Beyond Single Plots: A Benchmark for Question Answering on Multi-Charts

Azher Ahmed Efat and **Seok Hwan Song** and **Wallapak Tavanapong**
Department of Computer Science, Iowa State University, Ames, Iowa, USA
{efat, song92, tavanapo}@iastate.edu

Abstract

Charts are widely used to present complex information. Deriving meaningful insights in real-world contexts often requires interpreting multiple related charts together. Research on understanding multi-chart images has not been extensively explored. We introduce PolyChartQA, a mid-scale dataset specifically designed for question answering over *multi-chart images*. PolyChartQA comprises 534 multi-chart images (with a total of 2,297 sub-charts) sourced from peer-reviewed computer science research publications and 2,694 QA pairs. We evaluate the performance of nine state-of-the-art Multimodal Language Models (MLMs) on PolyChartQA across question type, difficulty, question source, and key structural characteristics of multi-charts. Our results show a 27.4% LLM-based accuracy (L-Accuracy) drop on human-authored questions compared to MLM-generated questions, and a 5.39% L-accuracy gain with our proposed prompting method.

1 Introduction

Charts play a critical role in presenting data, trends, and information in real-world contexts, including scientific research, finance, and policy making (Beattie and Jones, 2002). Charts often give rise to questions that require substantial analytical skill, a deep understanding of visual features, precise data extraction, and multi-step reasoning operations based on the chart content (Kim et al., 2020; Masry et al., 2022; Hoque et al., 2022).

Multimodal Language Models (MLMs) have demonstrated strong performance across diverse real-world vision-language tasks, including visual question answering (Li et al., 2023; Masry et al., 2022), image captioning and generation (Rotstein et al., 2024; Koh et al., 2023), summarization (Islam et al., 2024), and many chart understanding tasks (Liu et al., 2023a; Akhtar et al., 2023; Masry et al., 2025b; Zhang et al., 2024).

Existing research on chart understanding has been limited to single chart images. However, complex decision-making in real-world contexts also requires an understanding of a multi-chart image with at least two relevant sub-chart images. Understanding multi-chart images poses unique challenges, requiring integration across related sub-charts, interpretation of diverse visual structures, and resolution of visual ambiguities (Hoque et al., 2022).

MultiChartQA (Zhu et al., 2025) and CharXiv (Wang et al., 2024) are two major datasets for evaluation of MLM performance on multi-chart question answering (QA). MultiChartQA has multi-chart sets, each comprising two or three related single- or multi-chart images to simulate multi-chart scenarios. Its questions include explicit chart references (e.g., "first chart") and answer-format instructions, unlike how end-users phrase questions or how composite figures naturally appear. Performance drops when such references are removed or when actual multi-chart images are used (Zhu et al., 2025). However, the joint effect of these factors on performance relative to single-chart settings has not been investigated. MultiChartQA lacks annotations that describe multi-chart characteristics (e.g., chart type, homogeneity, and the sub-chart count).

Similarly, the authors of CharXiv explore descriptive and reasoning questions but omit data retrieval questions, which require precise extraction of numerical values, often under visual ambiguity (e.g., overlapping bars, unclear ticks). It also does not consider question difficulty or other multi-chart attributes such as homogeneity or chart type, nor provides corresponding annotations. Its questions include positional cues (e.g., "row 1, column 2") and short, direct answers, diverging from the open-ended nature of real-world multi-chart queries.

In addition, due to the scalability limitations of human-authored QA pairs, recent studies have increasingly relied on MLM-generated questions for chart QA tasks (Shinoda et al., 2024; Li et al., 2024;

Findings of the Association for Computational Linguistics: ACL 2026, pages 35374–35411

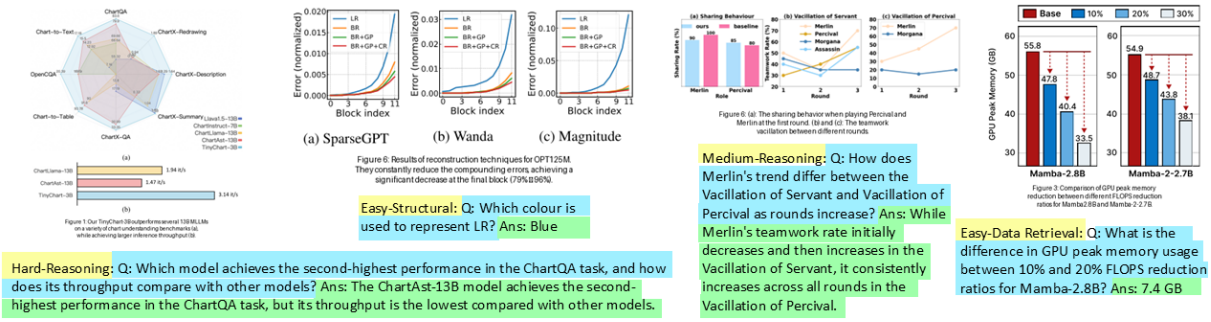


Figure 1: Example human-authored QA pairs from PolyChartQA. Multichart images are from Zhang et al. (2024); Shin et al. (2024); Lan et al. (2024); Zhan et al. (2024).

Pramanick et al., 2024). However, this growing reliance on MLM-generated QA raises the question of their suitability for evaluating multi-chart reasoning and how performance differs from human-authored questions in multi-chart contexts. Despite the importance and prevalence of multi-chart images, their complexity remains underexplored in existing research. To bridge this gap, we pose the following research questions.

- RQ1: How does MLM performance differ between single-chart and multi-chart images when questions lack explicit chart references?
- RQ2: How do question and multi-chart characteristics influence MLMs on multi-chart QA?
  - RQ2.1: To what extent does question difficulty (i.e., easy, medium, hard) influence MLM performance?
  - RQ2.2: Does the question type (structural, data retrieval, or reasoning) affect MLM performance?
  - RQ2.3: How does chart type homogeneity (sub-charts are of the same type versus mixed types) affect performance?
  - RQ 2.4: Does increasing sub-chart count impact MLM performance?
- RQ3. To what extent are human-authored multi-chart questions more challenging for MLMs than MLM-generated ones?
- RQ4. Does stepwise visual decomposition and self-verification prompting improve MLM performance and interpretability in multi-chart QA?

