3B, and LLaVA-v1.6-Mistral-7B. We applied LoRA (Hu et al., 2022) training with a rank of $r = 128$ and a learning rate of $1e^{-5}$; the vision encoder was kept frozen, and all models were trained for a single epoch.

As summarized in Table 5, fine-tuning on PolyChartQA-Train yields **substantial performance improvements across all models**. The average accuracy increases by approximately 20% for Qwen2.5-VL, 30% for InternVL3, and over 70% for PaliGemma2 and LLaVA-v1.6. Notably, Qwen2.5-VL-7B surpasses *GPT-4o* and reaches performance comparable to *Gemini-2.5-Pro* after fine-tuning. These results highlight **the strong generalizability and effectiveness of PolyChartQA** in enhancing multilingual chart understanding across diverse LVLM architectures.



(a) Error Composition Analysis (b) Error Rate by Chart Type (c) Error Rate by Question Type

Figure 5: Error analysis across error types, chart types, and question types.

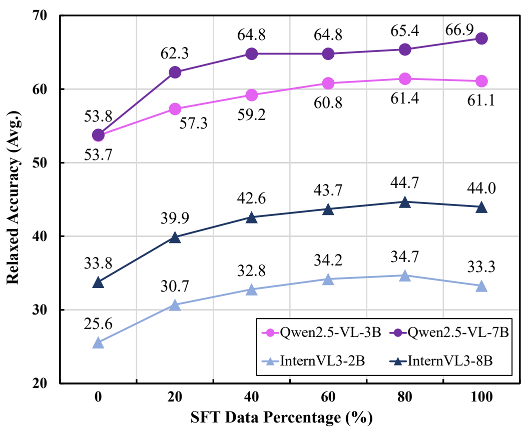


Figure 6: Performance on POLYCHARTQA with respect to the SFT data size across different model families.

Performance Scales with Training Data Size. We assess the impact of data scale by fine-tuning each model on 20%–100% of the POLYCHARTQA-Train. As shown in Figure 6, **performance scales positively with data size** across all model families and capacities. The most substantial gains occur within the initial 20% of data, indicating that early exposure provides the greatest learning benefit (Shaham et al., 2024). Smaller models such as InternVL3-2B and Qwen2.5-VL 3B tend to reach performance saturation earlier, at around 80%. Whereas stronger models such as Qwen2.5-VL-7B continue to benefit from additional data, demonstrating greater scalability and data utilization efficiency. These results suggest that larger models better capture diverse multilingual chart patterns, whereas smaller ones may benefit from more targeted or curriculum-based training.

4.3 Error Analysis

To further investigate model limitations, we selected four representative models and conducted a multi-dimensional error analysis. We first evaluated error rates across different chart and question types. To further diagnose the root causes of these errors, we sampled 300 failure cases per model for each language and categorized them manually. Overall and language-wise breakdowns exhibit similar trends across languages.

As shown in Figure 5(a), analysis reveals that OCR failures (27.8%–51.4%) and Language Bias (26.8%–61.5%) are the dominant error sources, together accounting for the vast majority of incorrect predictions. Reasoning gaps constitute a moderate portion (9.7%–22.2%), while Hallucination remains a minor issue (<7%). Figure 5(b) shows that error rates increase with chart complexity, with multi-axes, 3D-bar, and candlestick charts exhibiting substantially higher failure rates than simpler formats. Figure 5(c) further reveals clear variation across question types: arithmetic questions are the hardest, while comparison and retrieval are easier.

5 Conclusion

In this paper, we introduce **POLYCHARTQA**, the first large-scale multilingual benchmark for chart question answering, covering 10 diverse languages. Built through a scalable and reproducible pipeline, POLYCHARTQA enables efficient multilingual chart generation and evaluation. Experiments reveal that existing LVLMs struggle with multilingual chart understanding, particularly in non-Latin languages. Applying fine-tuning on POLYCHARTQA-Train leads to substantial and consistent improvements across all model architectures, demonstrating the effectiveness and strong generalizability of our dataset. We hope this work inspires broader research into multilingual multimodal understanding and foster the development of more inclusive, globally accessible LVLMs.

Limitations

Despite introducing the first large-scale multilingual benchmark for chart question answering, POLYCHARTQA still has several limitations. While it includes a diverse set of major languages, it excludes many lesser-spoken or low-resource ones, limiting its global inclusivity. Secondly, since POLYCHARTQA builds on existing datasets, it may inherit framing biases or inaccuracies from the source datasets. Additionally, although we employ a multi-stage validation process with human review, the use of LLM-based generation and translation may still introduce subtle shifts in tone, cultural framing, or emphasis across languages. Future work may explore fully human-annotated datasets when feasible, extend POLYCHARTQA to additional chart understanding tasks beyond QA, and expand to more complex real-world visual formats such as infographics or interactive dashboards.

Ethics Statements

Our work aims to promote language inclusivity and accessibility in AI technologies by constructing a multilingual benchmark focused on chart understanding. By systematically evaluating model performance across diverse languages and scripts, especially those underrepresented in existing resources, we highlight current limitations and foster the development of more equitable large vision-language models. We believe this contributes to reducing the dominance of English in AI systems and supports the global community in accessing AI tools in their native languages. We acknowledge that our dataset, being derived from existing sources, may inherit biases or misinformation from the original charts. Furthermore, our use of LLMs for translation, despite a multi-stage validation process, may introduce subtle artifacts such as tonal shifts or cultural inaccuracies. We encourage future work to further improve multilingual data fidelity and broaden the linguistic inclusivity of AI systems.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024a. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024b. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, and 1 others. 2024. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Onechart: Purify the chart structural extraction via one auxiliary token. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 147–155.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024b. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024d. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024e. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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*.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2024. mblip: Efficient bootstrapping of multilingual vision-llms. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 7–25.

Gregor Geigle, Florian Schneider, Carolin Holtermann, Chris Biemann, Radu Timofte, Anne Lauscher, and Goran Glavaš. 2025. Centurio: On drivers of multilingual ability of large vision-language model. *arXiv preprint arXiv:2501.05122*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Ahmed Heakl, Abdullah Sohail, Mukul Ranjan, Rania Hossam, Ghazi Ahmed, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Khan, and Salman Khan. 2025. Kitab-bench: A comprehensive multi-domain benchmark for arabic ocr and document understanding. *arXiv preprint arXiv:2502.14949*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

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

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022a. Opencqa: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022b. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1287–1310.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. 2024. Palo: A polyglot large multimodal model for 5b people. *arXiv preprint arXiv:2402.14818*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025a. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension and reasoning. *arXiv preprint arXiv:2403.09028*.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2025b. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*.

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

Omar Moured, Sara Alzalabny, Anas Osman, Thorsten Schwarz, Karin Müller, and Rainer Stiefelhagen. 2024. Chartformer: A large vision language model for converting chart images into tactile accessible svgs. In *International Conference on Computers Helping People with Special Needs*, pages 299–305. Springer.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

Chen Qiu, Dan Oneață, Emanuele Bugliarello, Stella Frank, and Desmond Elliott. 2022. Multilingual multimodal learning with machine translated text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4178–4193.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620*.

David Romero, Chenyang Lyu, Haryo Wibowo, Santiago Góngora, Aishik Mandal, Sukannya Purkayastha, Jesus-German Ortiz-Barajas, Emilio Cueva, Jinheon Baek, Soyeong Jeong, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Advances in Neural Information Processing Systems*, 37:11479–11505.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

Risa Shinoda, Kuniaki Saito, Shohei Tanaka, Tosho Hirasawa, and Yoshitaka Ushiku. 2024. Sbs figures: Pre-training figure qa from stage-by-stage synthesized images. *arXiv preprint arXiv:2412.17606*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, and 1 others. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, and 1 others. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024a. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024c. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Jingxuan Wei, Nan Xu, Junnan Zhu, Gaowei Wu, Qi Chen, Bihui Yu, Lei Wang, and 1 others. 2025. Chartmind: A comprehensive benchmark for complex real-world multimodal chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4555–4569.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, and Junchi Yan. 2023. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, and 1 others. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, and 1 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran XU, Xinyu Zhu, Siheng Li, Yuxiang Zhang, and 1 others. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. In *The Thirteenth International Conference on Learning Representations*.

Xinmiao Yu, Xiaocheng Feng, Yun Li, Minghui Liao, Ya-Qi Yu, Xiachong Feng, Weihong Zhong, Ruihan Chen, Mengkang Hu, Jihao Wu, and 1 others. 2025. Cross-lingual text-rich visual comprehension: An information theory perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9680–9688.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A fully open multilingual multimodal llm for 39 languages. In *The Thirteenth International Conference on Learning Representations*.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Data Construction Pipeline Details

This section provides extended technical details on our data construction pipeline, clarifying design choices, dataset selection, and quality assurance processes. It also addresses common concerns regarding technical contributions, source datasets selection, and filtering statistics.

A.1 Source Dataset Selection

To validate our choice of source datasets, Table 6 compares existing English chart QA datasets in terms of realism, diversity, and scale. We selected ChartQA and ChartX because they together provide an optimal combination of coverage, real-world grounding, and annotation quality, forming a strong foundation for multilingual extension.

