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

**Anonymous ACL submission**

## Abstract

Charts are a universally adopted medium for interpreting and communicating data. However, existing chart understanding benchmarks are predominantly English-centric, limiting their accessibility and applicability to global audiences. In this paper, we present **PolyChartQA**, the first large-scale multilingual chart question answering benchmark covering 22,606 charts and 26,151 question-answering pairs across 10 diverse languages. PolyChartQA is built using a decoupled pipeline that separates chart data from rendering code, allowing multilingual charts to be flexibly generated by simply translating the data and reusing the code. We leverage state-of-the-art LLM-based translation and enforce rigorous quality control in the pipeline to ensure the linguistic and semantic consistency of the generated multilingual charts. PolyChartQA facilitates systematic evaluation of multilingual chart understanding. Experiments on both open- and closed-source large vision-language models reveal a significant performance gap between English and other languages, especially low-resource ones with non-Latin scripts. This benchmark lays a foundation for advancing globally inclusive vision-language models.

## 1 Introduction

Charts are ubiquitous tools for communicating quantitative information and supporting analytical reasoning in science, business, journalism, and daily life. Accurate chart interpretation is fundamental to data-driven decision-making. With the advent of large vision-language models (LVLMs), significant progress has been made in perceiving and reasoning over charts, plots, and diagrams. Recent advances have demonstrated that these models can answer complex questions (Masry et al., 2022; Xia et al., 2024; Wang et al., 2024c; Masry et al., 2025), summarize important contents (Rahman et al., 2022; Tang et al., 2023), and even regenerate chart images (Moured et al., 2024; Yang et al., 2024) based on chart contents.

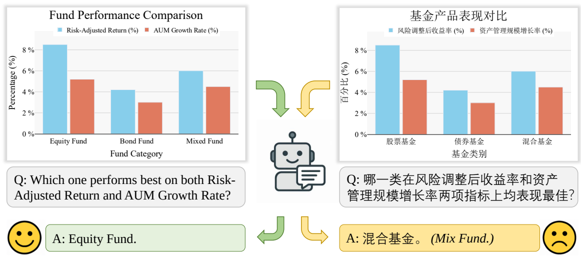


Figure 1: Example of inconsistent chart understanding by LVLMs in English and Chinese. The model provides the correct answer for the English chart question, but fails to generalize to the Chinese equivalent.

Despite these advancements, existing datasets and benchmarks for chart understanding remain overwhelmingly English-centric, overlooking the unique challenges posed by multilingual and cross-script chart comprehension. As illustrated in Figure 1, leading LVLMs may succeed on English chart questions but fail to generalize to their counterparts in other languages. This English-only bias presents a substantial barrier to the global deployment of chart understanding models, limiting their accessibility for speakers of underrepresented languages. While recent efforts (Chen et al., 2024a; Heakl et al., 2025) have introduced bilingual chart-related data, they remain limited in both scale and linguistic diversity. To date, no comprehensive benchmark exists for evaluating multilingual chart understanding capabilities in LVLMs.

Meanwhile, most multilingual and multimodal benchmarks focus on understanding and reasoning over natural images (Pfeiffer et al., 2021; Romero et al., 2024; Liu et al., 2021; Yu et al., 2025; Liu et al., 2024b; Xuan et al., 2025; Wang et al., 2024a), offering only limited attention to images with structured information such as charts. Although benchmarks such as M3Exam (Zhang et al., 2023), xMMMU (Yue et al., 2024), and SMPQA (Geigle et al.,

2025) have incorporated charts into their datasets, these tasks are often limited to recognizing characters or shallow pattern matching, lacking the depth of holistic reasoning. Thus, they fail to provide a systematic framework for evaluating chart understanding across diverse languages.

Constructing multilingual chart understanding datasets presents two major challenges: (1) manual annotation of high-quality question–answer pairs across multiple languages is prohibitively expensive and time-consuming, particularly for low-resource languages. (2) ensuring precise semantic and visual consistency between translated questions and chart images demands careful control, as misalignment can easily arise during translation or rendering.

To address these challenges, we propose a robust and extensible data pipeline comprising three core components. First, we **decouple** each chart image into two representations: a structured JSON file that encodes chart data and metadata, and a corresponding executable template code for rendering the chart. Second, we employ advanced large language models to perform **joint translation** of both the chart JSON specification and all associated QA pairs. This approach keeps terminology and semantics aligned across all languages and scripts. Third, we **automatically render** high-fidelity chart images in each target language using the translated JSON and templates, followed by a multi-stage quality control combining automated checks and human verification. This pipeline ensures high-quality, semantically faithful, and visually consistent chart datasets across multiple languages.

Using this pipeline, we construct POLYCHARTQA, the first large-scale benchmark for multilingual chart understanding. POLYCHARTQA comprises approximately 22K chart images and 26K QA pairs in ten widely spoken languages: English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, and Japanese, collectively covering over 65% of the global population (Maaz et al., 2024). The benchmark includes both real-world and synthetic charts, offering a diverse and rigorously curated resource for evaluating and advancing multilingual chart understanding.

With POLYCHARTQA, we conduct the first systematic, large-scale evaluation of multilingual chart understanding in LVLMs. Our experiments across both open- and closed-source LVLMs—covering general-purpose and multilingual-specialized models—yields three key findings: (1) All current models show clear limitations in chart understanding for non-English languages. (2) Significant performance gaps remain across different scripts and language families. (3) Persistent challenges in cross-lingual alignment and visual reasoning are not adequately captured by existing multimodal benchmarks. These results highlight the need for renewed research focus on robust multilingual chart understanding and provide a foundation for future progress.

**In summary, our main contributions are:**

- We propose a unified and reproducible pipeline for building high-quality, large-scale multilingual chart QA datasets, leveraging LLM-based translation and code-driven chart generation.
- We present POLYCHARTQA, the first benchmark to enable systematic evaluation of LVLMs on chart understanding in ten diverse languages.
- We perform extensive empirical studies on a variety of state-of-the-art models, offering new insights into current limitations and future directions in multilingual multimodal research.

## 2 Related Work

### 2.1 Chart Understanding Datasets

Chart understanding tasks challenge models to interpret both visual and textual information within charts and to provide accurate responses to a range of instructions. In recent years, several benchmarks have been introduced to systematically evaluate the capabilities of Large Vision-Language Models (LVLMs) across tasks such as chart question answering (Masry et al., 2022; Methani et al., 2020; Kantharaj et al., 2022a), chart summarization (Tang et al., 2023; Kantharaj et al., 2022b; Rahman et al., 2022), chart-to-table conversion (Xia et al., 2023, 2024; Chen et al., 2024a), and chart rerendering (Moured et al., 2024; Yang et al., 2024). Of these, chart question answering has emerged as a central metric for assessing a model's ability to perform fine-grained chart understanding.

Early datasets such as FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018) and PlotQA (Methani et al., 2020) featured synthetic charts and template questions, limiting their diversity and real-world applicability. More recent benchmarks, including ChartQA (Masry et al., 2022), ChartX (Xia et al., 2024), MMC (Liu et al.,

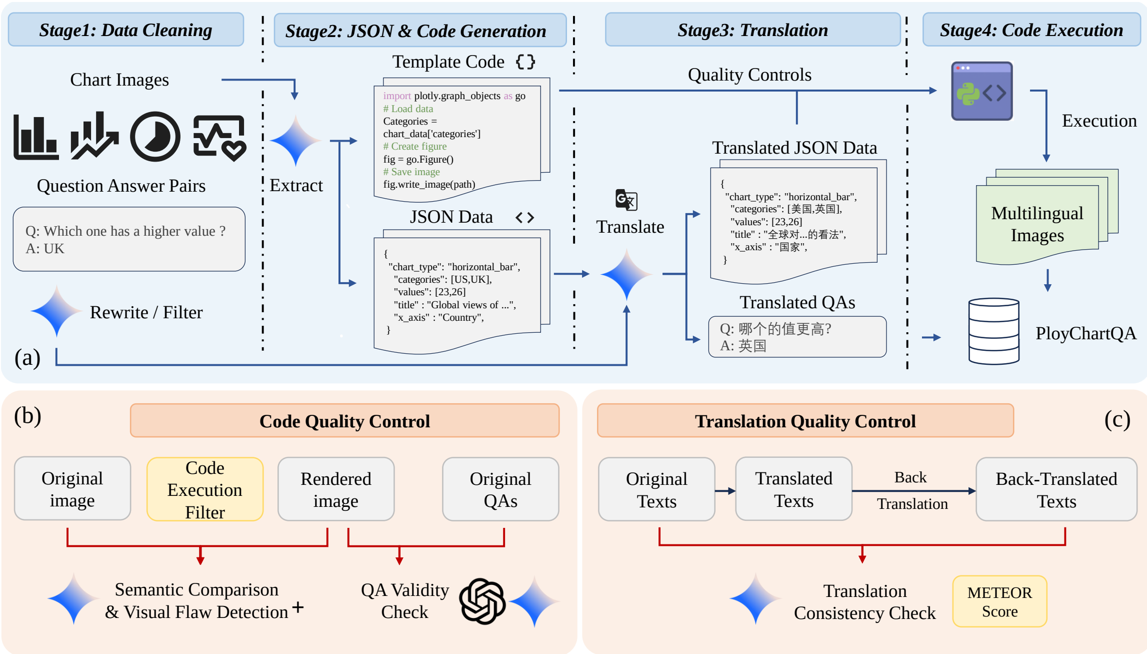

Figure 2: Overview of the POLYCHARTQA data pipeline with quality control.