To systematically explore these research questions, we introduce **PolyChartQA**[1], for question answering over multi-chart images. PolyChartQA comprises 534 multi-chart images containing a total of 2,297 sub-charts sourced from peer-reviewed computer science research papers and 2,694 question-answer pairs annotated across diverse question types and difficulty levels. Figure 1 shows example human-authored QA pairs.

PolyChartQA differs fundamentally from MultiChartQA (Zhu et al., 2025) and CharXiv (Wang et al., 2024) in both annotation depth and task formulation. First, PolyChartQA includes explicit multi-chart annotations that are absent in the existing datasets, such as chart type at the sub-chart level, and chart homogeneity. These annotations enable controlled analysis of multi-chart structure and reasoning complexity, which cannot be supported by the existing datasets. Second, PolyChartQA does not have explicit chart references in the questions (e.g., "first chart"). In contrast, existing datasets rely on such identifiers. This design choice forces MLMs to localize and select the relevant sub-chart(s) before reasoning, making the task closer to how questions naturally arise and substantially more challenging. Third, PolyChartQA preserves composite multi-chart figures as single images with embedded sub-charts, whereas MultiChartQA primarily represents multi-chart examples as sets of separate chart images. Finally, PolyChartQA explicitly categorizes questions into difficulty levels. PolyChartQA contains 941 more open-ended questions than MultiChartQA, expanding both task diversity and evaluation coverage.

In summary, we make five primary contributions: (1) Formulation and investigation of the aforementioned research questions on MLM performance in multi-chart QA, focusing on chart structure, ques-

[1]PolyChartQA Dataset: https://github.com/NRT-D4/PolyChartQA

tion difficulty, and human-authored vs. MLM-generated QA; (2) PolyChartQA, a benchmark dataset specifically designed to evaluate question answering over multi-chart images; (3) A comprehensive evaluation of nine state-of-the-art MLMs on the multi-chart QA task on two datasets; (4) A stepwise visual decomposition and self-verification pipeline to enhance MLM performance and interpretability on multi-chart QA; and (5) Key findings that MLMs perform significantly worse on multi-chart images compared to single-chart ones (up to 36.98% L-accuracy drop), and that human-authored questions are substantially more challenging than MLM-generated ones, with performance gaps as large as 27.4% L-accuracy. Our results highlight the challenges that MLMs face when reasoning over multi-chart images. Our curated dataset complements existing datasets to evaluate and improve MLM performance on multi-charts.

2 Literature Review

2.1 Chart Question-Answering Datasets

Single-chart QA datasets: Most chart QA datasets contain single-chart images as summarized in Table 1. The questions of these datasets can be broadly categorized into (1) template-based (Kafle et al., 2018; Chaudhry et al., 2020; Kahou et al., 2017; Methani et al., 2020), (2) human-authored (Masry et al., 2022; Huang et al., 2025), and (3) MLM-generated questions (Shinoda et al., 2024; Li et al., 2024).

Multi-chart QA datasets: MultiChartQA (Zhu et al., 2025) has 1,753 open-ended questions and 247 multiple-choice questions over 500 multi-chart sets (≈2,168 sub-charts). However, approximately 75.2% of the multi-chart sets consist solely of two or three single charts. MMC-Benchmark (Liu et al., 2024a) contains 52 multi-chart images, which is inadequate to thoroughly assess model performance on multi-chart tasks (Zhu et al., 2025). MMC-benchmark also lacks an in-depth exploration of the question-level and chart-level complexity in multi-chart tasks. CharXiv (Wang et al., 2024) contains 1,427 multi-chart images.

2.2 Existing Models for Chart Understanding

Early chart understanding primarily relied on Optical Character Recognition and deep neural networks (Kafle et al., 2018; Chaudhry et al., 2020), followed by transformer-based models (Singh and Shekhar, 2020). Works on MLMs for chart-to-table conversion and evaluation include (Methani et al., 2020; Masry et al., 2023; Liu et al., 2023a; Kim et al., 2025; Masry et al., 2022; Zhou et al., 2023; Liu et al., 2025; Meng et al., 2024). Recent works focus on developing MLMs tailored for chart understanding (Zhang et al., 2024; Masry et al., 2024; Liu et al., 2024a; Masry et al., 2025b) and evaluation of MLMs on chart reasoning tasks (Islam et al., 2024; Zhu et al., 2025; Mukhopadhyay et al., 2024; Huang et al., 2025; Wang et al., 2024; Tang et al., 2025; Masry et al., 2025a; Wu et al., 2024b; Hutchinson et al., 2025).

3 Proposed PolyChartQA

We present PolyChartQA, a benchmark dataset for multi-chart question answering. PolyChartQA comprises 534 multi-chart images and 2,694 QA pairs (Human-authored: 519, MLM-generated: 2,175). MLM-generated QA pairs were generated using GPT-4.1 and manually verified. Table 1 compares our dataset with the existing datasets.

3.1 Charts Collection

We curated a collection of multi-chart images from 168 research papers published in 2024 across 4 top-tier computer science conferences: EMNLP, ICSE, ICML, and SIGCOMM. We then used PDFFigures 2.0 (Clark and Divvala, 2016) to extract all the figures from these articles. Following figure extraction, a manual filtering process was conducted to identify and retain only multi-chart images, resulting in a dataset of 534 multi-chart images. Each multi-chart image was manually annotated with the number of sub-charts, chart type homogeneity, and specific sub-chart types by the first author. Of these images, 85.58% are homogeneous multi-chart images where all sub-charts are of the same type. The remaining images have multiple types of sub-charts. PolyChartQA includes 11 chart types. Although this reflects an imbalance in chart homogeneity, the imbalance originates from real-world sources.