Dataset	Chart Types	Real-World Charts	#Charts	#QAs
PlotQA	3	✗	224K	28M
ChartQA	3	✓	21.9K	32.7K
OpenCQA	5	✓	–	–
ChartBench	9	✗	66.6K	599.6K
ChartX	18	✗	6K	6K

Table 6: Comparison of major English chart QA datasets.

ChartQA contributes high-quality, human-annotated real-world QA pairs, while ChartX adds diversity through synthetic chart types. Together, they balance realism, diversity, and usability, which is crucial for developing a representative multilingual benchmark.

A.2 Source Dataset Licenses

We use three existing chart QA datasets as part of our data construction pipeline. CHARTQA is released under the GPL-3.0 license[2], CHARTX under the CC-BY-4.0 license[3], and CHARTLLAMA under the MIT license[4]. All datasets are publicly available via HuggingFace and used in accordance with their respective licenses.

A.3 Language Definition

We follow the language selection and the definition of high/low-resource languages in (Maaz et al., 2024), which identifies Arabic, Urdu, Hindi, and Bengali as low-resource languages among the ten included in our benchmark.

A.4 Filtering Statistics and Data Retention

We report detailed filtering ratios and retained item counts across all stages of data construction when constructing POLYCHARTQA to ensure transparency and reproducibility:

- **Source Dataset Cleaning & Validation:** 11.2% filtered and 0.5% corrected through automated validation; all items passed normalization (remaining: 7,545).
- **Seed Data Generation (with Quality Control):** 35.9% filtered during JSON/code extraction and chart-type balancing (remaining: 4,840 core seed items).
- **Text Translation:** 23.2% filtered across 10 languages after automated validation (remaining per language: ~3,716).
- **Chart Image Translation:** 11.4% removed after rendering validation (remaining total: 32,897).
- **Final Visual Inspection (in Multilingual Data Quality control):** 20.5% filtered through manual inspection, resulting in a final dataset of 26,151 multilingual QA pairs.

These statistics demonstrate that each stage enforced strict quality thresholds, ensuring the reliability and linguistic–visual consistency of the final benchmark dataset. Since POLYCHARTQA-TRAIN serves as the training set, we did not record detailed statistics for it.

B Detailed Dataset Statistics of POLYCHARTQA

This section provides detailed data statistics of POLYCHARTQA. It covers Data Statistics by Language and Chart Type, Question and Answer Length Statistics, Per-language Distribution of Images and Questions, as well as the Distribution of Images, Questions, JSON, and Code for the English seed data in POLYCHARTQA-Test (§ B.1), and Data Statistics by Language and Chart Type for POLYCHARTQA-Train (§ B.2).

B.1 POLYCHARTQA-Test

Data Statistics by Language and Chart Type. We show the detailed statistics of POLYCHARTQA in Tables 8 and 9, including per-language and per-chart-type breakdowns for both images and QA pairs. Note that "EN" here does not refer to the original English dataset; instead, it was regenerated and processed through the same pipeline as other languages, with the only exception being the translation step.

Question and Answer Length Statistics. We report statistics of question and answer lengths across all ten languages in POLYCHARTQA, using token counts computed with the GPT-4o tokenizer. The distribution for each language, aggregated over

[2] https://huggingface.co/datasets/ahmed-masry/ChartQA
[3] https://huggingface.co/datasets/U4R/ChartX
[4] https://huggingface.co/datasets/listen2you002/ChartLlama-Dataset

training and test splits, is illustrated in Figure 8. These results highlight significant variation in textual length, which reflects both linguistic and orthographic diversity across languages.

Distribution of Images, Questions, JSON, and Code for English Seed Data. We also provide a detailed analysis of the English subset, which serves as the seed data for POLYCHARTQA. Figure 11 shows t-SNE visualizations of image and question embeddings, with points colored by chart type to reveal clustering based on visual and semantic chart characteristics. Figure 12 presents t-SNE plots of embeddings from the JSON data underlying the charts and the Python code used to generate them, again colored by chart type. These analyses illustrate the extent to which chart types can be distinguished within visual, textual, and structural representations.

Distribution of Images and Questions by Language. We further examine the distribution of images and questions in each language. Figure 9 presents a t-SNE visualization of CLIP image embeddings, while Figure 10 visualizes CLIP text embeddings of questions. In both cases, each subplot corresponds to a specific language. All points are uniformly colored to emphasize intra-language distribution rather than inter-category variation. These visualizations reveal the diversity and clustering patterns present in the multilingual data.

B.2 POLYCHARTQA-Train

Data Statistics by Language and Chart Type We show the detailed statistics of POLYCHARTQA-Train in Tables 10 and 11, including per-language and per-chart-type breakdowns for both images and QA pairs.

C Human Evaluation Details

C.1 Information of Human Annotators

We conducted a rigorous human evaluation to measure the quality of multilingual chart images and their question-answering pairs in POLYCHARTQA. All annotators are either native speakers with over 15 years of experience in the target language or individuals holding a bachelor's degree and official certification in the corresponding language. We recruit two annotators for each language.

C.2 Annotation Process

All annotations were collected via crowdsourcing. Annotators reviewed HTML-rendered charts and questions, and recorded their responses in structured Excel spreadsheets. Full instructions provided to human annotators are detailed below.

Full Human Evaluation Instructions

Evaluation Dimensions & Criteria:
(1) Image Quality Assessment: Assess the visual quality of the target language chart. Evaluate its clarity, the legibility and correctness of all text and graphical elements, and its overall professional integrity.

- 3: The image is clear, professional, and undistorted. All text and graphical elements are correctly displayed and legible. The chart type accurately reflects the data.
- 2: The chart has minor flaws, such as slight blurriness or minor display issues, but these do not significantly hinder comprehension.
- 1: The chart has major issues (e.g., distortion, illegible text, incorrect chart type) that hinder or prevent comprehension.

(2) QA Correctness Assessment: Assess if the question is relevant to the chart and if the answer is factually correct and fully supported by the information presented in the target language chart.

- 3: The question is relevant, and the answer is correct and fully supported by the chart data.
- 2: The QA pair has minor errors or ambiguities. The question might be slightly unclear, or the answer may have small inaccuracies.
- 1: The question is irrelevant to the chart, or the answer is factually incorrect or unsupported by the chart.

(3) Translation Accuracy: Evaluate the quality of the image and QA translation from English to the target language. Assess its fidelity, semantic consistency, and natural fluency, and check if it conforms to the target language's idiomatic expressions. Crucially, determine if the translation introduces any bias, misinformation, or framing.

- 3: The translation is accurate, fluent, and natural, conforming perfectly to the target language's conventions. It preserves the original meaning and key information without introducing any bias, misinformation, or framing.
- 2: The translation is mostly correct and preserves the core meaning, but has minor issues like awkward phrasing or does not feel fully idiomatic. It may subtly introduce minor bias or framing, but does not significantly mislead.
- 1: The translation has major errors, is semantically inconsistent, or is highly unnatural. Additionally, or as a primary issue, it introduces clear bias, misinformation, or framing that distorts the original message.

Figure 20 shows an example of the custom annotation interface designed for this task, enabling annotators to efficiently compare original and translated chart images as well as their corresponding question-answer pairs.

C.3 Annotation Results Details

We present the complete results of human annotations in Table 7. For each language, we report the average human score, inter-annotator agreement, and the weighted Cohen's κ between annotators. These consistently high scores indicate strong annotator consistency and confidence, further validating the overall quality and reliability of our dataset.

D More Implementation Details

D.1 Metric Details

For METEOR metric, we use its official code from huggingface[5].

D.2 Models Details

The general open-source LVLMs include Qwen2-VL (Wang et al., 2024b), Qwen2.5-VL (Bai et al., 2025), InternVL-2.5 (Chen et al., 2024c), InternVL-3 (Zhu et al., 2025), Phi-3 Vision (Abdin et al., 2024a), Phi-4 Multimodal (Abdin et al., 2024b), PaliGemma 2 (Team et al., 2024b), LLaVA-v1.6 (Liu et al., 2024b), LLaVA-OneVision (Li et al., 2024), Llama-3.2-Vision (Grattafiori et al., 2024), and DeepSeek-VL2 (Wu et al., 2024). For open-source multilingual LVLMs, we evaluate PALO (Maaz et al., 2024), Maya (Alam et al., 2024), Pangea (Yue et al., 2024), and Centurio (Geigle et al., 2025). The chart-specific category includes TinyChart (Zhang et al., 2024), ChartGemma (Masry et al., 2025b), ChartInstruct (Masry et al., 2024), ChartLlama (Han et al., 2023), and ChartAssistant (Meng et al., 2024). Closed-source category comprises Gemini-2.5-Pro (Comanici et al., 2025) and GPT-4o (Hurst et al., 2024). Closed-source models are accessed via their official APIs, while open-source models are run using their instruct versions available on the Hugging Face Model Hub.

E More Experiments

We further conduct a series of experiments on model inference and training, including ablations on English data ratio (§E.1), two-stage post-training (§E.2) and fine-tuning settings. These analyses reveal key insights into the weaknesses of current models and provide guidance for improving their multilingual chart understanding capabilities. We also present the complete experimental results corresponding to the main paper, including fine-tuning results (§E.3), and ablation on training data percentage (§E.4).