2023), and ChartXiv (Wang et al., 2024c), incorporate real-world charts and human-authored questions, broadening the range of chart types and question complexities represented in the evaluation.

Despite recent advances, most chart datasets are English-only, with limited multilingual benchmarks (Chen et al., 2024a; Heakl et al., 2025). This lack of coverage prevents comprehensive evaluation of LVLMs on multilingual chart understanding and limits their real-world applicability.

## 2.2 Multilingual LVLMs

Building on the progress of foundational monolingual models (Li et al., 2023; Team et al., 2024a,b), researchers have developed a wide range of LVLMs with multilingual capabilities. Early influential works include PaLI (Chen et al., 2022), mBLIP (Geigle et al., 2023), and PaliGemma (Beyer et al., 2024; Steiner et al., 2024), which pioneered scalable multilingual vision-language alignment.

More recently, open-source models such as PALO (Maaz et al., 2024), Maya (Alam et al., 2024), Pangea (Yue et al., 2024), and Centurio (Geigle et al., 2025) have significantly broadened linguistic coverage and improved cross-lingual visual understanding. In parallel, open-source families like QwenVL (Bai et al., 2023, 2025; Wang et al., 2024b), InternVL (Chen et al., 2024b,c,d), and Phi-Vision (Abdin et al., 2024a,b) have demonstrated strong performance on multilingual multimodal tasks. Despite these advances, the ability of these models to process complex, text-rich visual inputs such as charts in multiple languages remains an underexplored challenge.

## 2.3 Multilingual Evaluations on LVLMs

The rapid development of multilingual LVLMs has driven the creation of a diverse set of benchmarks to assess their performance across multimodal tasks. These benchmarks typically cover general cross-lingual VQA (Pfeiffer et al., 2021; Changpinyo et al., 2022), text-centric VQA (Tang et al., 2024; Yu et al., 2025), and culturally diverse VQA (Romero et al., 2024; Liu et al., 2021; Vayani et al., 2024). In addition to task-specific resources, several comprehensive evaluation suites have been proposed to measure broader multilingual and multimodal capabilities of LVLMs. Benchmarks such as MMBench (Liu et al., 2024b), MMLU-Prox (Xuan et al., 2025), and M4U (Wang et al., 2024a) span a wide array of tasks including multimodal reasoning, open-domain chat, image captioning, and math problem solving. Similarly, M3Exam (Zhang et al., 2023) and Exams-V (Das et al., 2024) offer large-scale, real-world exam-style evaluations for LVLMs in multilingual and multimodal settings.

Despite these advances, most of the above benchmarks pay little attention to structured data representations like charts. Although datasets such as M3Exam (Zhang et al., 2023) and SMPQA (Geigle et al., 2025) contain chart relevent data, but still limited in both scale and task diversity, often emphasizing subtasks like OCR rather than holistic chart reasoning.

## 3 PolyChartQA

We present **PolyChartQA**, a large-scale multilingual chart question answering benchmark designed to address the lack of multilingual resources in this field. Figure 2 summarizes the data pipeline we adopt to construct PolyChartQA. It begins with the selection and refinement of high-quality English chart corpora and systematically expands to other languages via LLM-based translation and rigorous quality control. The following subsections detail each stage of this process: Section 3.1 describes monolingual data preparation; Section 3.2 covers multilingual expansion and evaluation; and Section 3.2 presents key dataset statistics of PolyChartQA benchmark.

### 3.1 Monolingual Corpora Construction

**English Dataset Selection** We began by surveying publicly available English-language chart question answering benchmarks, evaluating each candidate against three primary criteria: (*i*) overall data quality and chart diversity; (*ii*) the breadth and clarity of question types; and (*iii*) demonstrated adoption within the research community. Early large-scale datasets such as DVQA (Kafle et al., 2018) and PlotQA (Methani et al., 2020) rely heavily on rigid templates and synthetic chart designs, making them less suitable for a comprehensive benchmark dataset. In contrast, more recent corpora such as ChartQA-Pro (Masry et al., 2025) and CharXiv (Wang et al., 2024c) feature realistic, domain-rich visualizations, but present significant processing challenges due to their complexity. To achieve a balance between realism, diversity, and practical usability, we selected ChartQA (Masry et al., 2022) and ChartX (Xia et al., 2024) as the English seed corpora to construct our benchmark. To achieve a balance between realism, diversity, and practical usability, we selected ChartQA (Masry et al., 2022)[1] and ChartX (Xia et al., 2024)[2] as the English seed corpora to construct our benchmark.

**Data Cleaning and Validation** We applied a two-step quality control pipeline to both *ChartQA* and *ChartX*. First, each image–question–answer triplet was automatically checked with *Gemini 2.5 Pro-exp-0325* to verify whether the answer could be reliably inferred from the chart given the question; triplets flagged as unanswerable or inconsistent were removed, eliminating approximately 17–20% of the original data.

In the second stage, we manually normalized answers that were excessively long or verbose, standardizing them into concise formats—such as single words, numbers, or short phrases—while preserving their original semantics. This normalization step facilitates consistent multilingual translation and downstream evaluation.

To assess residual noise, we randomly sampled 10% of the cleaned dataset for human review and achieved a pass rate exceeding 98%, confirming the corpus's low noise and high consistency for downstream multilingual expansion.

**Seed Data Generation** A central innovation of our pipeline is the decoupling of chart content and visual rendering. For each cleaned chart, we prompted *Gemini 2.5 Pro* to generate two key artifacts: a structured JSON file that encodes the data table, chart type, text layout, color scheme, and other essential visual attributes, and an executable Python script that reconstructs the chart using Plotly. This decoupled design enables direct ingestion and regeneration of arbitrary chart images, enhances compatibility and reusability. Special attention was given to ensuring compatibility with multilingual rendering, including support for right-to-left (RTL) languages such as Arabic and other non-Latin scripts. Complex chart categories (e.g., box plots, bubble charts, multi-axis plots, etc.) were handled with four-shot prompting to guide generation. We selected *Plotly* as the main coding language over other packages because of better support for multilingual text rendering.

To ensure the integrity and usability of the seed dataset, we implemented a rigorous three-stage validation process. (i) **Code executability**: First, we verified that every generated Plotly script could be executed without errors to produce a chart image; any triplets that failed at this stage were manually inspected and either repaired or discarded if unfixable. (ii) **Visual fidelity**: Second, we assessed the visual fidelity of each regenerated chart by comparing it to the original, using *Gemini 2.5 Pro* to detect semantic and stylistic discrepancies. Charts with notable inconsistencies in type, values, or layout were removed. (iii) **QA validity**: Finally, we checked that all questions remained answerable on the reconstructed charts, using both *Gemini 2.5*

[1] ChartQA is released under license GPL-3.0.
[2] ChartX is released under license CC-BY-4.0.

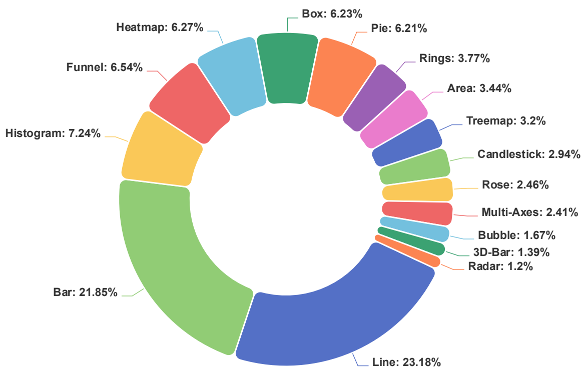


Figure 3: **Distribution of chart types in PolyChartQA.** The dataset covers sixteen chart types, including both common formats (e.g., bar, line, pie) and rarer forms (e.g., box, radar, rings, candlestick).

*Pro* and *GPT-4.1* as independent validators, and retained only those samples confirmed as correct by both models.