To validate the annotations, given each multi-chart image, we prompted GPT-4o to annotate with the number of charts, the type of sub-charts, and the chart type homogeneity. We validated annotations for all 534 multi-chart images. The agreement between the first author and the GPT-4o annotations was 91.95%. The disagreement was due to sub-chart count (4.49%), sub-chart type (3.93%), and chart homogeneity (0.75%), with some instances involving overlaps across these categories. The

Chart Setting	Dataset	Image Type	# Images	# QA	Question Generation	Sub-chart Ref. in Q	Chart Annotation
Single-chart	LEAF-QA (Chaudhry et al., 2020)	Synthetic from real data	250K	2M	Template-based	-	Chart type, Bounding box
	PlotQA (Methani et al., 2020)	Synthetic from real data	224.3K	28.9M	Template-based	-	Chart type Bounding box
	ChartQA (Masry et al., 2022)	Real-world	21.9K	32.7K	Human-authored + Machine-generated	-	None
	SBS Figures (Shinoda et al., 2024)	Synthetic	1M	4.2M	LLM-generated (Verified 100 QAs)	-	Chart type
	SynChart (Liu et al., 2024b)	Synthetic	4M	59.7M	LLM-generated	-	Chart type
	EvoChart-QA (Huang et al., 2025)	Real-world	625	1,250	Human-authored	-	Chart type
	CharXiv (Wang et al., 2024)	Real-world	896	3,584*	Human-authored + Verified GPT-inspired & generated	-	# of sub-chart
Multi-chart (Simulated)	MultiChartQA (Zhu et al., 2025)	Real-world	500 multi-chart sets (≈ 2,168 sub-charts)	2K	Human-authored	Yes	None
Multi-chart	CharXiv (Wang et al., 2024)	Real-world	1,427[1]	5,708*	Human-authored + Verified GPT-inspired & generated	Yes	# of sub-chart
	PolyChartQA (Ours)	**Real-world Computer Science paper**	**534 divided into 2,297 sub-charts**	**2.69K**	**Human-authored + Verified MLM-generated**	**No**	**# of sub-chart, sub-chart type, chart homogeneity**

Table 1: Comparison of PolyChartQA with existing chart question answering datasets; [1]of which 161 images and 161 reasoning QA are from computer science; * denotes #reasoning and #descriptive questions.

first author ensured the accuracy and consistency of the analysis. As GPT-4.1 became available later, we used it for question-and-answer generation. Appendix A.1 summarizes multi-chart image distributions by conference, sub-chart count, and chart type.

3.2 Question Generation

We define three difficulty levels and three question types for our experiments. Our questions for each multi-chart image are open-ended and can be answered from the image. All question criteria were established before the annotation process began and were applied uniformly during both QA generation and subsequent manual verification. This strict adherence to predefined guidelines ensured consistency in question formulation, reducing individual annotator stylistic differences and intent-level bias.

Question Difficulty Levels

- **Easy:** Easy questions focus on a single sub-chart with straightforward retrieval or trend analysis. Our Easy questions are similar to the Direct questions from MultiChartQA.
- **Medium:** Medium questions require comparing, aggregating, or synthesizing information spread across two or more sub-charts. The reasoning is relatively direct and does not require recursive visual inference.
- **Hard:** Hard questions demand multi-step reasoning across multiple sub-charts. These include tasks that require chaining information across sub-charts, for example, extracting a label or legend from one chart and applying it to interpret symbols or data points in another. Our Hard questions are inspired by the Sequential Reasoning questions in MultiChartQA (Zhu et al., 2025). However, MultiChartQA's sequential reasoning questions only focus on content-based tasks, such as extracting specific numbers or phrases from given clues, and do not include structural questions. In contrast, our Hard questions span structural, data retrieval, and reasoning types, all without providing explicit chart references.

Question Types: We use the types of questions following existing works (Methani et al., 2020).

- **Structural:** Focuses on visual or structural elements, e.g., colors and labels.
- **Data Retrieval:** Focuses on retrieving explicit numeric or categorical values, e.g., reading data points, and extracting counts.
- **Reasoning:** Involves inference, comparison, aggregation, trend analysis, or contradiction.

Human-authored PolyChartQA: The human-authored QA pairs were manually crafted by the first author, a computer science graduate student. Initially, 520 QA pairs were created from 238 multi-chart images. These QA pairs were then evenly distributed between the second author (a computer science graduate student) and the third author (a professor of computer science) for validation. Each question was validated by a single annotator (either second or third author). The validation focused on two criteria: (i) clarity of the question and (ii) correctness of the QA pair. Based on the feedback, one image and its associated one question–answer pair were excluded as both the question and the answer were incorrect. Additionally, 98 QA pairs were revised for clarity and correctness.

MLM-generated PolyChartQA: We prompted GPT-4.1 to generate QA pairs from all the multi-chart images. Our prompt includes the criteria for the desired question difficulty level, question type, multi-chart image, and instructions to generate QA pairs per the given criteria. Initially, we generated a total of 4,308 QA pairs from 534 images. **Manual Verification:** All the generated QA pairs were checked by the first author to ensure correctness according to our criteria given in A.2 Table 6. About 49.5% of the QA pairs (2,133 QA pairs) and 3 multi-chart images were removed. Next, the first author verified the correctness of the question difficulty, question type, and relevant chart annotations for the remaining questions according to our definitions of difficulty and question types. After manual checking, annotation of 18.99%, 11.26%, and 7.59% QA pairs was revised for the question difficulty, question type, and relevant chart. For the relevant charts, most of the edits were changing specific chart names to 'Any chart'. Additionally, we found 16 QA pairs for which GPT-4.1 assigned more than one question type (e.g., Structural/Reasoning instead of Reasoning). We asked an independent human annotator (an undergraduate CS research assistant) to correct the labeling mistakes for these 16 QA pairs. The annotator was briefed with detailed definitions of question types, difficulty levels, and task instructions.

Although scaling the dataset size through MLM-generated questions is an option, we prioritized quality by manually verifying all QA pairs generated by the MLM. This decision inevitably limited scalability but ensured correctness, which is particularly critical given the substantially greater complexity of multi-chart images and harder question types compared to single-chart. Prior large-scale efforts illustrate this inherent trade-off between manual validation and scalability. QA pairs from less than 1% of the number of papers in the SPIQA dataset were manually verified (Pramanick et al., 2024), of which about 11% were discarded. Similarly, about 52.21% of all QA pairs were discarded after manual verification (Hutchinson et al., 2025). We chose to retain fewer higher-quality questions by validating every pair. Table 2 presents the statistics of our dataset. The Distinct-2 (Li et al., 2016) (unique bigram ratio) of the human-authored questions (0.4516) in Table 2 is higher than those of the MLM-generated ones and MultiChartQA (0.2257), indicating greater lexical diversity. We compute Distinct-2 using only the core question text, excluding answer instructions. Beyond Figure 1, additional human- and MLM-generated examples are in A.16 - A.17. A.18 presents the prompts for validating chart annotation, and generating QA pairs.