E.1 Ablation on English Data Ratio

To investigate the impact of English data proportion in multilingual fine-tuning, we conduct an ablation study by varying the ratio of English samples from 0% to 100% while keeping the total dataset size fixed at 70K QA pairs. The remaining proportion (i.e., non-English data) is evenly distributed across the other nine languages to ensure balanced multilingual representation. As shown in Figure 7 and Table 12, increasing the proportion of English data does not consistently enhance multilingual performance. Larger models like Qwen2.5-VL-7B maintain stable accuracy across all ratios, suggesting strong multilingual robustness, whereas smaller models such as InternVL3 exhibit slight degradation when English data dominates, likely due to reduced exposure to multilingual contexts. Overall, excessive reliance on English offers limited benefit and may even weaken cross-lingual generalization.

E.2 Two-stage Fine-tuning on Qwen-2.5-VL

In this section, we investigate whether the multilingual chart understanding ability of models can be further improved through a two-stage training strategy. We choose Qwen2.5-VL as our base model. In the first stage, we construct an alignment dataset using POLYCHARTQA-Train and other open-source resources. We then perform alignment training followed by fine-tuning on POLYCHARTQA-Train. Additionally, we examine the impact of unfreezing the vision encoder in each stage on overall performance. We further discuss the results and provide training insights below.

Data Construction for Alignment Stage In the alignment stage, we aim to achieve multilingual alignment through a chart-to-JSON prediction task using the chart metadata from POLYCHARTQA-Train. To further strengthen multilingual visual–textual grounding, we incorporate additional document and chart OCR tasks from external datasets, including MTVQA, PangeaOCR, and SMPQA. In total, this stage involves approximately 850K samples, comprising:

1. POLYCHARTQA-Train. We extract image–JSON pairs from POLYCHARTQA-Train,

[5] https://huggingface.co/spaces/evaluate-metric/meteor

Language	Image Quality			QA Relevance			Translation Accuracy		
	Avg. Score	Disag.	κ_w	Avg. Score	Disag.	κ_w	Avg. Score	Disag.	κ_w
Arabic	2.94	2	0.929	2.97	5	0.656	2.79	3	0.964
Urdu	2.71	3	0.971	2.92	4	0.891	2.71	7	0.932
Hindi	2.93	3	0.908	3.00	0	–*	2.95	3	0.874
Bengali	2.91	7	0.829	2.98	2	0.796	2.92	6	0.837
Chinese	2.96	2	0.896	2.98	2	0.796	2.95	3	0.874
French	2.92	4	0.891	2.95	5	0.789	2.91	1	0.976
Spanish	2.84	2	0.970	2.95	5	0.789	2.87	5	0.912
Russian	2.65	5	0.956	2.71	3	0.971	2.92	2	0.946
Japanese	2.86	2	0.967	2.90	6	0.867	2.95	5	0.789
English	2.95	1	0.958	2.98	2	0.796	2.96	4	0.792
Average	**2.87**	**3.6**	**0.927**	**2.93**	**3.9**	**0.817**	**2.89**	**3.9**	**0.889**

*Kappa is undefined due to zero variance (100% agreement). This entry was excluded from the average calculation.

Table 7: Detailed human scores and inter-annotator agreement scores for each language and evaluation dimension. Scores are based on 250 items per language rated by two annotators. "Disag." shows the raw count of differing ratings and κ_w denotes weighted Cohen's κ.

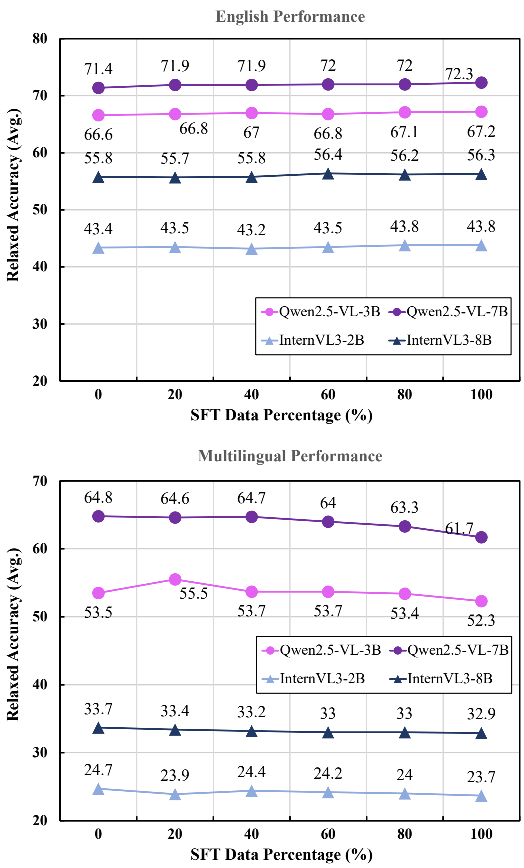


Figure 7: Performance on POLYCHARTQA with respect to the English data ratio across different model families.

yielding approximately 131K instances.

2. **MTVQA.** We incorporate the full training split of MTVQA (Tang et al., 2024), which contains 21K chart–QA pairs.

3. **Pangea.** We include 300K OCR data samples from the Pangea-OCR dataset (Yue et al., 2024).

4. **SMPQA-Reconstructed.** Following Geigle et al. (2025), we adapt SMPQA to our 10-language setting by reconstructing 410K synthetic chart-OCR training examples.

Two-stage Training Results We apply LoRA (Hu et al., 2022) in both stages with a fixed $r = 128$. The alignment stage uses a learning rate of $5e^{-}5$, while the instruction tuning stage uses a learning rate of $1e^{-}5$. Each stage is trained for one epoch.

Table 13 and Table 14 present the full ablation results of Qwen2.5-VL-3B and 7B, respectively. Across both model sizes, we observe consistent patterns: (i) fine-tuning alone provides substantial gains over the baseline, and (ii) incorporating an additional alignment stage further improves performance. Notably, the configuration where the vision encoder is unfrozen during alignment but frozen during instruction tuning achieves the highest accuracy in both models (63.6 for 3B, 68.0 for 7B). These results confirm that gradual visual adaptation followed by stabilization is a robust strategy for enhancing multilingual chart understanding across different model scales. This also indicates that the ability of models to understand multilingual charts can be further enhanced through additional training strategies.

E.3 Full Results of Fine-tuning on PolyChartQA-Train

We provide the complete fine-tuning results of various multilingual LVLMs on PolyChartQA-Train. This extended analysis reports per-language accuracy across all ten languages, offering a detailed view of how fine-tuning impacts different linguistic settings and model scales. As shown in Table 15, all models exhibit consistent improvements after fine-tuning, with particularly large gains for smaller or previously weaker models. Results also show that fine-tuning yields the most significant relative improvements in low-resource languages such as Urdu, Bengali, and Hindi, where accuracies often increase by over 100%, reflecting the strong transferability of multilingual chart instruction data. In contrast, high-resource languages such as English, Chinese, and French experience smaller yet consistent improvements, suggesting a saturation effect from stronger pretraining. Overall, these results indicate that fine-tuning primarily bridges multilingual reasoning gaps, especially in linguistically underrepresented settings.

E.4 Full Results of Ablation on Training Data Percentage

The full results in Table 16 confirm a consistent positive correlation between data volume and model performance across all architectures. The most substantial gains occur within the first 20–40% of training data, after which improvements gradually plateau. Notably, smaller models (e.g., InternVL3-2B) reach saturation earlier, while larger ones such as Qwen2.5-VL-7B continue to benefit steadily from additional data, underscoring their stronger data utilization capacity.

F Full Prompt Templates Used in Our Study

In this section, we present all prompt templates used throughout our PolyChartQA data pipeline. This includes the pipeline prompts for data cleaning, generation, translation, and consistency checking.

F.1 Prompts Used in Seed Data Preparation

The question-answer pair rewriting prompt used for **answer verification** of source datasets is shown in Figure 13. The question-answer pair rating prompt used for **answer standardization** of source datasets is shown in Figure 14. The prompt used for structured JSON extraction and visualization code generation during seed data construction is shown in Figure 15. The **visual fidelity** prompt used for quality control in seed data generation is shown in Figure 16. The **QA validity** prompt used for quality control in seed data generation is shown in Figure 17.

F.2 Prompts Used in Multilingual Chart Generation

The **translation** prompt used for multilingual text translation is shown in Figure 18. The **translation consistency** prompt used for back-translation verification is shown in Figure 19.

Chart Type	EN	AR	BN	ES	FR	HI	JA	RU	UR	ZH	Total
3d-bar	40	31	27	35	35	30	26	30	30	26	**310**
area	106	79	76	84	78	86	61	65	68	63	**766**
bar	600	447	507	505	471	547	409	477	514	393	**4870**
box	171	144	155	148	144	153	131	132	153	134	**1465**
bubble	81	32	39	38	38	40	33	35	37	35	**408**
candlestick	86	62	67	74	62	70	50	56	61	56	**644**
funnel	211	148	155	158	154	165	121	142	137	117	**1508**
heatmap	183	133	149	149	153	160	120	134	153	125	**1459**
histogram	219	167	177	180	187	182	141	162	181	137	**1733**
line	600	491	500	551	521	539	436	516	509	402	**5065**
multi-axes	77	49	53	52	58	55	42	48	58	45	**537**
pie	190	133	148	150	148	162	120	130	146	93	**1420**
radar	42	23	25	26	24	29	27	24	26	21	**267**
rings	123	80	83	91	95	92	72	66	85	76	**863**
rose	84	46	58	53	61	64	36	44	54	34	**534**
treemap	104	74	78	85	75	78	68	63	72	60	**757**
Total	**2917**	**2139**	**2297**	**2379**	**2304**	**2452**	**1893**	**2124**	**2284**	**1817**	**22606**

Table 8: Detailed statistics of Image counts per chart type across all languages in POLYCHARTQA.