### 3.2 Translating the Corpora

**Text Data Translation** To enable robust multilingual chart question answering, we expand the high-quality English seed dataset to ten typologically diverse languages, following the PALO typology (Maaz et al., 2024). These languages includes: English, Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, and Japanese.

Standard machine translation services, which typically treat input as plain text, struggle to preserve the tightly coupled structure of chart JSON files and their associated QA pairs, often leading to inconsistencies in field names, terminology, and semantics. To address this, we adopt an LLM-based workflow using *Gemini 2.5 Pro*, which jointly translates the chart specification and all corresponding questions and answers within a unified, structured prompt. This approach ensures consistent terminology, high linguistic fluency, and cultural appropriateness across languages. In addition, our prompt template explicitly guides the model to retain key data attributes, preserve layout logic, and avoid semantic drift, markedly reducing the translation errors observed with conventional methods. As a result, the translated corpora maintain strong semantic and structural alignment with the original English data, enabling reliable cross-lingual evaluation and large-scale benchmarking of multilingual chart understanding.

**Multilingual Data Generation** The translated JSON files are then paired with the corresponding

Table 1: Average scores and exact match (EM) agreement between human and model evaluations on a English-Chinese subset ($N$=250). Scores range from 1 (worst) to 3 (best).

| Eval Dimension | Human Avg. | Model Avg. | EM (%) |
|---|---|---|---|
| Image Quality | 2.904 | 2.956 | 91.2 |
| QA Correctness | 2.948 | 2.984 | 94.4 |
| Translation Accuracy | 2.908 | 2.892 | 90.8 |

template code to generate chart images in each target language. As in the monolingual stage, we discard any samples for which the code fails to execute successfully. All remaining multilingual chart images are manually inspected to eliminate instances with significant visual defects, such as text clipping, layout shifts, or rendering errors. This rigorous process ensures that the final dataset maintains high visual quality and cross-lingual consistency.

**Translation Quality Control** We employ a rigorous, two-stage quality control process to guarantee the translation process produces high-quality translations, as illustrated in Figure 2 (c).

First, we conduct automated consistency checks for all translated samples. Each instance is back-translated into English and compared against its original counterparts. We elvauate the consisitency between these two parts by using METEOR scores and qualitatively using semantic judgments from *Gemini 2.5 Pro*. Only samples that meet stringent criteria—namely, a METEOR score above 0.6 and above 4 out of 5 score from positive alignment in Gemini's assessment—are retained. This step effectively eliminates translations with semantic drift or poor linguistic quality.

Second, for a after check to prove our automated checking method is useful, we conduct comprehensive human evaluation on the English-Chinese subset, assessing translation fidelity, visual accuracy, and QA consistency. Bilingual annotators jointly examine: (i) pairs of original and translated chart images for visual-semantic alignment; (ii) pairs of original and translated QA texts for linguistic accuracy and fluency; and (iii) the internal consistency between each translated chart and its QA pairs. Each dimension is rated on a three-point scale (with 3 indicating perfect and 1 indicating substantial errors), with disagreements adjudicated by a third annotator.

Given the practical difficulty of extending manual evaluations to all languages, we evaluated whether *Gemini 2.5 Pro* could reliably simulate human evaluation. We repeated the above evalua-

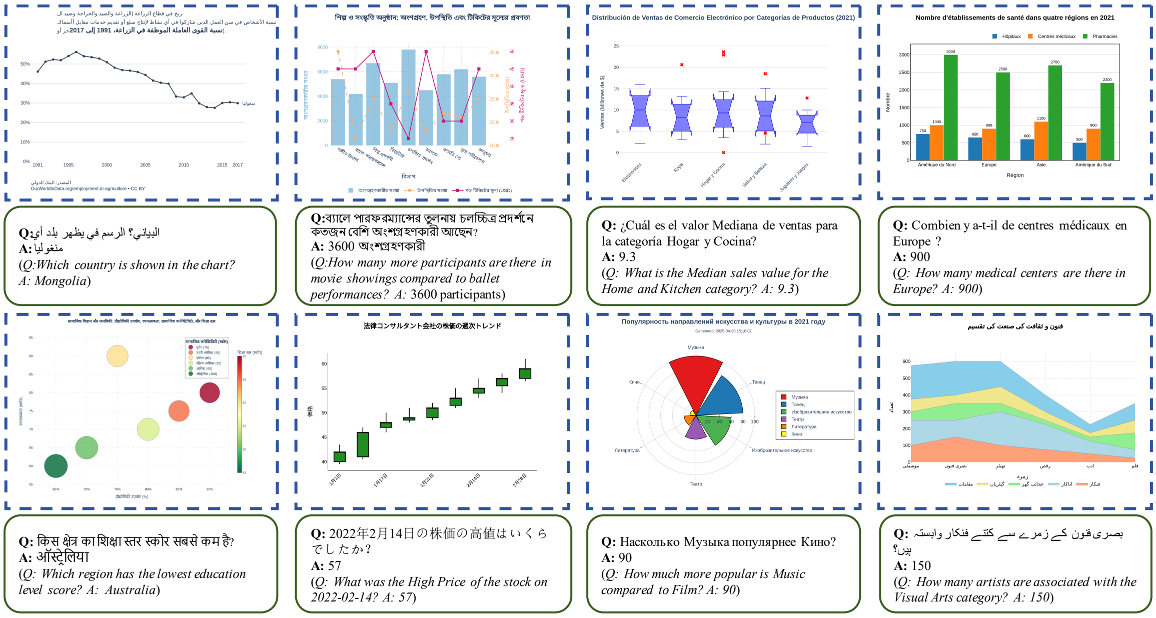


Figure 4: **Multilingual chart question answering visualizations selected from PolyChartQA.** First row, from left to right: Arabic, Bengali, Spanish, French. Second row, from left to right: Hindi, Japanese, Russian, Urdu.

tion on the same sampled Chinese-English subset using Gemini as a stand-in annotator, producing a comparable score distribution. As shown in Table 1, the distributions derived from human and Gemini evaluations exhibit strong alignment, validating Gemini's capability as an effective proxy for human annotation.

Consequently, we applied the Gemini-based evaluation approach to all remaining target languages and obtained consistently high translation quality scores across the multilingual corpus. The results indicate that our translations received high ratings in all aspects—including image texts, QA pairs, and general translation quality, with over 90% of cases achieving perfect scores. These findings demonstrate that our human evaluation pipeline confirms the high quality and semantic consistency of our translations, ensuring robust multilingual representations suitable for downstream QA tasks. Detailed results of human evaluation scores can be found at Appendix B

**Data Statistics** PolyChartQA comprises 22,606 chart images and 26,151 corresponding question–answer pairs across 10 languages. As a test-only benchmark, it encompasses 16 distinct chart types with a balanced distribution, as shown in Figure 3. Representative examples from the dataset are presented in Figure 4. Appendix A provides a detailed analysis of PolyChartQA, covering per-language instance counts, question and answer lengths, and their distributions.

# 4 Experiments

## 4.1 Baseline Models

**Evaluated Multilingual Vision-Language Models** To thoroughly assess the multilingual perception and reasoning abilities of modern LVLMs on our multilingual chart benchmark, we select representative state-of-the-art models from three categories: open-source general MLLMs, open-source multilingual LVLMs, and closed-source LVLMs.

The general open-source LVLMs include Qwen2-VL (Bai et al., 2023), InternVL 2.5 (Chen et al., 2024b), 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-1.6 (Liu et al., 2024a), 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 closed-source category comprises Gemini-2.5-Pro-03-25 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.