Statistic	Number
Human-authored PolyChartQA	
Images	237
Sub-charts	894
Homogeneous charts	209
Non-homogeneous charts	28
Total questions	519
Question Difficulty:	
Easy	293
Medium	168
Hard	58
Question Type:	
Structural	89
Data Retrieval	141
Reasoning	289
Median questions per image	1.0
Distinct-2 (Question)	0.4516
Avg. Token Length (Question)	15.48
Min - Max Token Length (Question)	5 - 51
Verified MLM-generated PolyChartQA	
Images	531
Sub-charts	2,290
Homogeneous charts	454
Non-homogeneous charts	77
Total questions	2,175
Question Difficulty:	
Easy	1,208
Medium	867
Hard	100
Question Type:	
Structural	763
Data Retrieval	302
Reasoning	1,109
Median questions per image	4.0
Distinct-2 (Question)	0.3356
Avg. Token Length (Question)	18.49
Min - Max Token Length (Question)	8 - 56
Total (Combined)	
Images	534
Sub-charts	2,297
Median sub-charts per image	3
Homogeneous charts	457
Non-homogeneous charts	77
Unique papers	168
Unique chart types	11
Total questions	2,694

Table 2: PolyChartQA dataset statistics.

4 Proposed Visual Decomposition and Self-Verification Pipeline (VDSP)

To improve MLM performance and interpretability in complex multi-chart QA, we propose a modular 3-stage prompt-based pipeline that decomposes the reasoning process while leveraging the visual understanding capabilities of MLMs. Figure 2 shows our pipeline with explicit visual decomposition,

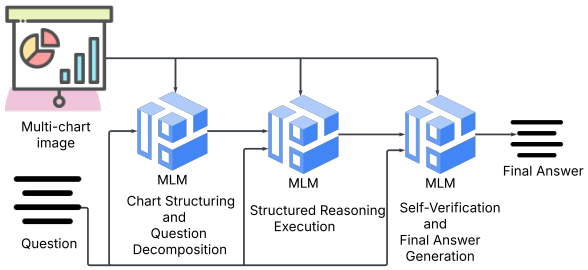


Figure 2: VDSP pipeline for multi-chart QA.

structured reasoning, and self-verification, aligned with the observed error types.

Stage 1. Chart Structuring and Question Decomposition: Given a multi-chart image and the question, the MLM is prompted to first parse the image into individual sub-charts, identifying their types (e.g., bar, line) and key characteristics (e.g., axis labels, legends). Next, the question is decomposed into a sequence of sub-tasks, grounded in chart structure. This mimics a form of visual Chain of Thought reasoning, but with explicit mapping between visual elements and reasoning steps.

Stage 2. Structured Reasoning Execution: Given the multi-chart image, question, and Stage 1 output, the MLM extracts relevant values from specific visual components (e.g., bar heights) and performs computations or comparisons as needed. Finally, the MLM combines intermediate outcomes to formulate a preliminary answer.

Stage 3. Self-Verification and Final Answer Generation: Given the multi-chart image, the question, and the output from Stage 2, the model performs a self-reflective check on the reasoning in Stage 2 by verifying whether all chart references and visual details are accurate, followed by detecting common reasoning errors (e.g., incorrect axis readings, incorrect chart selection). It returns either a validated or a corrected final answer.

5 Experimental Setup

Nine compared models: We evaluated three state-of-the-art closed-source MLMs (Claude-3.7-Sonnet, GPT-4.1, and Gemini-2.0-Flash), four open-source MLMs (Pixtral-12B, Llama-3.2-11B-Vision, Llava-1.5-7b, and Gemma-3-4b-it), and two chart-specialized MLMs, ChartGemma (Masry et al., 2025b) and MatCha (Liu et al., 2023b). See the configurations in A.5.

Two datasets: PolyChartQA was used for testing RQ2-4 while MultiChartQA-RQ1 was used for RQ1. MultiChartQA-RQ1 consists of all 365 "Direct" questions from the MultiChartQA (Zhu et al., 2025) dataset in which all the images in the multi-chart sets are single-chart images. Direct questions refer to those for which an MLM can derive the correct answer using information from a single chart. We replaced explicit sub-chart identifiers (e.g., "first chart") in the question with the generic term "chart". The same QA pairs were used for both the single-chart and multi-chart conditions. The two conditions differ only in the input image configuration. In the single-chart setting, the MLM was provided with only the relevant sub-chart needed to answer the question. In the multi-chart setting, we constructed a composite multi-chart image by merging all charts from the corresponding set into a single image. As the questions and answers were identical across both settings, any observed performance differences directly reflect the added difficulty of sub-chart localization and reasoning in a multi-chart visual context, thereby isolating the effect of multi-chart complexity.

Prompting methods: We used Zero-shot prompting for RQ1, Zero-shot and Chain-of-Thought (CoT) for RQ2–RQ4, and also the proposed VDSP prompting on the human-authored QA pairs of PolyChartQA for RQ4.

Evaluation Metrics: We use H-Accuracy (Human-evaluation), L-Accuracy (LLM-based accuracy) introduced by the previous works (Pramanick et al., 2024; Liu et al., 2024a; Tang et al., 2025; Wang et al., 2024), and BERTScore (Zhang et al., 2020). BERTScore measures semantic similarity between model predictions and references. H-accuracy was obtained through human evaluation of model-predicted answers for human-authored questions in a Zero-shot setting. As human evaluation is difficult to scale, for all other cases, we use L-accuracy by prompting a selected LLM as a judge to assess whether the ground-truth and model-generated answers are similar or not.

Judge selection: We randomly sampled 100 QA pairs from PolyChartQA, where seven general-purpose MLM-predicted answers were evaluated independently by a human and three LLMs: Claude-3.7-Sonnet, GPT-4.1, and Gemini-2.0-flash. The results showed substantial to almost perfect Cohen's Kappa agreement between the human and Claude-3.7-Sonnet (0.73–0.92), the human and GPT-4.1 (0.65–0.94), and moderate to substantial agreement between the human and Gemini-2.0-flash (0.46-0.84). Given its highest alignment with

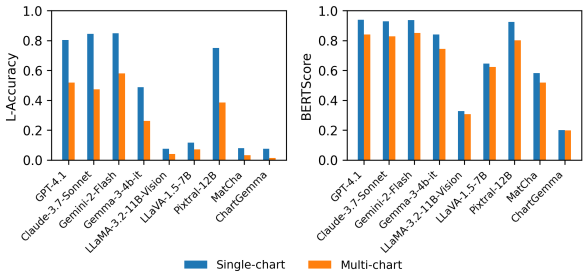


Figure 3: Single-chart vs multi-chart QA performance.