Chart Type	EN	AR	BN	ES	FR	HI	JA	RU	UR	ZH	Total
3d-bar	40	31	27	35	35	30	26	30	30	26	**310**
area	107	80	77	85	79	87	62	66	69	64	**776**
bar	696	592	670	669	627	733	535	638	685	517	**6362**
box	171	144	155	148	144	153	131	132	153	134	**1465**
bubble	81	32	39	38	38	40	33	35	37	35	**408**
candlestick	86	62	67	74	62	70	50	56	61	56	**644**
funnel	211	148	155	158	154	165	121	142	137	117	**1508**
heatmap	183	133	149	149	153	160	120	134	153	125	**1459**
histogram	219	167	177	180	187	182	141	162	181	137	**1733**
line	646	689	718	794	739	770	602	734	720	551	**6963**
multi-axes	77	49	53	52	58	55	42	48	58	45	**537**
pie	210	146	164	165	163	178	129	145	159	106	**1565**
radar	42	23	25	26	24	29	27	24	26	21	**267**
rings	123	80	83	91	95	92	72	66	85	76	**863**
rose	84	46	58	53	61	64	36	44	54	34	**534**
treemap	104	74	78	85	75	78	68	63	72	60	**757**
Total	**3080**	**2496**	**2695**	**2802**	**2694**	**2886**	**2195**	**2519**	**2680**	**2104**	**26151**

Table 9: Detailed statistics of Question-Answer (QA) pair counts per chart type across all languages in POLY-CHARTQA

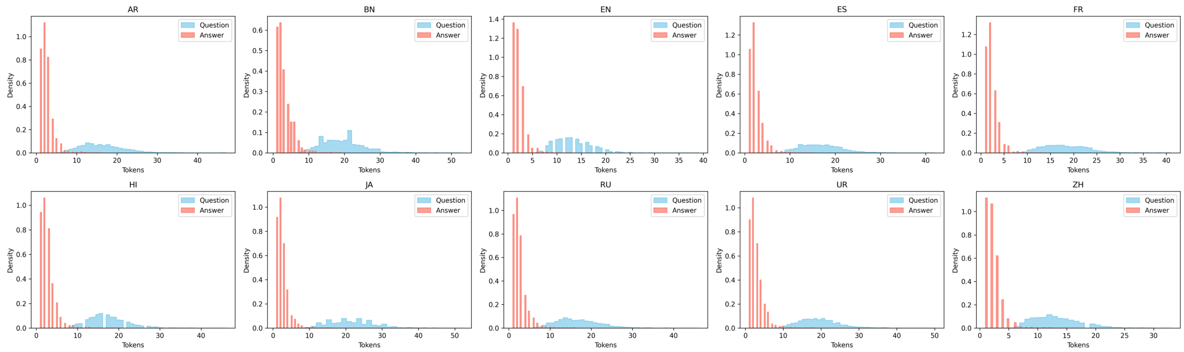


Figure 8: Question and answer length statistics in POLYCHARTQA.

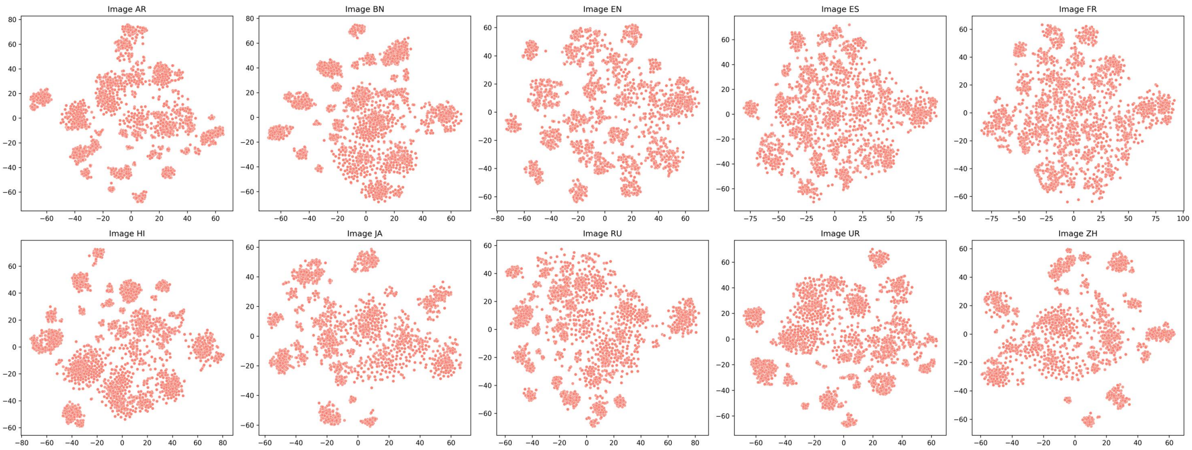


Figure 9: Distribution of images in POLYCHARTQA by language.

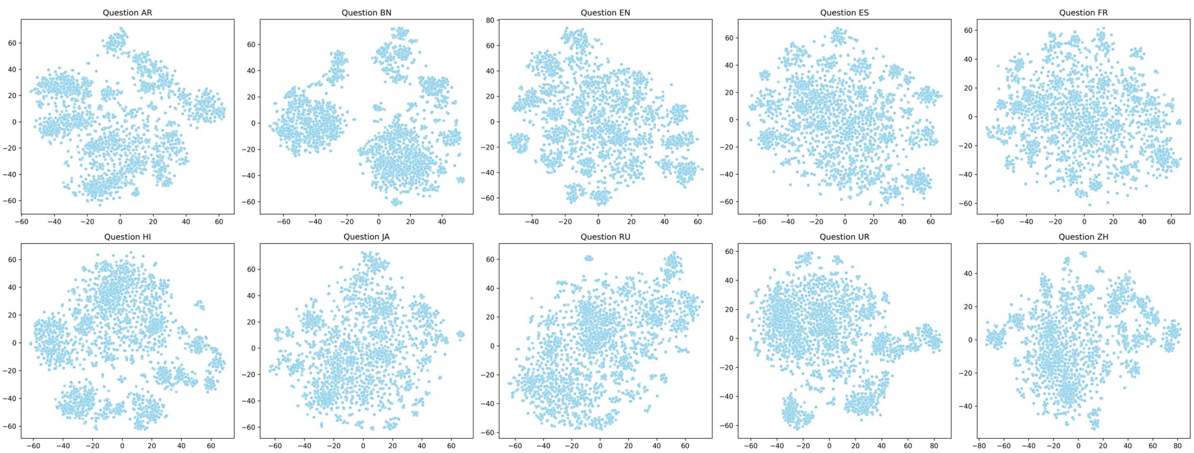


Figure 10: Distribution of questions in POLYCHARTQA by language.

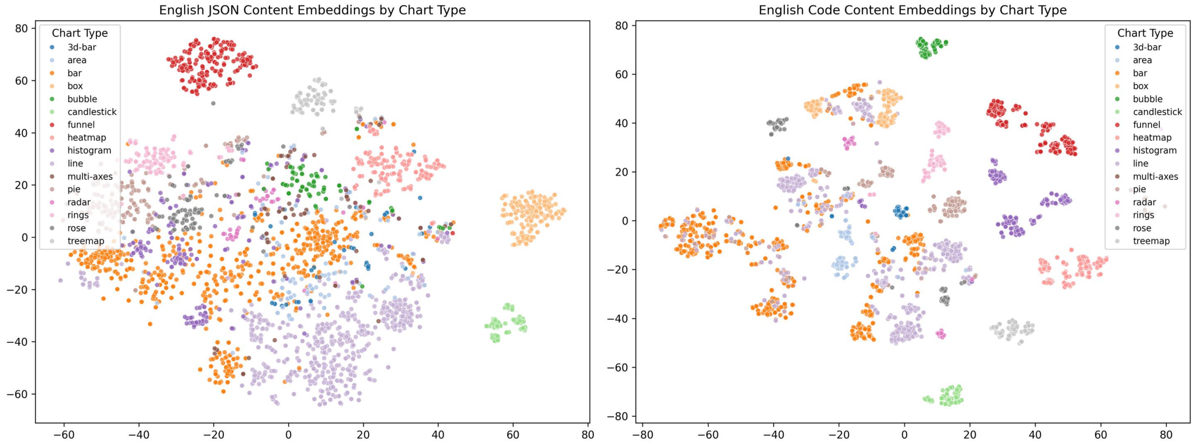


Figure 11: Distribution of images and questions in English by chart type in POLYCHARTQA.