## 4.2 Implementation Details

**Metrics** Following (Methani et al., 2020) and (Masry et al., 2022), we use a relaxed accuracy

Table 2: Model Average Relaxed Accuracy (%) by Language and Overall. Best score per column is in **bold**, second best is underlined. Models are grouped, with open source models further sub-grouped by approximate size, and then sorted by average performance (w EN).

| Model | EN | ZH | FR | ES | RU | JA | AR | UR | HI | BN | Avg. (w EN) | Avg. (w/o EN) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Closed Source Models*** | | | | | | | | | | | | |
| GPT-4o | 0.559 | 0.460 | 0.534 | 0.544 | 0.524 | 0.454 | 0.505 | 0.487 | 0.513 | 0.482 | 0.509 | 0.502 |
| Gemini-2.5-Pro | **0.706** | **0.677** | **0.690** | **0.693** | **0.676** | **0.686** | **0.691** | **0.675** | **0.686** | **0.660** | **0.685** | **0.682** |
| ***Open Source Models*** | | | | | | | | | | | | |
| PaliGemma2-3B | 0.266 | 0.147 | 0.197 | 0.215 | 0.139 | 0.107 | 0.159 | 0.122 | 0.143 | 0.102 | 0.163 | 0.149 |
| Phi-3 Vision | 0.451 | 0.175 | 0.372 | 0.369 | 0.269 | 0.157 | 0.093 | 0.047 | 0.106 | 0.106 | 0.232 | 0.202 |
| InternVL-2.5-2B | 0.278 | 0.033 | 0.147 | 0.092 | 0.095 | 0.020 | 0.043 | 0.003 | 0.012 | 0.001 | 0.078 | 0.051 |
| InternVL-3-2B | 0.437 | 0.353 | 0.308 | 0.335 | 0.256 | 0.269 | 0.171 | 0.146 | 0.157 | 0.119 | 0.256 | 0.231 |
| Qwen2-VL-2B | 0.423 | 0.336 | 0.376 | 0.377 | 0.359 | 0.222 | 0.288 | 0.191 | 0.244 | 0.230 | 0.307 | 0.291 |
| Qwen2.5-VL-3B | **0.674** | **0.596** | **0.618** | **0.625** | **0.580** | **0.488** | **0.514** | **0.372** | **0.457** | **0.430** | **0.537** | **0.518** |
| LLaVA-OneVision-7B | 0.187 | 0.101 | 0.131 | 0.142 | 0.094 | 0.083 | 0.075 | 0.052 | 0.071 | 0.057 | 0.101 | 0.090 |
| LLaVA-v1.6-7B-Vicuna | 0.275 | 0.055 | 0.186 | 0.171 | 0.149 | 0.034 | 0.129 | 0.066 | 0.093 | 0.053 | 0.126 | 0.106 |
| LLaVA-v1.6-7B-Mistral | 0.248 | 0.129 | 0.189 | 0.182 | 0.135 | 0.115 | 0.120 | 0.077 | 0.100 | 0.067 | 0.139 | 0.124 |
| Llama-3.2-11B-Vision | 0.155 | 0.169 | 0.141 | 0.129 | 0.154 | 0.096 | 0.131 | 0.144 | 0.213 | 0.175 | 0.152 | 0.152 |
| DeepSeek-VL2 | 0.401 | 0.388 | 0.264 | 0.341 | 0.199 | 0.000 | 0.142 | 0.138 | 0.191 | 0.163 | 0.248 | 0.225 |
| InternVL-2.5-8B | 0.392 | 0.263 | 0.324 | 0.335 | 0.295 | 0.226 | 0.109 | 0.112 | 0.140 | 0.134 | 0.235 | 0.214 |
| InternVL-3-8B | 0.541 | 0.394 | 0.434 | 0.458 | 0.381 | 0.397 | 0.214 | 0.172 | 0.202 | 0.175 | 0.338 | 0.310 |
| Phi-4 Vision | **0.623** | 0.460 | 0.559 | 0.446 | 0.487 | 0.416 | 0.297 | 0.234 | 0.334 | 0.183 | 0.406 | 0.377 |
| Qwen2-VL-7B | 0.564 | 0.543 | 0.534 | 0.527 | 0.522 | 0.473 | 0.405 | 0.320 | 0.439 | 0.403 | 0.473 | 0.461 |
| Qwen2.5-VL-7B | 0.605 | **0.583** | **0.572** | **0.590** | **0.568** | **0.556** | **0.520** | **0.437** | **0.494** | **0.464** | **0.538** | **0.530** |
| ***Multilingual Models*** | | | | | | | | | | | | |
| Pangea-7B | **0.247** | **0.136** | **0.198** | **0.213** | **0.158** | **0.115** | **0.131** | **0.121** | **0.131** | **0.131** | **0.161** | **0.149** |
| PALO-7B | 0.115 | 0.060 | 0.105 | 0.099 | 0.070 | 0.059 | 0.070 | 0.050 | 0.052 | 0.036 | 0.073 | 0.067 |
| Maya | 0.087 | 0.064 | 0.076 | 0.072 | 0.068 | 0.060 | 0.071 | 0.057 | 0.069 | 0.056 | 0.068 | 0.066 |
| Centurio-Qwen | 0.079 | 0.040 | 0.036 | 0.030 | 0.015 | 0.025 | 0.020 | 0.015 | 0.015 | 0.010 | 0.029 | 0.022 |

measure for the numeric answers to allow a minor inaccuracy that may result from the automatic data extraction process. We consider an answer to be correct if it is within 5% of the gold answer. For non-numeric answers, we still need an exact match to consider an answer to be correct.

**Evaluation Procedure** We evaluate all baseline LVLMs using their default configurations. For all models, we set the decoding temperature to 0.01 to ensure deterministic outputs. To enhance evaluation efficiency, the following concise prompt template is adopted uniformly: `"Answer the question using a word or phrase in <target_language> or a number in digits. <Question>"`, where `<Question>` is replaced with the actual test question from PolyChartQA. All results are from a signle run. All evaluations are conducted on NVIDIA A100 40G GPUs.

Table 3: Few-shot inference performance (Relaxed Accuracy %) showing English scores and overall averages.

| Model | Shots | EN | Avg. (w EN) | Avg. (w/o EN) |
|---|---|---|---|---|
| Qwen2.5-VL-3B | 0 | **0.674** | **0.537** | **0.518** |
| | 2 | 0.592 | 0.505 | 0.493 |
| | 4 | 0.635 | 0.507 | 0.490 |
| | 8 | 0.638 | 0.515 | 0.499 |
| Qwen2.5-VL-7B | 0 | 0.605 | 0.538 | 0.530 |
| | 2 | 0.625 | 0.538 | 0.526 |
| | 4 | **0.663** | 0.559 | 0.545 |
| | 8 | 0.653 | **0.560** | **0.548** |
| Qwen2-VL-2B | 0 | **0.423** | 0.307 | 0.291 |
| | 2 | 0.403 | 0.315 | 0.304 |
| | 4 | 0.421 | **0.321** | **0.307** |
| | 8 | 0.418 | 0.318 | 0.305 |
| Qwen2-VL-7B | 0 | 0.564 | **0.473** | **0.461** |
| | 2 | 0.517 | 0.448 | 0.439 |
| | 4 | **0.570** | 0.471 | 0.457 |
| | 8 | 0.538 | 0.468 | 0.459 |

### 4.3 Evaluation Results

**Zero-shot Evaluation** Table 2 presents zero-shot relaxed accuracy for a range of multilingual LVLMs on PolyChartQA. There is a clear and substantial performance gap between closed-source and open-source models. `Gemini-2.5-Pro` achieves the best results across all languages, with average accuracy reaching 0.685 (w EN) and 0.682 (w/o EN). In contrast, `GPT-4o` lags significantly behind, with an average accuracy of only 0.509.

`Qwen2.5-VL-Series` is the top performer among open-source models, consistently outperforming its peers in both high- and low-resource languages. Notably, its 7B and 3B variants achieve leading scores, especially on challenging languages like Urdu and Bengali, highlighting strong multilingual generalization. By comparison, other open models—including both smaller (e.g., `PaliGemma2-3B`) and similarly sized (e.g., `Qwen2-VL-7B`) competitors—show clear deficits, often struggling on non-English data. These differences can be traced to insufficient multilingual chart-specific training data or a lack of targeted adaptation for chart reasoning.

General-purpose multilingual models such as `Pangea-7B` show limited effectiveness on chart QA, and the rest of this category perform even worse. This demonstrates that broad multilingual training alone does not equip models for structured visual reasoning.

In terms of language coverage, English consistently yields the highest accuracy, often serving as the benchmark for model performance. However, the gap between English and other languages, especially low-resource ones, is evident. For instance, models that perform well on English can see their accuracy drop below 0.1 for languages like Bengali and Urdu. This significant disparity highlights challenges in cross-lingual robustness, particularly in underrepresented scripts. High-resource languages tend to show better performance, yet there remains a noticeable gap in performance for low-resource languages, which could be indicative of the model's inability to generalize well across linguistic and script variations.

**Few-shot Evaluation** Table 3 presents few-shot relaxed accuracy of four Qwen-based LVLMs under 0, 2, 4, and 8-shot settings. We observe that few-shot prompting does not reliably improve multilingual performance. Neither English nor non-English scores show a consistent upward trend as the number of shots increases. In particular, smaller models such as `Qwen2.5-VL-3B` and `Qwen2-VL-2B` exhibit fluctuations or even regressions in both English and overall average scores. Only `Qwen2.5-VL-7B`, the largest model in this comparison, shows modest gains with more shots. These results suggest that few-shot prompting alone is insufficient to address the multilingual transfer gap in current LVLMs and may require stronger model capacity and task adaptation strategies.