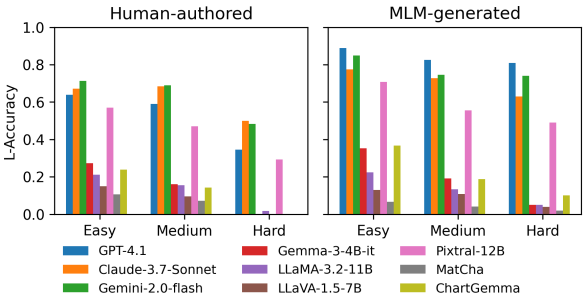


Figure 4: L-Accuracy on PolyChartQA under Zero-shot prompting across difficulty levels.

human evaluation, Claude-3.7-Sonnet was selected as the judge to compute L-Accuracy. Detailed agreement statistics are provided in A.3.

Following previous studies (Wang et al., 2024; Pramanick et al., 2024; Liu et al., 2024a; Tang et al., 2025), we report all metrics based on a single evaluation run. While minor variability may occur across runs or model versions, Claude-3.7-Sonnet showed strong alignment with human judgments, ensuring reliable evaluation. The prompt for L-Accuracy is provided in A.18.

6 Experimental Results

Underlined scores indicate the best performance within each category. Performance drops are computed as the score difference across categories for each individual MLM.

RQ1 Findings: Several MLMs perform significantly worse on multi-chart QA compared to single-chart QA. Figure 3 shows that L-accuracy drops from single- to multi-chart settings by 26.85%–36.98% for closed-source, 3.56%–36.44% for open-source, and 4.66%–6.30% for chart-specific models. MatCha and ChartGemma perform poorly, likely due to their pre-training on single-chart datasets. BERTScore decreases by up to 12.38% when processing multi-chart images, indicating reduced semantic alignment.

RQ2.1 Findings: ***Performance consistently declines with increasing question difficulty.*** This trend is consistent across all MLMs and question sources, underscoring the difficulty MLMs face with multi-step reasoning and cross-chart aggregation. Figure 4 reports Zero-shot L-Accuracy, with other results in A.6. **Human-authored questions:** Both H- and L-accuracy consistently decline from Easy to Hard across all MLMs. For L-Accuracy under Zero-shot, the drop ranges are 17.24%–29.34% (closed-source MLMs), 15.02%–27.69% (open-source MLMs), and 10.58%–23.89% (chart-specific MLMs). For H-Accuracy, the corresponding drops are 23.09%–27.94% (closed-source MLMs), 16.04%-31.10% (open-source MLMs), and 10.92%–26.28% (chart-specific MLMs). CoT prompting slightly mitigates the decline, but substantial gaps remain. **MLM-generated questions:** A similar trend is observed, with declines of 4.62%–30.35% (Zero-shot), and 5.79%-28.27% (CoT) L-Accuracy.

RQ2.2 Findings: ***MLM performances vary by question types.*** The models consistently perform best on structural and worst on data retrieval questions. **Human-authored questions:** Both H- and L-accuracy consistently decline from structural, reasoning, to data retrieval types across MLMs except for MatCha for both Zero-shot and CoT prompting. Figure 5 and A.7 show detailed results. In summary, for L-Accuracy under Zero-shot, the drop ranges are 20.5%-35.09% (closed-source MLMs), 17.16%-28.35% (open-source MLMs), and 9.88%-34.31% (chart-specific MLMs). For H-Accuracy, the corresponding drops are 20.5%-37.76% (closed-source MLMs), 10.42%-28.89% (open-source MLMs), and 8.46%-35.13% (chart-specific MLMs). **MLM-generated questions:** A similar trend is observed for all closed-source models and Pixtral, with declines of 10.74%–22.89% (Zero-shot), 11.73%-25.59% (CoT). The rest of the MLMs perform the best on structural questions, but sometimes better on data retrieval than on reasoning questions.

RQ2.3 Findings: ***Most MLMs perform better on non-homogeneous charts.*** The detailed results are in A.8. In summary, under Zero-shot on human-authored questions, both H- and L-Accuracy decline from non-homogeneous to homogeneous charts, with drops of up to 9.31% (L-Accuracy) and 8.40% (H-Accuracy) for closed-source MLMs,

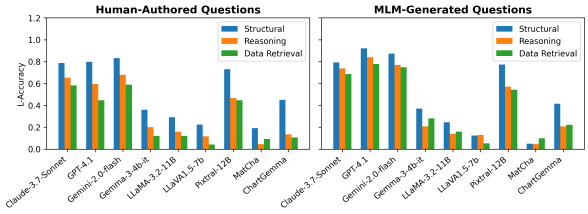


Figure 5: L-Accuracy on PolyChartQA under Zero-shot setting across question types.

Model	Zero-shot		CoT	
	Human	MLM	Human	MLM
Claude-3.7-Sonnet	**0.6570**	**0.7494**	**0.6647**	**0.7471**
GPT-4.1	**0.5896**	**0.8598**	**0.5954**	**0.8694**
Gemini-2.0-flash	**0.6802**	**0.8028**	**0.6936**	**0.8106**
Gemma-3-4b-it	**0.2062**	**0.2749**	0.2158	0.2529
LLaMA-3.2-11B-Vision	0.1715	0.1794	**0.0848**	**0.1559**
LLaVA1.5-7b	0.1156	0.1168	0.1195	0.1159
Pixtral-12B	**0.5067**	**0.6377**	**0.5337**	**0.6354**
MatCha	**0.0829**	**0.0543**	0.0617	0.0441
ChartGemma	**0.1811**	**0.2828**	0.1911	0.2336

Table 3: L-Accuracy on Human vs. MLM-generated questions. Bold pairs indicate statistically significant differences at a significance level of 0.05.

17.41% (L-Accuracy) and 14.58% (H-Accuracy) for open-source MLMs, and 1.26% (H-Accuracy) for chart-specific MLMs except GPT-4.1, Gemma, and ChartGemma. The downward trend holds for MLM-generated questions, up to 9.02% (L-Accuracy). CoT mitigates the gap mostly.

RQ2.4 Findings: ***MLM performance generally declines as the number of sub-charts increases.*** This trend is more pronounced for human-authored QAs than MLM-generated QAs. Detailed results are in A.9.