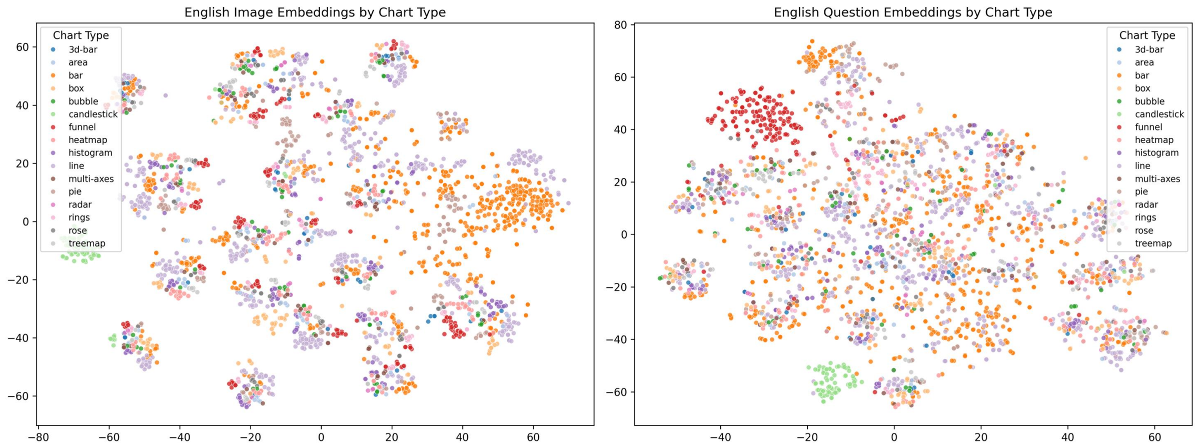


Figure 12: Distribution of JSON data and code in English by chart type in POLYCHARTQA.

Chart Type	**AR**	**BN**	**EN**	**ES**	**FR**	**HI**	**JA**	**RU**	**UR**	**ZH**	**Total**
3d-bar	4	4	4	2	3	3	3	2	4	4	**33**
area	1	1	1	1	1	1	1	1	1	1	**10**
bar	7834	7978	8876	7726	7878	8049	7883	7955	7804	8000	**79983**
box	50	47	57	47	48	46	46	48	49	50	**488**
candlestick	231	224	267	226	240	244	222	223	223	231	**2331**
funnel	103	107	118	101	96	107	102	106	100	102	**1042**
gantt	110	101	143	122	119	122	114	117	99	114	**1161**
heatmap	154	160	218	162	155	167	168	169	174	153	**1680**
line	3281	3383	3937	3220	3281	3374	3294	3374	3348	3340	**33832**
other	13	17	17	14	16	17	16	15	14	13	**152**
pie	630	629	781	602	593	645	632	643	631	630	**6416**
radar	184	176	203	166	165	185	165	177	167	173	**1761**
rings	66	68	88	67	69	68	68	69	68	70	**699**
scatter	186	193	222	182	185	185	188	200	187	196	**1924**
Total	**12847**	**13088**	**14932**	**12638**	**12849**	**13213**	**12902**	**13099**	**12869**	**13078**	**131515**

Table 10: Detailed statistics of Image counts per chart type across all languages in POLYCHARTQA-Train.

Chart Type	AR	BN	EN	ES	FR	HI	JA	RU	UR	ZH	Total
3d-bar	41	41	41	21	30	31	30	20	41	41	**317**
area	1	1	1	1	1	1	1	1	1	1	**10**
bar	33161	33764	38339	32794	33463	33940	33385	33626	32962	33998	**339432**
box	510	479	580	478	491	467	468	488	500	510	**4971**
candlestick	2279	2209	2639	2231	2369	2409	2190	2208	2201	2288	**23023**
funnel	1055	1097	1202	1042	982	1096	1055	1086	1009	1044	**10724**
gantt	1098	1008	1428	1218	1188	1219	1139	1169	989	1138	**11592**
heatmap	1547	1609	2190	1629	1558	1679	1688	1698	1750	1539	**16887**
line	25110	25942	30793	24645	25110	26013	25196	25998	25539	25763	**260109**
other	98	129	129	91	128	129	119	109	98	115	**1145**
pie	3833	3779	4901	3617	3615	3947	3863	3909	3806	3856	**39126**
radar	1845	1766	2042	1660	1662	1860	1663	1774	1671	1745	**17688**
rings	669	688	893	680	700	688	691	700	688	711	**7108**
scatter	1857	1933	2221	1824	1856	1857	1885	1997	1880	1955	**19265**
Total	**73104**	**74445**	**87399**	**71931**	**73153**	**75336**	**73373**	**74783**	**73135**	**74704**	**751363**

Table 11: Detailed statistics of QA pair counts per chart type across all languages in POLYCHARTQA-Train.

Model	% EN Data	EN	ZH	FR	ES	RU	JA	AR	HI	UR	BN	Avg. (w EN)	Avg. (w/o EN)
InternVL3-2B	0	43.4	**35.9**	**34.3**	**36.0**	29.2	**26.3**	**18.3**	16.5	**15.6**	**13.1**	**26.9**	**24.7**
	20	43.5	33.1	32.1	35.6	29.6	24.4	**18.3**	16.2	15.4	12.4	26.2	23.9
	40	43.2	34.8	33.9	35.8	29.4	25.8	18.2	**16.3**	15.7	12.6	26.6	24.4
	60	43.5	34.5	33.5	35.5	29.1	25.0	18.1	16.0	15.4	12.8	26.4	24.2
	80	**43.8**	33.7	33.1	35.8	28.8	24.1	18.2	16.0	15.6	12.7	26.3	24.0
	100	**43.8**	32.9	32.3	35.3	28.9	23.8	18.2	15.9	15.6	12.5	26.1	23.7
InternVL3-8B	0	55.8	**45.4**	**47.6**	**51.0**	**41.3**	**40.3**	**22.4**	**21.3**	**18.4**	19.1	**36.3**	**33.7**
	20	55.7	44.8	47.0	50.5	**41.4**	40.0	22.2	21.2	18.2	**19.1**	36.1	33.4
	40	55.8	43.8	46.4	50.4	41.3	39.7	22.3	21.1	18.1	18.7	35.8	33.2
	60	56.4	42.9	46.3	50.3	41.4	39.5	22.1	21.1	**18.2**	18.6	35.8	33.0
	80	56.2	42.0	46.1	50.2	41.4	39.7	22.1	21.3	18.2	18.6	35.7	33.0
	100	**56.3**	40.8	46.4	50.2	41.0	40.0	22.3	21.3	18.0	18.7	35.6	32.9
Qwen2.5-VL-3B	0	66.6	60.6	63.0	62.7	59.7	53.8	**54.0**	47.2	**39.6**	43.1	55.0	53.5
	20	66.8	**61.7**	**63.9**	62.8	**61.9**	**57.3**	**55.5**	**50.5**	42.7	**45.4**	**56.8**	**55.5**
	40	67.0	60.5	63.5	62.6	60.6	53.6	54.0	47.6	40.2	43.0	55.3	53.7
	60	66.8	60.8	63.4	62.8	61.0	53.0	53.9	47.6	40.0	43.0	55.3	53.7
	80	67.1	60.7	63.7	63.2	61.1	51.1	53.4	47.4	38.7	42.6	55.0	53.4
	100	**67.2**	59.6	63.4	63.2	60.3	48.6	51.6	46.4	37.1	42.0	54.1	52.3
Qwen2.5-VL-7B	0	71.4	**68.0**	70.2	**69.5**	**68.9**	**66.2**	**63.4**	**62.4**	**56.6**	**58.8**	**65.5**	**64.8**
	20	71.9	68.3	69.5	69.4	67.9	67.2	63.1	62.1	56.2	58.8	65.4	64.6
	40	71.9	67.9	**70.5**	69.7	68.1	66.5	63.1	62.3	56.3	58.6	65.5	64.7
	60	72.0	67.3	69.9	69.8	67.6	66.1	62.2	61.0	54.9	58.5	64.9	64.0
	80	72.0	66.5	69.2	69.4	67.2	64.6	61.2	60.6	54.3	57.5	64.3	63.3
	100	**72.3**	63.5	69.7	69.4	67.1	59.6	61.5	58.6	51.1	54.8	62.9	61.7

Table 12: Overall performance on the POLYCHARTQA benchmark under different **English data ratios**. For each model category, the best score per column is in **bold** and the second-best is underlined.

Training Strategy	Stage1	Stage2	EN	ZH	FR	ES	RU	JA	AR	UR	HI	BN	Avg. (w EN)	Avg. (w/o EN)
Baseline	✗	✗	67.4	59.6	61.8	62.5	58.0	48.8	51.4	37.2	45.7	43.0	53.7	51.8
SFT only	✗	❄	68.2	64.1	66.1	65.4	64.9	63.1	59.0	49.8	56.8	54.0	61.1	60.2
	✗	🔥	68.2	64.0	**66.3**	65.9	65.0	63.3	60.7	51.5	58.8	55.5	61.9	61.1
Align+SFT	❄	❄	68.8	64.2	66.1	**66.2**	64.3	62.7	61.3	53.4	57.9	53.5	61.9	60.9
	🔥	❄	69.0	**64.8**	65.5	**66.2**	65.2	**64.5**	**63.9**	**56.5**	**61.3**	**58.4**	**63.6**	**62.8**
	🔥	🔥	**69.3**	64.1	64.9	66.0	**65.5**	**64.5**	63.8	55.6	61.1	**58.4**	63.4	62.6

Table 13: Performance of different training strategies on Qwen2.5-VL-3B across various languages. ❄ and 🔥 indicate that the vision encoder is frozen or unfrozen, respectively, during each stage. ✗ denotes that the stage is skipped. Bold values denote the best performance.