**Cross-lingual Evaluation** Table 4 compares three inference settings to evaluate cross-lingual transfer: (1) using English-language charts with translated questions (*Img*), (2) using translated charts with English questions (*QA*), and (3) fully localized inputs with both charts and questions in the target language (*Native*). We find that the native setting consistently achieves the highest accuracy across all models, highlighting the inherent difficulty of cross-lingual transfer. Among the cross-lingual setups, using English questions with localized charts (QA) outperforms the reverse (Img), suggesting that language consistency on the question side is more critical than on the visual side. These findings indicate that current LVLMs struggle to generalize across linguistic mismatches in visual content, underscoring the importance of maintaining consistent language modalities for cross-lingual understanding.

Table 4: Cross-lingual inference results (Relaxed Accuracy %) with English source contexts.

| Model | Source | Avg. (w/EN) | Avg. (w/o EN) |
|---|---|---|---|
| Qwen2.5-VL-3B | Img | 0.496 | 0.473 |
| | QA | 0.521 | 0.499 |
| | Native | **0.537** | **0.518** |
| Qwen2.5-VL-7B | Img | 0.483 | 0.466 |
| | QA | 0.510 | 0.495 |
| | Native | **0.538** | **0.530** |
| Qwen2-VL-2B | Img | 0.275 | 0.255 |
| | QA | **0.309** | **0.292** |
| | Native | 0.307 | 0.291 |
| Qwen2-VL-7B | Img | 0.412 | 0.391 |
| | QA | 0.430 | 0.409 |
| | Native | **0.473** | **0.461** |

## 5 Conclusion

In this paper, we introduce PolyChartQA, a multilingual chart question-answering dataset comprising 22,606 charts and 26,151 QA pairs across 10 diverse languages. To construct the dataset, we develop a scalable data pipeline that decouples chart data from rendering code, enabling efficient multilingual chart generation. Evaluation results on PolyChartQA reveal that current models face significant challenges in multilingual chart understanding, especially for languages with non-Latin scripts. We hope this work draws greater attention to the multilingual capabilities of LVLMs and serves as a foundation for developing more language-inclusive and globally accessible models.

## Limitations

Despite introducing the first large-scale multilingual benchmark for chart question answering, PolyChartQA has several limitations. First, PolyChartQA currently covers only ten major world languages. Though they cover over 65% of the global population, there are still many low-source languages not covered. Since our data pipeline decouples chart data from template codes, it can be flexibly extended to more typologically diverse and low-resource languages. Second, our benchmark covers only the question-answering task over multilingual charts, future work would focus on more diverse chart understanding tasks such as chart summary generation, chart fact-checking, and future trend analysis under multilingual scenarios.

## 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. While our dataset relies partly on machine translation, we take care to ensure quality through post-filtering and design practices. We encourage future research to further improve multilingual 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, S M Iftekhar Uddin, Shayekh Bin Islam, Roshan Santhosh, Snegha A, Drishti Sharma, Chen Liu, Isha Chaturvedi, Genta Indra Winata, Ashvanth. S, Snehanshu Mukherjee, and Alham Fikri Aji. 2024. Maya: An instruction finetuned multilingual multimodal model. *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*.

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.

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. 2024b. 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. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024d. 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.

Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

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*.

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*.

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 Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022a. Opencqa: Open-ended question answering with charts. *arXiv preprint arXiv:2210.06628*.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022b. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

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. *arXiv preprint arXiv:2109.13238*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: 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. 2024b. 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, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. *Preprint*, arXiv:2504.05506.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

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 O Steitz, Stefan Roth, Ivan Vulić, and Iryna

Gurevych. 2021. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*.

Raian Rahman, Rizvi Hasan, and Abdullah Al Farhad. 2022. *ChartSumm: A large scale benchmark for Chart to Text Summarization*. Ph.D. thesis, Department of Computer Science and Engineering (CSE), Islamic University of ....

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and 1 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

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 J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.

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 Kuckreja, and 1 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

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.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *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. 2024. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*.

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*.

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 Detailed Dataset Statistics

**A.1. Image and QA Pair Distribution by Language and Chart Type** We show the detailed statistics of POLYCHARTQA in Tables 5 and 6, including per-language and per-chart-type breakdowns for both images and QA pairs.

**A.2. 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 training and test splits, is illustrated in Figure 5. These results highlight significant variation in textual length, which reflects both linguistic and orthographic diversity across languages.

**A.3. Distribution of Images and Questions by Language** We further examine the distribution of images and questions in each language. Figure 6 presents a t-SNE visualization of CLIP image embeddings, while Figure 7 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.

**A.4. Distribution of Images, Questions, JSON, and Code for English Data** We also provide a detailed analysis of the English subset, which serves as the seed data for POLYCHARTQA. Figure 8 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 9 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.

# B Human Evaluation Details

**B.1. Human Annotators and Annotation Process** We conducted a rigorous human evaluation to measure the quality of multilingual chart question-answering pairs in POLYCHARTQA. Figure 10 shows 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. Annotators were instructed to assess each dimension—image quality, QA correctness, and translation accuracy—according to our predefined scoring guidelines, ensuring consistent and reliable evaluations across all samples.

**B.2. Full Analysis Results of Human and Proxy Model** To support transparency and reproducibility, we provide comprehensive evaluation results from both human annotators and the Gemini proxy model. Table 7 presents a detailed breakdown of human annotation scores for a random sample of Chinese-English chart QA pairs. Table 8 further summarizes the model-based evaluation scores across all target languages in POLYCHARTQA, highlighting the effectiveness of the proxy model in simulating human assessment quality.

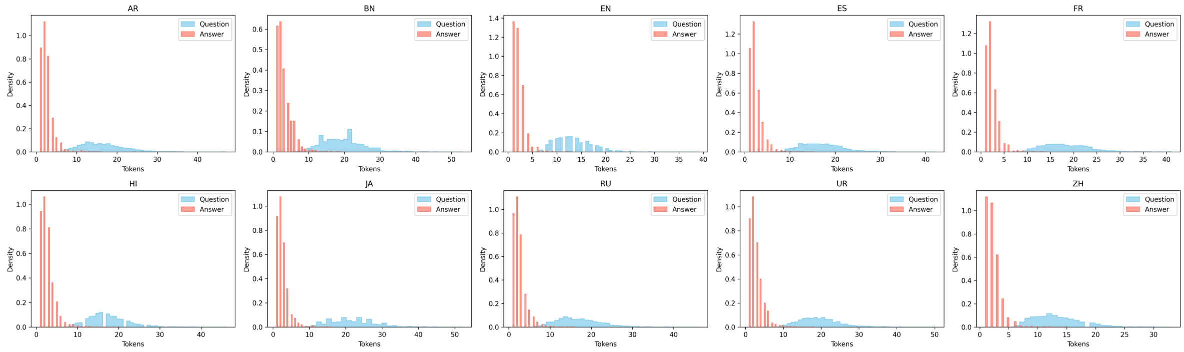


Figure 5: **Question and answer length statistics in PolyChartQA.** Token counts are calculated using the GPT-4o tokenizer and aggregated over all splits. The results reveal considerable variation in length distributions across languages, indicative of linguistic and orthographic differences.

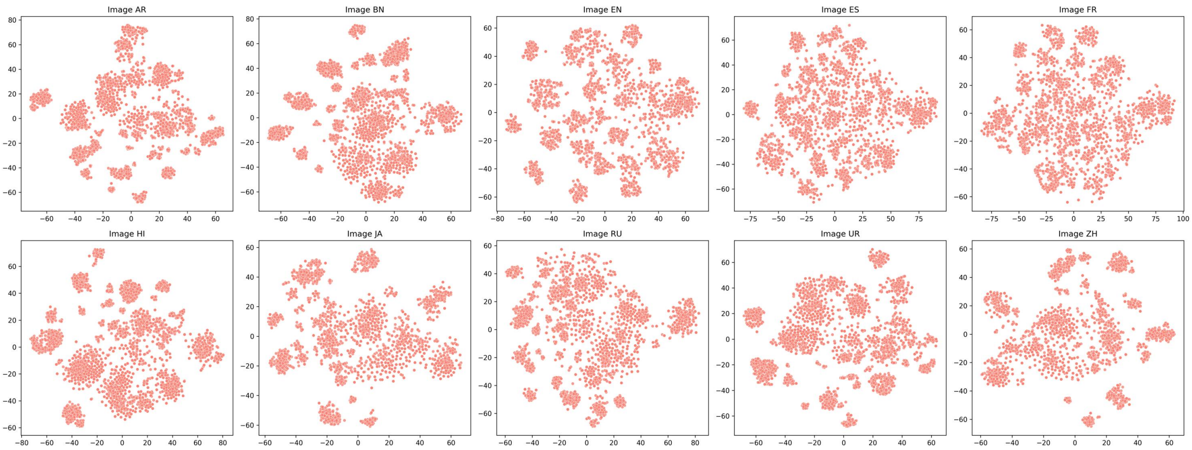


Figure 6: **Distribution of images in PolyChartQA by language.** t-SNE visualization of CLIP image embeddings for each language. Each subplot depicts the distribution of visual features for one language, with all points colored uniformly (skyblue) to emphasize the general embedding spread within each language.