RQ3 Findings: Human-authored questions are significantly harder than MLM-generated ones. Table 3 reveals a statistically significant accuracy gap between human-authored and MLM-generated questions for most MLMs using the Pearson's chi-squared statistical test. With Zero-shot prompting, L-accuracy drops from MLM-generated QAs to human-authored QAs by 9.24–27.02% (closed-source), up to 13.1% (open-source), and up to 10.16% (chart-based). With CoT, L-accuracy decreases is similar (up to 27.4%). BERTScore is presented in A.10.

RQ4 Findings: Incorporating stepwise visual decomposition and self-verification enhances the L-accuracy and interpretability. Table 4 shows L-accuracy of the closed-source models and Pixtral-12B, the best-performing open-source

Model	Zero-shot	CoT	VDSP
GPT-4.1	0.5896	0.5954	0.5954
Claude-3.7-Sonnet	0.6570	0.6647	0.7110
Gemini-2.0-flash	0.6802	0.6936	0.7013
Pixtral-12B	0.5067	0.5337	0.5145

Table 4: L-Accuracy of Zero-shot, CoT, and VDSP.

model on human-authored questions from PolyChartQA. VDSP improves L-accuracy by up to 5.39% over Zero-shot and 4.62% over CoT. While GPT-4.1 exhibits similar accuracy under CoT and VDSP, VDSP provides interpretability. First, VDSP explicitly decomposes the task into the 1) sub-chart identification; 2) reasoning; and 3) verification, making the model's failure points transparent (e.g., mis-localization, incorrect reasoning). Second, VDSP reveals exactly which sub-chart the model selected and how it used it. This level of structured interpretability and diagnostic clarity is not available in CoT and is particularly valuable for multi-chart tasks, where transparency and traceability are important. A qualitative example illustrating the interpretability benefit of VDSP is given below.

Qualitative Example: Question: *"Which colour is used to represent the optimized prompt for phi-2 model?"*; CoT: *"Blue"*; VDSP: *"The reasoning process is correct. The bottom-left chart is labeled 'phi-2', and the legend clearly shows that 'Optimized' is represented by blue. The blue bars in the phi-2 chart correspond to the optimized prompt. Therefore, the color used to represent the optimized prompt for the phi-2 model is blue."*

Figures 45–46 in A.19 show an example intermediate output. BERTScore is reported in A.11. The prompts for Zero-shot, CoT, and VDSP used across RQ1-4, are listed in Figures 37-41 in A.18.

Error Analysis on Human-Authored Questions from PolyChartQA: We analyzed one-third of the human-authored questions that all MLMs answered incorrectly under human evaluation. Most errors stem from line selection errors in line plots (21–26%), visual value misreading (15–26%), and sub-chart misidentification (15–21%). Other error types include axis misalignment, calculation mistakes, legend association errors, multi-step reasoning failures, and incorrect value comparisons. Open-source MLMs often produce incomplete answers. Examples are provided in A.20.

7 Ablation Studies

Variants of the VDSP prompting strategy: In the first variant, Stage 1 was modified to generate structured tables instead of textual descriptions, while Stages 2 and 3 remained unchanged, resulting in a performance drop of up to 4.8% L-Accuracy. In the second variant, all three prompts were revised, though Stages 1 and 3 retained their original roles, Stage 2 adopted a dual-persona (Analyst–Reviewer) design. This led to performance comparable to the Zero-shot setting, with a 2.7% L-Accuracy drop in Pixtral-12B, confirming Stage 1 as the most critical component. Detailed results are provided in A.13 - A.14.

Overall H-Accuracy: H-Accuracy ranges from 9.06%-70.52% under Zero-shot setting, and 6.17%-71.87% under CoT setting on human-authored PolyChartQA as presented in Table 5. As shown in Table 11 in A.12, except for Gemini-2.0-flash and ChartGemma, the absolute difference in accuracy between the human and the LLM judges is $\leq 1.2\%$ across all seven models under the CoT setting.

Model	Zero-shot	CoT
Claude-3.7-Sonnet	0.6532	0.6609
GPT-4.1	0.6108	0.5954
Gemini-2.0-flash	0.7052	0.7187
Pixtral-12B	0.5318	0.5376
LLaMA-3.2-11B	0.1522	0.0809
LLaVA-1.5-7B	0.1233	0.1175
Gemma-3-4b-it	0.2119	0.2274
MatCha	0.0906	0.0617
ChartGemma	0.1927	0.2351

Table 5: H-Accuracy on human-authored PolyChartQA under the Zero-shot and Chain-of-Thought settings.

Robustness of our findings across LLM judges: The human–vs–MLM QA performance trend remains consistent when using GPT-4.1 and Gemini-2.0-flash as LLM-judges. In Zero-shot human QA, closed-source MLMs reach 63.2%–72.1% (GPT-4.1 judge) and 67.4%–74.8% (Gemini-2.0-flash judge), while open-source MLMs reach 12.9%–53.4% and 17.0%–56.3%, respectively. On MLM-generated QA, closed-source MLMs reach 78.9%–89.7% (GPT-4.1) and 81.5%–90.8% (Gemini-2.0-flash), and open-source MLMs obtain 15.8%–69.2% and 23.1%–72.0%. Detailed results are in A.4 Table 9.

8 Conclusion

We introduce PolyChartQA, a benchmark for multi-chart QA, showing that the task is substantially harder and exhibits large performance gaps between human- and MLM-authored questions. Task decomposition and self-verification prompting help improve MLM performance and highlight key reasoning and failures. Our work facilitates further studies on multi-chart QA.

9 Limitations

Despite its contributions, PolyChartQA has a few limitations. First, human-authored QA creation is costly and time-intensive; thus, the dataset includes 519 human-authored QAs, complemented by MLM-generated QAs that are fully manually verified, prioritizing quality over scale.

Second, our dataset exhibits limited diversity along two dimensions. (1) All images are drawn from computer science research papers published in 2024, which may constrain external validity beyond academic figures and other disciplines. However, the generated questions do not require domain-specific knowledge and can be answered solely using the information presented in the charts. Moreover, the chart types used (e.g., bar charts, line charts) are common across domains and widely used in real-world applications. (2) The dataset is predominantly homogeneous by chart type (85.58%). Although this distribution reflects the real-world distribution in contemporary research papers, our findings are therefore more conclusive for homogeneous multi-chart images, with relatively limited coverage of heterogeneous multi-chart reasoning. Addressing broader disciplinary sources and increasing chart-type diversity are important directions for future extensions of PolyChartQA.