Training Strategy	Stage1	Stage2	EN	ZH	FR	ES	RU	JA	AR	UR	HI	BN	Avg. (w EN)	Avg. (w/o EN)
Baseline	✗	✗	53.8	53.0	53.0	53.0	53.0	53.0	53.0	53.0	53.0	53.0	53.8	53.0
SFT only	✗	❄	73.1	68.5	71.1	70.0	68.5	67.7	65.5	58.6	64.9	60.9	66.9	66.1
	✗	🔥	72.6	68.8	70.6	70.0	**68.6**	67.9	65.1	60.0	65.2	61.6	67.0	66.2
Align+SFT	❄	❄	**73.6**	**69.6**	70.9	**70.8**	67.8	67.8	65.6	61.0	65.1	62.2	67.5	66.7
	🔥	❄	73.7	69.2	**71.3**	70.7	**68.0**	**68.0**	**66.1**	**62.7**	**66.6**	**62.9**	**68.0**	**67.2**
	🔥	🔥	73.5	69.1	70.4	70.2	68.2	68.1	65.3	59.6	64.4	62.1	67.1	66.3

Table 14: Performance of different training strategies on Qwen2.5-VL-7B across various languages. ❄ and 🔥 indicate that the vision encoder is frozen or unfrozen, respectively, during each stage. ✗ denotes that the stage is skipped. Bold values denote the best performance.

Model	EN	ZH	FR	ES	RU	JA
Qwen2.5-VL-3B	67.4	59.6	61.8	62.5	58.0	48.8
w/ fine-tuning	68.2 (+1.2%)	64.1 (+7.6%)	66.1 (+7.0%)	65.4 (+4.6%)	64.9 (+11.9%)	63.1 (+29.3%)
Qwen2.5-VL-7B	60.5	58.3	57.2	59.0	56.8	55.6
w/ fine-tuning	73.1 (+20.8%)	68.5 (+17.5%)	71.1 (+24.3%)	70.0 (+18.6%)	68.5 (+20.6%)	67.7 (+21.8%)
InternVL-3-2B	43.7	35.3	30.8	33.5	25.6	26.9
w/ fine-tuning	48.9 (+11.9%)	46.5 (+31.7%)	43.1 (+39.9%)	41.6 (+24.2%)	36.6 (+43.0%)	39.4 (+46.5%)
InternVL-3-8B	54.1	39.4	43.4	45.8	38.1	39.7
w/ fine-tuning	63.1 (+16.6%)	57.3 (+45.4%)	57.7 (+32.9%)	58.0 (+26.6%)	50.7 (+33.1%)	53.1 (+33.8%)
PaliGemma2-3B	26.6	14.7	19.7	21.5	13.9	10.7
w/ fine-tuning	33.9 (+27.4%)	28.5 (+93.9%)	32.3 (+64.0%)	33.1 (+54.0%)	30.0 (+115.8%)	28.9 (+170.1%)
LLaVA-v1.6-7B	24.8	12.9	18.9	18.2	13.5	11.5
w/ fine-tuning	36.6 (+47.6%)	22.2 (+72.1%)	33.6 (+77.8%)	33.8 (+85.7%)	24.6 (+82.2%)	20.9 (+81.7%)

Model	AR	UR	HI	BN	Avg. (w EN)	Avg. (w/o EN)
Qwen2.5-VL-3B	51.4	37.2	45.7	43.0	53.7	51.8
w/ fine-tuning	59.0 (+14.8%)	49.8 (+33.9%)	56.8 (+24.3%)	54.0 (+25.6%)	61.1 (+13.8%)	60.2 (+16.2%)
Qwen2.5-VL-7B	52.0	43.7	49.4	46.4	53.8	53.0
w/ fine-tuning	65.5 (+26.0%)	58.6 (+34.1%)	64.9 (+31.4%)	60.9 (+31.3%)	66.9 (+24.3%)	66.1 (+24.7%)
InternVL-3-2B	17.1	14.6	15.7	11.9	25.6	23.1
w/ fine-tuning	21.6 (+26.3%)	18.3 (+25.3%)	20.7 (+31.8%)	18.2 (+52.9%)	33.3 (+30.1%)	31.2 (+35.1%)
InternVL-3-8B	21.4	17.2	20.2	17.5	33.8	31.0
w/ fine-tuning	26.6 (+24.3%)	24.3 (+41.3%)	26.4 (+30.7%)	24.2 (+38.3%)	44.0 (+30.2%)	41.4 (+33.5%)
PaliGemma2-3B	15.9	12.2	14.3	10.2	16.3	14.9
w/ fine-tuning	26.5 (+66.7%)	26.2 (+114.8%)	27.1 (+89.5%)	22.7 (+122.5%)	29.0 (+77.9%)	28.4 (+90.6%)
LLaVA-v1.6-7B	12.0	7.7	10.0	6.7	13.9	12.4
w/ fine-tuning	20.3 (+69.2%)	20.2 (+162.3%)	19.5 (+95.0%)	19.2 (+186.6%)	25.5 (+83.5%)	24.0 (+93.5%)

Table 15: Fine-tuning Results using POLYCHARTQA-Train across different model families and sizes. Performance gains are highlighted in green.

Model	% Data	EN	ZH	FR	ES	RU	JA	AR	UR	HI	BN	Avg. (w EN)	Avg. (w/o EN)
Qwen2.5-VL-3B	0	67.4	59.6	61.8	62.5	58.0	48.8	51.4	37.2	45.7	43.0	53.7	51.8
	20	67.0	61.8	64.1	63.0	62.0	57.1	56.3	43.6	51.1	46.5	57.3	56.0
	40	67.5	62.6	65.4	64.5	63.3	60.4	57.5	46.5	53.8	50.4	59.2	58.1
	60	68.4	63.8	66.1	65.4	64.9	62.1	58.7	49.0	56.6	53.5	60.8	59.8
	80	**68.5**	**64.3**	**66.4**	**65.6**	64.9	**63.6**	**59.1**	**50.3**	**57.3**	**54.2**	**61.4**	**60.5**
	100	68.2	64.1	66.1	65.4	**64.9**	63.1	59.0	49.8	56.8	54.0	61.1	60.2
Qwen2.5-VL-7B	0	60.5	58.3	57.2	59.0	56.8	55.6	52.0	43.7	49.4	46.4	53.8	53.0
	20	69.8	64.4	67.0	67.2	66.1	62.6	59.5	52.5	58.5	54.7	62.3	61.3
	40	71.9	67.2	69.7	69.2	67.6	65.6	61.9	55.7	61.7	57.3	64.8	63.9
	60	72.2	66.9	69.7	69.0	67.7	65.6	61.6	55.4	61.5	57.5	64.8	63.8
	80	72.1	68.1	70.1	69.2	**68.2**	66.2	62.8	56.8	62.2	58.2	65.4	64.5
	100	**73.1**	**68.5**	**71.1**	**70.0**	68.5	**67.7**	**65.5**	**58.6**	**64.9**	**60.9**	**66.9**	**66.1**
InternVL3-2B	0	43.7	35.3	30.8	33.5	25.6	26.9	17.1	14.6	15.7	11.9	25.6	23.1
	20	47.3	41.6	38.9	39.5	32.5	33.9	19.4	18.1	19.5	17.0	30.7	28.5
	40	48.0	45.6	42.2	41.1	35.6	39.2	21.1	18.4	20.6	18.2	32.8	30.8
	60	50.0	46.9	44.4	43.0	37.3	40.4	22.6	19.2	21.2	19.0	34.2	32.1
	80	**50.1**	**47.5**	**45.3**	**43.6**	**38.2**	**41.5**	**22.7**	**19.3**	**21.3**	**19.6**	**34.7**	**32.7**
	100	48.9	46.5	43.1	41.6	36.6	39.4	21.6	18.3	20.7	18.2	33.3	31.2
InternVL3-8B	0	54.1	39.4	43.4	45.8	38.1	39.7	21.4	17.2	20.2	17.5	33.8	31.0
	20	59.7	50.6	53.9	54.1	45.9	44.7	24.2	21.0	23.6	21.9	39.9	37.3
	40	61.7	55.6	56.9	56.5	49.0	49.4	26.6	23.1	24.8	23.7	42.6	40.1
	60	63.1	56.8	57.3	57.3	50.0	52.9	26.6	24.3	26.0	24.6	43.7	41.2
	80	**63.7**	**58.3**	**58.2**	**58.4**	**51.3**	**54.7**	**27.3**	**25.6**	**26.9**	**24.7**	**44.7**	**42.2**
	100	63.1	57.3	57.7	58.0	50.7	53.1	26.6	24.3	26.4	24.2	44.0	41.4

Table 16: Overall performance on PolyChartQA benchmark across different fine-tuning data proportions. For each model category, the best score per column is in **bold** and the second-best is underlined.

Prompt for Question-Answer Pair Rewriting

You are a data processing expert specializing in refining chart Question-Answering pairs for automated evaluation. Your goal is to process provided Question-Answer examples, classifying them (KEPT, MODIFIED, DELETE) and potentially shortening the label (answer) to a concise format suitable for exact match (or numerical match with tolerance) evaluation.