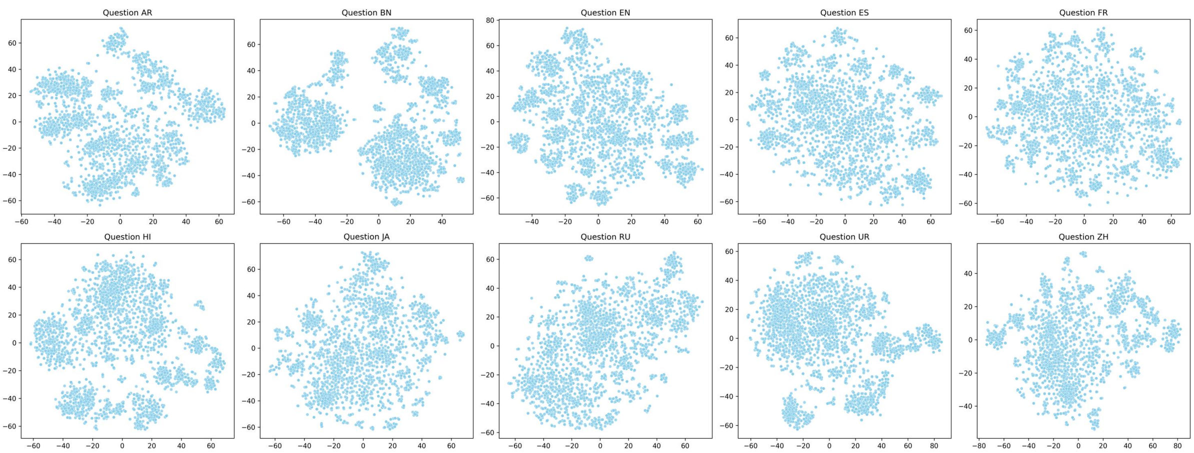


Figure 7: **Distribution of questions in PolyChartQA by language.** t-SNE visualization of CLIP question text embeddings for each language. Subplots display the clustering patterns of question semantics, with points colored uniformly (salmon) to highlight language-specific distributions.

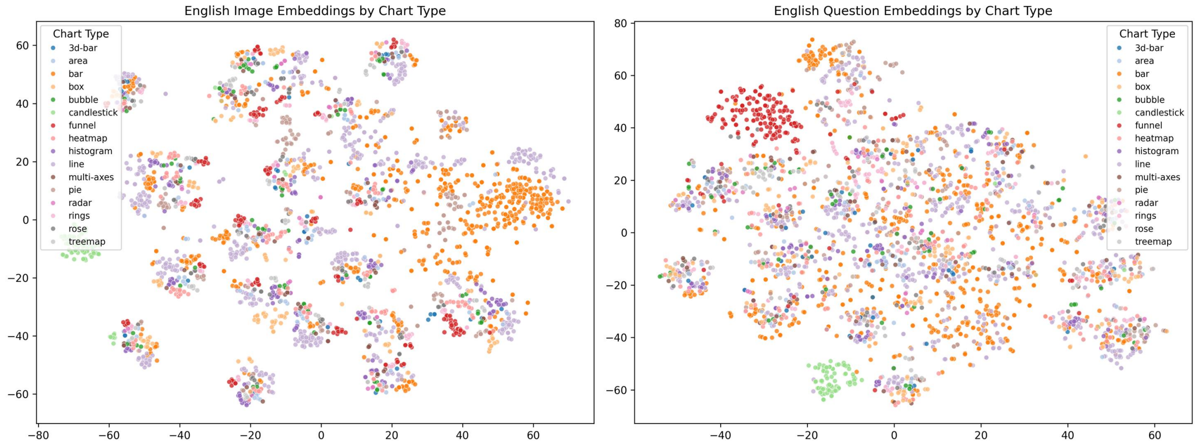


Figure 8: **Distribution of images and questions in English by chart type in POLYCHARTQA.** t-SNE visualizations of CLIP embeddings for the English subset, with points colored by chart type. (a) Image embeddings show clustering by chart visual characteristics; (b) Question embeddings reveal semantic groupings by chart type.

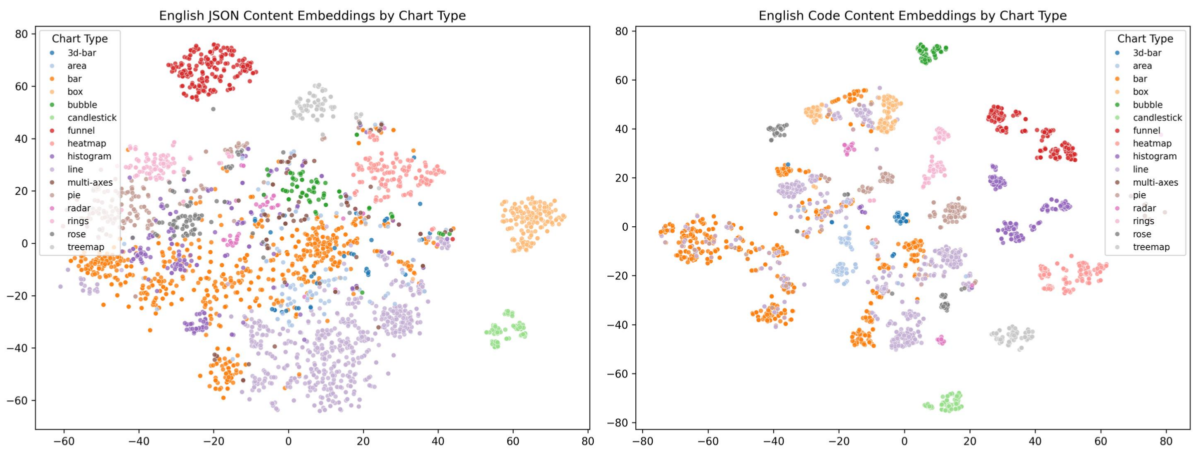


Figure 9: **Distribution of JSON data and code in English by chart type in POLYCHARTQA.** t-SNE visualizations of CLIP embeddings from (a) the underlying JSON data and (b) the Python code generating each chart, colored by chart type. These visualizations examine the structural and logical distinctions among chart types in the dataset.

Table 5: Detailed statistics of Image counts per language. Total image counts and specific chart type counts are shown. Note: CS=candlestick, hist.=histogram, MA=multi-axes, TM=treemap.

| Language | Total | 3d-bar | area | bar | box | bubble | CS | funnel | heatmap |
|---|---|---|---|---|---|---|---|---|---|
| English (en) | 2917 | 40 | 106 | 600 | 171 | 81 | 86 | 211 | 183 |
| Arabic (ar) | 2139 | 31 | 79 | 447 | 144 | 32 | 62 | 148 | 133 |
| Bengali (bn) | 2297 | 27 | 76 | 507 | 155 | 39 | 67 | 155 | 149 |
| Spanish (es) | 2379 | 35 | 84 | 505 | 148 | 38 | 74 | 158 | 149 |
| French (fr) | 2304 | 35 | 78 | 471 | 144 | 38 | 62 | 154 | 153 |
| Hindi (hi) | 2452 | 30 | 86 | 547 | 153 | 40 | 70 | 165 | 160 |
| Japanese (ja) | 1893 | 26 | 61 | 409 | 131 | 33 | 50 | 121 | 120 |
| Russian (ru) | 2124 | 30 | 65 | 477 | 132 | 35 | 56 | 142 | 134 |
| Urdu (ur) | 2284 | 30 | 68 | 514 | 153 | 37 | 61 | 137 | 153 |
| Chinese (zh) | 1817 | 26 | 63 | 393 | 134 | 35 | 56 | 117 | 125 |

| Language | Total | hist. | line | MA | pie | radar | rings | rose | TM |
|---|---|---|---|---|---|---|---|---|---|
| English (en) | 2917 | 219 | 600 | 77 | 190 | 42 | 123 | 84 | 104 |
| Arabic (ar) | 2139 | 167 | 491 | 49 | 133 | 23 | 80 | 46 | 74 |
| Bengali (bn) | 2297 | 177 | 500 | 53 | 148 | 25 | 83 | 58 | 78 |
| Spanish (es) | 2379 | 180 | 551 | 52 | 150 | 26 | 91 | 53 | 85 |
| French (fr) | 2304 | 187 | 521 | 58 | 148 | 24 | 95 | 61 | 75 |
| Hindi (hi) | 2452 | 182 | 539 | 55 | 162 | 29 | 92 | 64 | 78 |
| Japanese (ja) | 1893 | 141 | 436 | 42 | 120 | 27 | 72 | 36 | 68 |
| Russian (ru) | 2124 | 162 | 516 | 48 | 130 | 24 | 66 | 44 | 63 |
| Urdu (ur) | 2284 | 181 | 509 | 58 | 146 | 26 | 85 | 54 | 72 |
| Chinese (zh) | 1817 | 137 | 402 | 45 | 93 | 21 | 76 | 34 | 60 |