10 Ethical Considerations and Potential Risks

All charts used in our study were sourced from open-access computer science research papers. No personally identifiable information or sensitive content was used in the construction of the dataset.
ACL materials are licensed under a Creative Commons Attribution 4.0 International License, which covers papers collected from EMNLP. We selected multi-chart images from open-access EMNLP, ICML, ICSE, and SIGCOMM publications or corresponding preprints in arXiv with a Creative Com-

mons International License. We list the image names from the articles in a separate file and give credit to the original authors.

While we make our question-answers available under Commons Attribution License (CC BY 4.0), all chart images are subject to their respective copyrights by the authors of the respective papers. Additionally, we used PDFFigures 2.0 to extract the multi-chart images from the papers. PDFFigures 2.0 is publicly available under Apache License 2.0, allowing free use. We have cited the creators. The MultiChartQA dataset is available under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license that permits noncommercial use, sharing, and adaptation of the material, provided appropriate credit is given, changes are indicated, and the use remains noncommercial. We cited the creator of the artifacts and discussed the changes we made for our use.

Beyond the generation of QA pairs using MLMs, AI assistants were employed solely for non-substantive tasks such as language refinement, formatting, and basic debugging.

Acknowledgments

This work is partially supported by the National Science Foundation under Grant No. 2152117. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We acknowledge the valuable contribution of our independent annotator Tanisha Ravindran for the effort in reviewing and correcting the question-answer annotations.

References

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2023. Reading and reasoning over chart images for evidence-based automated fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia. Association for Computational Linguistics.

Randall Balestriero and Yann Lecun. 2024. How learning by reconstruction produces uninformative features for perception. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2566–2585. PMLR.

Vivien Beattie and Michael John Jones. 2002. The impact of graph slope on rate of change judgments in corporate reports. *Abacus*, 38(2):177–199.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. Leaf-qa: Locate, encode attend for figure question answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3501–3510.

Zhongwu Chen, Long Bai, Zixuan Li, Zhen Huang, Xiaolong Jin, and Yong Dou. 2024. A new pipeline for knowledge graph reasoning enhanced by large language models without fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1366–1381, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, JCDL '16, page 143–152, New York, NY, USA. Association for Computing Machinery.

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. "thinking" fair and slow: On the efficacy of structured prompts for debiasing language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, Miami, Florida, USA. Association for Computational Linguistics.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. 2024. ImageInWords: Unlocking hyper-detailed image descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 93–127, Miami, Florida, USA. Association for Computational Linguistics.

E. Hoque, P. Kavehzadeh, and A. Masry. 2022. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41(3):555–572.

Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2025. Evochart: A benchmark and a self-training approach towards real-world chart understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3680–3688.

Maeve Hutchinson, Radu Jianu, Aidan Slingsby, Jo Wood, and Pranava Madhyastha. 2025. Chart question answering from real-world analytical narratives. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 760–773, Vienna, Austria. Association for Computational Linguistics.

Mohammed Saidul Islam, Raian Rahman, Ahmed Masry, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, and Enamul Hoque. 2024. Are large vision language models up to the challenge of chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3334–3368, Miami, Florida, USA. Association for Computational Linguistics.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *2018 IEEE/CVF 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*.

Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Wonjoong Kim, Sangwu Park, Yeonjun In, Seokwon Han, and Chanyoung Park. 2025. SIMPLOT: Enhancing chart question answering by distilling essentials. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 573–593, Albuquerque, New Mexico. Association for Computational Linguistics.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 21487–21506. Curran Associates, Inc.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-based agent society investigation: Collaboration and confrontation in avalon gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 128–145, Miami, Florida, USA. Association for Computational Linguistics.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. 2024. Synthesize Step-by-Step: Tools, Templates and LLMs as Data Generators for Reasoning-Based Chart VQA . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13613–13623, Los Alamitos, CA, USA. IEEE Computer Society.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. 2023b. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada. Association for Computational Linguistics.

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, Mexico City, Mexico. Association for Computational Linguistics.

Jiapeng Liu, Liang Li, Shihao Rao, Xiyan Gao, Weixin Guan, Bing Li, and Can Ma. 2025. Union is strength! unite the power of llms and mllms for chart question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5487–5495.

Mengchen Liu, Qixiu Li, Dongdong Chen, Dong Chen, Jianmin Bao, and Yunsheng Li. 2024b. Synchart: Synthesizing charts from language models. *Preprint*, arXiv:2409.16517.

Zihang Liu, Yuanzhe Hu, Tianyu Pang, Yefan Zhou, Pu Ren, and Yaoqing Yang. 2024c. Model balancing helps low-data training and fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1311–1331, Miami, Florida, USA. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

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. 2025a. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria. Association for Computational Linguistics.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore. Association for Computational Linguistics.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. ChartInstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10387–10409, Bangkok, Thailand. Association for Computational Linguistics.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7775–7803, Bangkok, Thailand. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1525.

Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. 2024. Unraveling the truth: Do VLMs really understand charts? a deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16696–16717, Miami, Florida, USA. Association for Computational Linguistics.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*, volume 37, pages 118807–118833. Curran Associates, Inc.

Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. FuseCap: Leveraging Large Language Models for Enriched Fused Image Captions . In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5677–5688, Los Alamitos, CA, USA. IEEE Computer Society.

Sungbin Shin, Wonpyo Park, Jaeho Lee, and Namhoon Lee. 2024. Rethinking pruning large language models: Benefits and pitfalls of reconstruction error minimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1182–1191, Miami, Florida, USA. Association for Computational Linguistics.

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

Hrituraj Singh and Sumit Shekhar. 2020. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284, Online. Association for Computational Linguistics.

Ammar Tahir, Prateesh Goyal, Ilias Marinos, Mike Evans, and Radhika Mittal. 2024. Efficient policy-rich rate enforcement with phantom queues. In *Proceedings of the ACM SIGCOMM 2024 Conference*, ACM SIGCOMM '24, page 1000–1013, New York, NY, USA. Association for Computing Machinery.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Sprague, Ramya Namuduri, Bodun Hu, Juan Diego Rodriguez, Puyuan Peng, and Greg Durrett. 2025. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *Preprint*, arXiv:2505.13444.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.