CORE INSTRUCTION: Assess the provided label in the context of the query. You MUST base the new_label strictly on information present in the original label. Do NOT generate new information or answers.

Input:
1. query: The question asked about a chart.
2. label: The original answer.

Task Steps (Follow Strictly):
1. Assess Query Suitability (DELETE):
If the query requires an answer that cannot be concise (e.g., trend, explanation, subjective, or complex comparison), set action: "DELETE", new_label: "", and stop.
2. Assess Label Conciseness (KEPT):
If the original label is already concise (single number, name, yes/no, short list, or "Unanswerable"), set action: "KEPT", new_label: label (exact copy), and stop.
3. Perform Modification (MODIFIED):
If the query is suitable and the label is verbose, set action: "MODIFIED", extract ONLY the core factual answer(s), format concisely (list, units, standardize "Data not available" as "Unanswerable"), and set as new_label.

Final Output Format:
Respond ONLY with the following JSON object (no other text):

```
{
  "action": "KEPT" | "MODIFIED" | "DELETE",
  "new_label": "string"
}
```

Rules:
- If action is DELETE, new_label must be "".
- If action is KEPT, new_label is identical to the original label.
- If action is MODIFIED, new_label is your concise rewrite.

Now, process the following input:

```
{ "query": "{query}", "label": "{label}" }
```

Figure 13: Prompt for question–answer pair rewriting.

Prompt for Question-Answer Pair Rating

You are an expert evaluator for chart question-answering pairs.
Your task is to assess the quality and correctness of the provided Answer in response to the Question, based solely on the information presented in the accompanying chart image. Assign a rating from 1 to 5 based on the criteria below.

Do not use any external knowledge or make assumptions beyond what is visually represented or directly calculable from the chart.

Rating Scale and Criteria:
- 5: Excellent / Fully Correct
  The answer is completely accurate according to the chart data; directly and fully addresses the question; all information is visible or calculable from the chart; no ambiguities or unsupported inferences.
- 4: Good / Mostly Correct
  Substantially correct, with only very minor inaccuracies or omissions; main point addressed; clearly derived from the chart.
- 3: Fair / Partially Correct
  Contains both correct and incorrect elements, or answers the wrong question, or relies on inferences not explicitly supported; addresses the question only partially or inaccurately.
- 2: Poor / Mostly Incorrect
  Contains significant errors contradicted by the chart; fundamentally misunderstands the chart or question; core claim is wrong according to the chart.
- 1: Very Poor / Completely Incorrect or Irrelevant
  Entirely false or irrelevant to the chart or question; no connection between the answer and the visual evidence.

Input Context (User Prompt):
1. Chart Image
2. Chart Question
3. Proposed Answer

Output Format:
Respond ONLY with a valid JSON object containing:

```
{
  "rating": <integer 1-5>,
  "reason": "<brief justification, referencing specific chart elements or data points where possible>"
}
```

Example Output (Score 5):

```
{
  "rating": 5,
  "reason": "The answer accurately states the value for Q3 Revenue is $1.2M, which matches the bar labeled Q3 on the chart."
}
```

Example Output (Score 3):

```
{
  "rating": 3,
  "reason": "The answer correctly identifies Product A as having the highest value, but misstates the exact percentage shown on the chart."
}
```

Example Output (Score 1):

```
{
  "rating": 1,
  "reason": "The answer discusses stock market trends, which are not present in the provided chart."
}
```

Now evaluate the specific chart image, question, and answer provided in the user prompt based on the 1–5 scale. Respond ONLY with the JSON object.

Figure 14: Prompt for question–answer pair rating.

Prompt for JSON and Code Extraction

You MUST act as an expert Python data visualization assistant. Your primary objective is to meticulously analyze a given chart image, extract its data and text into a structured JSON format suitable for translation, and then generate a robust Python script using Plotly that accurately recreates the chart solely from that JSON data. The generated script must preserve the original data order and handle multilingual text input correctly, in addition to proactively addressing potential layout issues.

Input:
1. <image_description>: A reference to, or the content of, the input chart image file.
2. <image_filename_base>: The base filename string for the input image (e.g., "my_chart"). This base name is crucial for naming the JSON file read by the script and the output PNG image.

Your Tasks (Execute Sequentially):
1) Analyze Image and Generate JSON Data Structure:
- Identify chart type and store as chart_type if useful.
- Extract all data series and categories (order must match original visual presentation). Store as chart_data.
- Extract all visible text elements into a texts dictionary, preserving original English, capitalization, and line breaks (
). If an element is missing, set its value to null.
- Extract primary colors as hex codes in a colors list, aligned with data series order.
- Final JSON contains chart_data, texts, colors, and optionally chart_type.

2) Generate Robust Python Plotly Code:
- Data source: The script must read only from <filename>.json and use the unpacked JSON for all chart content and styling. Absolutely no hardcoded data or text.
- Use Plotly (plotly.graph_objects) to recreate the chart. Iterate through JSON data in order; apply colors and texts per JSON content.
- Combine titles/subtitles and source/note using HTML as specified.
- Multilingual/Unicode support: Code must be language-agnostic, display provided strings as-is, and handle non-Latin scripts without logic changes.
- Layout: Prevent clipping/overlap with careful margins, anchors, and text placement. Font must be Arial.
- Output PNG as <filename>.png, with scale=2.
- Clean code: no extra installs, no function definitions, no unnecessary comments, only minimal print.

Output Format:
Return the output in exactly two code blocks:
- A single JSON code block containing the full JSON object.
- A single Python code block containing the full script.

Here is the filename <FILENAME> and the chart image.

Figure 15: Prompt for JSON extraction and visualization code generation.

Prompt for Visual Fidelity Assessment

You are an expert visual comparison and chart quality evaluator. Your task is to assess two chart images (Original, Rendered) based on two criteria: Semantic Consistency and Visual Flaws.

Input:
1. Original Chart Image
2. Rendered Chart Image (generated from code based on the original)

Task 1: Evaluate Semantic Consistency (Rating 1–5)

Assess whether the Rendered Image represents the same core data and key information as the Original Image. Focus on:
- Data Values & Proportions: Are numerical values (bars, points, slices) substantially the same? Do relative proportions match?
- Categories & Series: Do labels, axes, and legend entries match the original data structure and order?
- Text Content: Are titles, axis titles, legend labels, and other key text elements semantically identical or extremely close to the original?
- Color Hue Consistency: While exact shades may differ, do the primary colors preserve the same hue category (e.g., reds remain red/orange, blues remain blue/cyan)? A swap across hue families is a major inconsistency.
- Overall Message/Trend: Does the rendered chart convey the same main insight or pattern?

Ignore minor stylistic differences (fonts, gridlines, spacing) unless they hinder interpretation or violate the criteria above.

Rating Scale (1–5):
- 5: Highly Consistent — Near-perfect semantic match in data, text, color hues, and overall message; only negligible, non-misleading differences.
- 4: Mostly Consistent — Core data, text, and message are accurate; minor inaccuracies or color shade differences (hue preserved) without changing interpretation.
- 3: Moderately Consistent — Noticeable discrepancies; some key values/text differ, hues mismatched, or message partially distorted.
- 2: Poorly Consistent — Significant data errors, trends misrepresented, misleading text, or confusing color usage; interpretation fundamentally altered.
- 1: Inconsistent / Unrelated — Completely different data, topic, or structure.

Task 2: Identify Visual Flaws (Yes/No)

Determine whether the Rendered Image has significant visual flaws that impede understanding or indicate generation errors. Check for:
- Severe Text Overlap: Critical labels, titles, or data points overlap illegibly.
- Element Clipping: Data, labels, or legends are cut off by chart boundaries.
- Unreadable Text: Text is too small, blurry, or contains unsupported characters.
- Data Obscurity: Data points are hidden behind other elements.
- Empty/Malformed Chart: Blank output, error messages, or non-meaningful chart.
- Gross Layout Issues: Elements placed nonsensically, making the chart hard to interpret.

Answer Yes if any major flaws are present; No if not. Minor imperfections that do not hinder interpretation should be marked No.

Output Format:
Respond ONLY with a valid JSON object containing FOUR keys:
{
"similarity_rating": <integer 1–5>,
"similarity_reason": "<brief explanation for the similarity rating>",
"has_visual_flaws": <true | false>,
"flaw_reason": "<brief explanation if flaws were found, otherwise 'No significant flaws detected.'>"
}

Now evaluate the Original and Rendered images based on BOTH tasks. Respond ONLY with the JSON object.

Figure 16: Prompt for visual fidelity checking.

Prompt for QA Validity Assessment

You are an expert evaluator for chart question-answering pairs.
Your task is to assess the quality and correctness of the provided Answer in response to the Question, based solely on the information presented in the accompanying chart image. Assign a rating from 1 to 5 based on the criteria below.

Do not use any external knowledge or make assumptions beyond what is visually represented or directly calculable from the chart.