Table 6: Detailed statistics of Question-Answer (QA) pair counts per language. Total QA counts and specific chart type counts are shown. Note: CS=candlestick, hist.=histogram, MA=multi-axes, TM=treemap.

| Language | Total | 3d-bar | area | bar | box | bubble | CS | funnel | heatmap |
|---|---|---|---|---|---|---|---|---|---|
| English (en) | 3080 | 40 | 107 | 696 | 171 | 81 | 86 | 211 | 183 |
| Arabic (ar) | 2496 | 31 | 80 | 592 | 144 | 32 | 62 | 148 | 133 |
| Bengali (bn) | 2695 | 27 | 77 | 670 | 155 | 39 | 67 | 155 | 149 |
| Spanish (es) | 2802 | 35 | 85 | 669 | 148 | 38 | 74 | 158 | 149 |
| French (fr) | 2694 | 35 | 79 | 627 | 144 | 38 | 62 | 154 | 153 |
| Hindi (hi) | 2886 | 30 | 87 | 733 | 153 | 40 | 70 | 165 | 160 |
| Japanese (ja) | 2195 | 26 | 62 | 535 | 131 | 33 | 50 | 121 | 120 |
| Russian (ru) | 2519 | 30 | 66 | 638 | 132 | 35 | 56 | 142 | 134 |
| Urdu (ur) | 2680 | 30 | 69 | 685 | 153 | 37 | 61 | 137 | 153 |
| Chinese (zh) | 2104 | 26 | 64 | 517 | 134 | 35 | 56 | 117 | 125 |

| Language | Total | hist. | line | MA | pie | radar | rings | rose | TM |
|---|---|---|---|---|---|---|---|---|---|
| English (en) | 3080 | 219 | 646 | 77 | 210 | 42 | 123 | 84 | 104 |
| Arabic (ar) | 2496 | 167 | 689 | 49 | 146 | 23 | 80 | 46 | 74 |
| Bengali (bn) | 2695 | 177 | 718 | 53 | 164 | 25 | 83 | 58 | 78 |
| Spanish (es) | 2802 | 180 | 794 | 52 | 165 | 26 | 91 | 53 | 85 |
| French (fr) | 2694 | 187 | 739 | 58 | 163 | 24 | 95 | 61 | 75 |
| Hindi (hi) | 2886 | 182 | 770 | 55 | 178 | 29 | 92 | 64 | 78 |
| Japanese (ja) | 2195 | 141 | 602 | 42 | 129 | 27 | 72 | 36 | 68 |
| Russian (ru) | 2519 | 162 | 734 | 48 | 145 | 24 | 66 | 44 | 63 |
| Urdu (ur) | 2680 | 181 | 720 | 58 | 159 | 26 | 85 | 54 | 72 |
| Chinese (zh) | 2104 | 137 | 551 | 45 | 106 | 21 | 76 | 34 | 60 |

Table 7: Detailed score distribution (%) and agreement metrics between human and model annotations on Chinese data ($N$=250). Scores range from 1 (worst) to 3 (best). Average score is calculated out of a maximum of 3.

| **Evaluation Dimension** | **Human Scores (%)** | | | **Avg** | **Model Scores (%)** | | | **Avg** | **Exact** |
|---|---|---|---|---|---|---|---|---|---|
| | 3 | 2 | 1 | | 3 | 2 | 1 | | **Agr. (%)** |
| Image Quality | 96.8 | 2.0 | 1.2 | 2.956 | 96.8 | 2.0 | 1.2 | 2.956 | 91.2 |
| QA Correctness | 98.8 | 0.8 | 0.4 | 2.984 | 98.8 | 0.8 | 0.4 | 2.984 | 94.4 |
| Translation Accuracy | 94.4 | 0.4 | 5.2 | 2.892 | 94.4 | 0.4 | 5.2 | 2.892 | 90.8 |

Table 8: Model-based evaluation scores for Image Quality, QA Correctness, and Translation Accuracy across all languages in PolyChartQA (N=250 per language). Each cell reports the average score (max = 3).

| **Quality Dimension** | **ZH** | **JA** | **UR** | **AR** | **FR** | **ES** | **RU** | **HI** | **BN** |
|---|---|---|---|---|---|---|---|---|---|
| Image Quality Avg. | 2.956 | 2.976 | 2.752 | 2.852 | 2.892 | 2.876 | 2.828 | 2.936 | 2.936 |
| QA Correctness Avg. | 2.984 | 2.980 | 2.948 | 2.894 | 2.972 | 2.960 | 2.924 | 2.948 | 2.936 |
| Translation Accuracy Avg. | 2.892 | 2.936 | 2.876 | 2.916 | 2.872 | 2.944 | 2.888 | 2.884 | 2.884 |


Figure 10: **Human evaluation interface.** Annotators review chart images and QA pairs in both source and target languages, providing quality ratings for translation fidelity, visual accuracy, and QA consistency.

## C Full Prompt Templates Used in Our Study

In this appendix, we present all prompt templates used throughout our POLYCHARTQA data pipeline and evaluation process, to support transparency and reproducibility. Section C.1 details the pipeline prompts for data cleaning, generation, translation, and consistency checking. Section C.2 provides all human and automatic evaluation prompts.

### C.1 Data Pipeline Prompt Templates

Here we presents the prompt of our data pipeline used to construct PLOYCHARTQA.

**Stage 1 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"
}
```

If action is DELETE, new_label must be ""; if KEPT, new_label is identical to the original label; if MODIFIED, new_label is your concise rewrite. Now, process the following input:

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

## Stage 1 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/question; core claim is wrong according to the chart.
- **1: Very Poor / Completely Incorrect or Irrelevant**
  Entirely false or irrelevant to the chart/question; no connection between 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 value ..."
}
```

**Example Output (Score 3):**

```
{
  "rating": 3,
  "reason": "The answer correctly identifies 'Product A' as having the highest value, but incorrectly states..."
}
```

**Example Output (Score 1):**

```
{
  "rating": 1,
  "reason": "The answer discusses stock market trends, which are completely absent from the provided..."
}
```

Now, evaluate the specific image, question, and answer provided in the user prompt based on the 1-5 scale. **Respond ONLY with the JSON object.** 992

## Stage 1 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:** Your primary task is to assess the provided `label` in the context of the `query`. You MUST base the `new_label` strictly on the 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 (Determine if DELETE is needed):**
   If the `query` requires an answer that cannot be concise (trend, explanation, subjective, complex comparison), set `action: "DELETE"`, `new_label: ""`, and stop.

2. **Assess Original Label Conciseness (Determine if KEPT is needed):**
   If the original `label` is already concise (single number, name, yes/no, short list, "Unanswerable"), set `action: "KEPT"`, `new_label: label` (exact copy), and stop.

3. **Perform Modification (Only if Verbose & Suitable Query):**
   If the query is suitable and 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"
}
```

If action is DELETE, new_label must be ""; if action is KEPT, new_label is identical to the original label; if MODIFIED, new_label is your concise rewrite.
**Examples:**

```
Input: {"query": "Highest ratio?",
        "label": "Mental Health Support and Healthcare Accessibility are the highest, both at 25%"}
Output: {"action": "MODIFIED",
        "new_label": "Mental Health Support, Healthcare Accessibility"}

Input: {"query": "Difference Twitter vs Facebook?",
        "label": "340 - 1.85 = 338.15 million more daily active users"}
Output: {"action": "MODIFIED",
        "new_label": "338.15 million"}

Input: {"query": "Trend for interest rates?",
        "label": "They fluctuated..."}
Output: {"action": "DELETE",
        "new_label": ""}

Input: {"query": "What is the value for Q2?",
        "label": "45.6"}
Output: {"action": "KEPT",
        "new_label": "45.6"}