Rating Scale and Criteria:
- 5: Excellent / Fully Correct
  The answer is completely accurate according to the chart data; directly and fully addresses the question; all information is visible or directly calculable from the chart; no ambiguities or unsupported inferences.
- 4: Good / Mostly Correct
  Substantially correct; addresses the main point; may contain very minor inaccuracies or omissions that do not significantly mislead.
- 3: Fair / Partially Correct
  Mix of correct and incorrect information; may extract data but fail to answer the question; may rely on unsupported inferences; partially or inaccurately addresses the question.
- 2: Poor / Mostly Incorrect
  Contains significant factual errors; fundamentally misunderstands the chart or the question; core claim is wrong based on chart evidence.
- 1: Very Poor / Completely Incorrect or Irrelevant
  Completely false or irrelevant; no connection between the answer and the chart content.

Input Context:
1. Chart Image
2. Chart Question
3. Proposed Answer

Output Format:
You MUST respond ONLY with a valid JSON object containing two keys:

```
{
  "rating": <integer 1–5>,
  "reason": "<brief explanation for the assigned rating, referencing chart elements or data points where possible>"
}
```

Example Output (Score 5):

```
{
  "rating": 5,
  "reason": "The answer accurately states the value for Q3 Revenue is $1.2M, which matches the bar labeled Q3 in the chart."
}
```

Example Output (Score 3):

```
{
  "rating": 3,
  "reason": "The answer correctly identifies Product A as having the highest value, but misstates the exact number shown."
}
```

Example Output (Score 1):

```
{
  "rating": 1,
  "reason": "The answer discusses stock market trends, which are not present in the provided chart."
}
```

Now evaluate the specific chart image, question, and answer provided in the user prompt based on the 1–5 scale. Respond ONLY with the JSON object.

Figure 17: Prompt for QA validity checking.

Prompt for Translation (Back-Translation)

You are an expert linguist and JSON data localization specialist simulating a translation process. Your task is to translate a given JSON object representing chart data and its associated question–answer pairs from {source_language_name} ({source_language_code}) to {target_language_name} ({target_language_code}). You must intelligently identify and translate only the user-facing text while preserving the JSON structure and non-textual data precisely.

Input Data:
You will receive a JSON object containing two keys:
1. chart_json_data: The JSON object extracted from a chart (variable structure).
2. qa_pairs_to_translate: A list of dictionaries, each with "query" and "label" strings in {source_language_code}.

CRITICAL Instructions for Translation:
1. Goal:
 Produce a translated version of the input suitable for displaying the chart and Q&A in {target_language_name}.

2. Translate chart_json_data Recursively:
 - Traverse the entire structure (nested dictionaries and lists).
 - ONLY translate string values meant for user display in {source_language_name} (e.g., titles, axis labels, legend entries, annotations).
 - DO NOT translate or modify:
  * JSON keys
  * Numerical values (integers or floats)
  * Strings containing only numbers (e.g., "2023", "1.5")
  * Strings containing only numbers with percent signs (e.g., "55.5%", "-10%")
  * Hex color codes (e.g., "#1f77b4")
  * URLs, file paths, or system identifiers
  * Boolean strings ("true", "false")
  * Type or configuration keywords (e.g., "stacked_bar", "Arial", "auto"); if unsure, do NOT translate
  * null values and empty strings
 - Preserve units and symbols unless a direct and standard equivalent is always used in {target_language_name}.
 - The output JSON MUST be identical in structure and data types to the input. ONLY translatable string values may change.

3. Translate qa_pairs_to_translate:
 - Translate both "query" and "label" for each QA pair.
 - Consistency requirement: Use the exact same translation for terms that appear in both the chart JSON and the QA pairs.

4. Translation Quality Requirements:
 - Accuracy and Fidelity: Preserve factual meaning.
 - Naturalness and Fluency: Use grammatically correct and natural phrasing.
 - Consistency: Identical source terms must have identical translations.
 - Cultural Appropriateness: Ensure suitability for the target audience.
 - Linguistic Integrity: Maintain correct grammar, syntax, and style.
 - Vocabulary Usage: Use accurate and context-appropriate terminology.
 - Non-Latin/BiDi Support: Produce correct Unicode; standard rendering will handle text direction.
 - HTML Tags: Preserve tags such as
 in their original positions.

Output Format:
You MUST respond ONLY with a single, valid JSON object containing:
- translated_chart_json: The processed chart JSON, identical in structure to the input, with translations applied ONLY to user-facing text.
- translated_qa_pairs: A list of translated QA pairs in the original order, each containing:
 * translated_query
 * translated_label

Input Data to Process:

Figure 18: Prompt for multilingual translation.

Prompt for Translation Consistency Evaluation

You are an expert linguistic evaluator comparing two versions of content in {source_language_name} ({source_language_code}). One is the Original Content, and the other is the Back-Translated Content (translated to another language and then back to {source_language_name}).

Your task is to evaluate the semantic equivalence between the Original and Back-Translated content based on the provided context, assigning ratings on a 1–5 scale. The required output format depends on the provided context.

Input Format (Provided in User Prompt):
You will receive a JSON object with three keys:
1. context: A string indicating the type of content. Either "Chart JSON Texts" or "Question-Answer Pair".
2. original_content: The original content in {source_language_name}. This is either:
  - a JSON object (for chart texts), or
  - a dictionary like {"query": "...", "label": "..."} (for a QA pair).
3. back_translated_content: The back-translated content in {source_language_name}, matching the structure of original_content.

Evaluation Criteria and Rating Scale (1–5):
- Focus: Semantic meaning and preservation of key information. Does the back-translation convey the same meaning as the original?
- Ignore: Minor grammatical variations, stylistic changes, or synonymous phrasing unless they significantly alter meaning, introduce ambiguity, or omit/distort critical information.
- 5: Excellent Equivalence — Perfect semantic match; only trivial or stylistic differences.
- 4: Good Equivalence — Main meaning and most key information preserved; minor acceptable differences.
- 3: Fair Equivalence — General topic preserved, but some important details or nuances are lost or altered.
- 2: Poor Equivalence — Significant errors; key information is lost, distorted, or contradicted.
- 1: No Equivalence / Unrelated — Meaning is completely different, nonsensical, or unrelated.

CRITICAL: Output Format Based on Context

A. If context is "Chart JSON Texts":
- Evaluate the overall semantic equivalence of the translatable text content in back_translated_content JSON compared to original_content JSON.
- Respond ONLY with a single valid JSON object with TWO keys:
  {
   "rating": <integer 1–5>,
   "reason": "<brief justification>"
  }

B. If context is "Question-Answer Pair":
- Evaluate the Query and the Label (Answer) separately.
- Respond ONLY with a valid JSON object containing FOUR keys:
  {
   "query_rating": <integer 1–5>,
   "query_reason": "<brief justification for query equivalence>",
   "label_rating": <integer 1–5>,
   "label_reason": "<brief justification for label equivalence>"
  }

Final Instruction:
Analyze the original_content and back_translated_content according to the specified context. Respond ONLY with the valid JSON object matching the required output format for that context.

Figure 19: Prompt for translation consistency.

Evaluation Guidelines

You will be provided with two chart images (one in English, one in the target language) and their corresponding Question-Answer (QA) pairs. Your task is to critically evaluate the target language materials based on the three dimensions below.

General Guidelines:

- **Reference:** Use the English materials as a reference for comparison.
- **Evaluation:** For each of the three dimensions, provide a score from 1 to 3.

Evaluation Dimensions & Criteria:

Score	Description
Image Quality Assessment: Assess the visual quality of the target language chart. Evaluate its clarity, the legibility and correctness of all text and graphical elements, and its overall professional integrity.	
3	The image is clear, professional, and undistorted. All text and graphical elements are correctly displayed and legible. The chart type accurately reflects the data.
2	The chart has minor flaws, such as slight blurriness or minor display issues, but these do not significantly hinder comprehension.
1	The chart has major issues (e.g., distortion, illegible text, incorrect chart type) that hinder or prevent comprehension.
QA Correctness Assessment: Assess if the question is relevant to the chart and if the answer is factually correct and fully supported by the information presented in the target language chart.	
3	The question is relevant, and the answer is correct and fully supported by the chart data.
2	The QA pair has minor errors or ambiguities. The question might be slightly unclear, or the answer may have small inaccuracies.
1	The question is irrelevant to the chart, or the answer is factually incorrect or unsupported by the chart.
Translation Accuracy: Evaluate the quality of the image and QA translation from English to the target language. Assess its fidelity, semantic consistency, and natural fluency, and check if it conforms to the target language's idiomatic expressions. Crucially, determine if the translation introduces any bias, misinformation, or framing.	
3	The translation is accurate, fluent, and natural, conforming perfectly to the target language's conventions. It preserves the original meaning and key information without introducing any bias, misinformation, or framing.
2	The translation is mostly correct and preserves the core meaning, but has minor issues like awkward phrasing or does not feel fully idiomatic. It may subtly introduce minor bias or framing, but does not significantly mislead.
1	The translation has major errors, is semantically inconsistent, or is highly unnatural. Additionally, or as a primary issue, it introduces clear bias, misinformation, or framing that distorts the original message.

Evaluation Samples

Sample ID: bar_102_ar_001

English Original

Q: How much data usage is reported in Asia?

A: 1800 GB

Translated Version

Q: ما هو حجم استخدام البيانات المسجل في آسيا؟

A: 1800 جيجابايت

Figure 20: Human evaluation interface. Annotators review chart images and QA pairs in both source and target languages, providing quality ratings for image quality, QA correctness and translation accuracy.