Input: {"query": "Lowest retailers month?",
        "label": "The dataset provided does not include..."}
Output: {"action": "MODIFIED",
        "new_label": "Unanswerable"}
```

Now, process the following input:

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

## Stage 2 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 (<br>). 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.

## Stage 2 Prompt for Visual Consistency Check and Visual Flaw Detection

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 if the Rendered Image accurately 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 Title, 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 for data series maintain the same *hue category* (e.g., reds stay red/orange, blues stay blue/cyan, greens stay green)? A swap from red to blue 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 checks 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 data inaccuracies, text variations, or color shade differences (hue preserved), but interpretation unchanged.
- 3: **Moderately Consistent**: Some aspects captured, but noticeable discrepancies. Key values may be inaccurate, important text differs, color hues mismatched, or message partially distorted.
- 2: **Poorly Consistent**: Significant data errors, trends misrepresented, text is wrong/misleading, or color usage creates confusion. Fundamentally different interpretation.
- 1: **Inconsistent / Unrelated**: Completely different data, topic, or structure.

**Task 2: Identify Visual Flaws (Yes/No)**
Determine if 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 overlapping illegibly.
- **Element Clipping**: Chart elements (data, labels, legends) cut off by boundaries.
- **Unreadable Text**: Text is too small, blurry, or has unsupported characters.
- **Data Obscurity**: Data points hidden behind other elements.
- **Empty/Malformed Chart**: Blank image, error messages, or not a meaningful chart.
- **Gross Layout Issues**: Elements positioned bizarrely, chart is nonsensical.

Answer **Yes** if any major flaws; **No** if not. Minor imperfections that do not hinder core interpretation = **No**.
**Output Format:**
Respond ONLY with a valid JSON object containing FOUR keys:

- `"similarity_rating"`: integer (1–5 based on Task 1)
- `"similarity_reason"`: string (brief explanation for the similarity rating)
- `"has_visual_flaws"`: boolean (true if significant flaws found, false otherwise)
- `"flaw_reason"`: string (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.

995

## Stage 2 Prompt for QA Validity Check

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 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 make unsupported inferences; partially or inaccurately addresses the question.
- **2: Poor / Mostly Incorrect:**
  Contains significant factual errors; fundamentally misunderstands chart or question; core claim is wrong based on chart evidence.
- **1: Very Poor / Completely Incorrect or Irrelevant:**
  Completely false or irrelevant; no connection between 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 to 5)
- `"reason"`: string (brief explanation for your 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 value ..."
}
```

**Example Output (Score 3):**

```
{
  "rating": 3,
 "reason": "The answer correctly identifies 'Product A' as having the highest value, but incorrectly states..."
}
```

**Example Output (Score 1):**

```
{
  "rating": 1,
 "reason": "The answer discusses stock market trends, which are completely absent from the provided..."
}
```

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

## Stage 3 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 & 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 dicts/lists).
   - **ONLY translate string values** meant for user display in `{source_language_name}` (titles, axis labels, legend entries, annotations, etc.).
   - **DO NOT translate/modify:**
     - JSON keys
     - Numerical values (int/float)
     - Strings only of numbers (e.g., "2023", "1.5")
     - Strings only of numbers with "%" (e.g., "55.5%", "-10%")
     - Hex color codes (e.g., "#1f77b4")
     - URLs, file paths, system identifiers
     - Boolean strings ("true", "false")
     - Type keywords (e.g., "stacked_bar", "Arial", "auto"). If unsure, do NOT translate.
     - `null` values and empty strings.
   - Preserve units and symbols unless a direct, standard equivalent is always used in `{target_language_name}`.
   - **Output JSON MUST be identical in structure and data types to input. ONLY translatable string values change.**
3. **Translate `qa_pairs_to_translate`:**
   - Translate "query" and "label" for each item.
   - **Consistency:** Use the exact same translation for terms that appear in both the chart JSON and QA pairs.
4. **Translation Quality Requirements:**
   - **Accuracy & Fidelity:** Preserve factual meaning.
   - **Naturalness & Fluency:** Use grammatically correct, natural phrasing.
   - **Consistency:** Identical source terms = identical translation.
   - **Cultural Appropriateness:** Target-audience appropriate.
   - **Linguistic Integrity:** Correct grammar, syntax, style.
   - **Vocabulary Usage:** Accurate and context-appropriate.
   - **Non-Latin/BiDi Support:** Generate correct Unicode. Standard rendering will handle text direction.
   - **HTML tags:** Preserve tags like <br> in correct position.

**Output Format:**
You MUST respond ONLY with a single, valid JSON object containing:

- `translated_chart_json`: The processed chart JSON, structure identical to input, translations ONLY on user-facing text.
- `translated_qa_pairs`: List of translated QA pairs in the original order, each with:
  - `translated_query`
  - `translated_label`

**Input Data to Process:**

## Stage 3 Prompt for Semantic Consistency Evaluation of English Content

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' (which was 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 (either a JSON object for chart texts or a dict like `{"query": "...", "label": "..." }` for a QA pair) in `{source_language_name}`.
3. `back_translated_content`: The back-translated content (matching the structure of `original_content`) in `{source_language_name}`.

**Evaluation Criteria and Rating Scale (1-5):**

- **Focus:** Semantic meaning and preservation of key information. Does the back-translation mean the same thing as the original?
- **Ignore:** Minor grammatical variations, stylistic choices, or synonymous phrasing common in translation *unless* they significantly alter the meaning, introduce ambiguity, or omit/distort critical information.
- 5: **Excellent Equivalence** — Perfect semantic match; only stylistic or trivial differences.
- 4: **Good Equivalence** — Main meaning and most key info conveyed accurately; minor acceptable differences.
- 3: **Fair Equivalence** — General topic captured, but some important details, nuance, or accuracy lost.
- 2: **Poor Equivalence** — Significant errors; key info 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 vs. `original_content` JSON.
- Respond ONLY with a single, valid JSON object with TWO keys:
  - `rating`: integer (1-5, overall JSON equivalence)
  - `reason`: string (brief justification)

Example:

```
{
  "rating": 4,
 "reason": "Overall JSON text equivalence is good. Most titles and labels match semantically, though..."
}
```

**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 equivalence)
  - `query_reason`: string (brief justification for query)
  - `label_rating`: integer (1-5, label/answer equivalence)
  - `label_reason`: string (brief justification for label)

Example:

```
{
  "query_rating": 5,
  "query_reason": "Back-translated query perfectly matches original meaning.",
  "label_rating": 5,
  "label_reason": "Back-translated label is identical to original."
}
```

**Final Instruction:** Analyze the `original_content` and `back_translated_content` based on `context`. Respond ONLY with the valid JSON object matching the required output format for that context.

### C.2 Human and Automatic Evaluation Prompts

**Prompt For Proxy Human Evaluation**

**System Prompt:**

You are a highly specialized AI assistant, combining the expertise of a *Professional Chart Domain Expert* and a *Professional Translator*.
You will be provided with two chart images (one in English, one in `[TARGET_LANGUAGE_NAME]`) and their corresponding Question-Answer (QA) pairs.

Your task is to critically evaluate three aspects of the `[TARGET_LANGUAGE_NAME]` materials, using the English materials as a reference when needed.
For each aspect, provide a score from 1 to 3, along with concise reasoning.

**Evaluation Dimensions and Criteria:**

**A. Image Quality Assessment (`[TARGET_LANGUAGE_NAME]`):**

- **Clarity & Accuracy:** Is the image clear, and does the chart type accurately reflect the data?
- **Text & Elements:** Are all textual and graphical elements in `[TARGET_LANGUAGE_NAME]` correctly displayed and legible?
- **Overall Integrity:** Is the chart visually professional and undistorted?

*Scoring:* 3 = Excellent; 2 = Minor flaws; 1 = Major issues hindering comprehension.

**B. QA Correctness Assessment (`[TARGET_LANGUAGE_NAME]`):**

- **Relevance:** Is the question relevant to the chart?
- **Accuracy:** Is the answer correct and fully supported by the chart?

*Scoring:* 3 = Excellent; 2 = Minor errors or ambiguity; 1 = Major errors or irrelevance.

**C. Translation Accuracy (English to `[TARGET_LANGUAGE_NAME]` QA):**

- **Fidelity:** Does the translation preserve key informational elements?
- **Semantic Equivalence:** Is the meaning consistent between languages?
- **Naturalness & Fluency:** Does the translation read naturally in `[TARGET_LANGUAGE_NAME]`?

*Scoring:* 3 = Excellent; 2 = Minor issues; 1 = Major errors or awkwardness.

**Output Format:**
Provide your evaluation as a JSON object:

```
{
  "image_quality_score": <score_A>,
  "image_quality_reasoning": "<Your reasoning for score_A>",
  "qa_correctness_score": <score_B>,
  "qa_correctness_reasoning": "<Your reasoning for score_B>",
  "translation_accuracy_score": <score_C>,
  "translation_accuracy_reasoning": "<Your reasoning for score_C>"
}
```

Replace <score_A>, <score_B>, <score_C> with integers (1, 2, or 3), and provide concise justifications.
Do not include any text outside this JSON. The first image is English; the second is `[TARGET_LANGUAGE_NAME]`.