Cheng-Kuang Wu, Zhi Rui Tam, Chao-Chung Wu, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024a. I need help! evaluating LLM's ability to ask for users' support: A case study on text-to-SQL generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2191–2199, Miami, Florida, USA. Association for Computational Linguistics.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024b. ChartInsights: Evaluating multimodal large language models for low-level chart question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200, Miami, Florida, USA. Association for Computational Linguistics.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024. EFUF: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1181, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Zhan, Yushu Wu, Zhenglun Kong, Changdi Yang, Yifan Gong, Xuan Shen, Xue Lin, Pu Zhao, and Yanzhi Wang. 2024. Rethinking token reduction for state space models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1686–1697, Miami, Florida, USA. Association for Computational Linguistics.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024. TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1898, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Ranchi Zhao, Zhen Leng Thai, Yifan Zhang, Shengding Hu, Jie Zhou, Yunqi Ba, Jie Cai, Zhiyuan Liu, and Maosong Sun. 2024. DecorateLM: Data engineering through corpus rating, tagging, and editing with language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1401–1418, Miami, Florida, USA. Association for Computational Linguistics.

Mingyang Zhou, Yi Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. 2023. Enhanced chart understanding via visual language pretraining on plot table pairs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1314–1326, Toronto, Canada. Association for Computational Linguistics.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2025. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11341–11359, Albuquerque, New Mexico. Association for Computational Linguistics.

A Appendix

A.1 Distribution of Multi-chart Images in PolychartQA

The dataset has these percentages of images from different conferences: EMNLP (65%), SIGCOMM (17%), ICSE (10%), and ICML (8%). Figure 6 presents the distribution of sub-chart counts in PolyChartQA. The highest number of sub-charts in an image is 72. PolyChartQA includes eleven chart types: Bar Chart, Scatter Plot, Spider Chart, Histogram, Line Chart, Box Plot, Point Plot, Surface Plot, Pie Chart, Area Chart, and Violin Plot. The distribution of chart types across homogeneous multi-chart images from PolyChartQA is shown in Figure 7.

A.2 Criteria for Removing MLM-generated Questions

Table 6 presents the criteria used to remove MLM-generated QA.

Criterion
QA explicitly mentions chart position (e.g., "left chart")
QA pair is partially or fully incorrect
Question is derived from the caption, not the chart itself
Question has multiple valid answers or is unclear
Errors in symbols (e.g., γ, λ) or formatting
Hallucinated content
QA does not follow the instructions
QA based on existing knowledge instead of chart

Table 6: Criteria for removing MLM-generated QA pairs.

A.3 Agreement between LLM Judges and Human Evaluation

Table 7 reports the Cohen's Kappa agreement between LLM-judge and human evaluation on randomly sampled 100 QA pairs from PolyChartQA. Table 8 reports the percent agreement between human evaluations and different LLM judges across all evaluated models, illustrating the reliability of each LLM judge.

A.4 L-Accuracy Comparison with Different Judges

Table 9 presents the L-accuracy on PolyChartQA when judged by GPT-4.1 and Gemini-2.0-flash.

A.5 Model Configuration

For reproducibility, we set the temperature = 0, and seed = 5 (GPT-4.1, Pixtral-12B) or sample

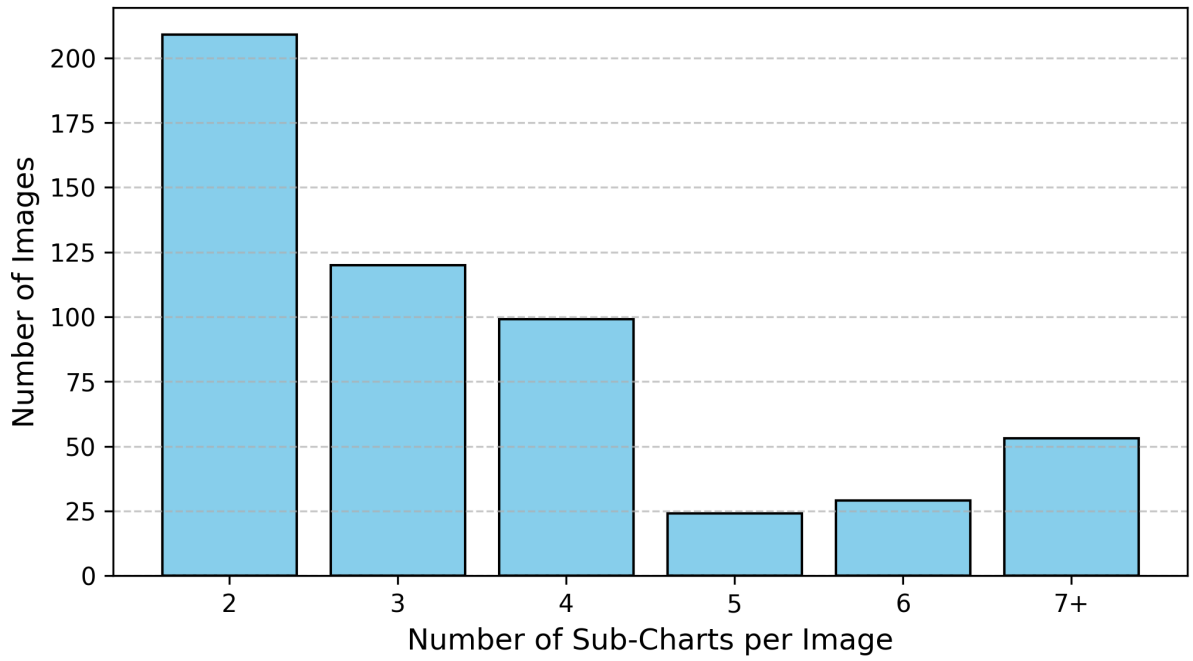


Figure 6: Distribution of sub-chart counts in PolyChartQA multi-chart images.

Judge Model	Claude-3.7-Sonnet	GPT-4.1	Gemini-2.0-flash	Gemma-3-4b-it	LLaMA-3.2-11B	LLaVA-1.5-7B	Pixtral-12B
Claude-3.7-Sonnet	0.852	0.920	0.884	0.834	0.734	0.864	0.900
GPT-4.1	0.870	0.774	0.829	0.777	0.649	0.942	0.860
Gemini-2.0-flash	0.780	0.730	0.753	0.455	0.481	0.684	0.840

Table 7: Cohen's Kappa agreement between LLM-judge and human accuracy under Zero-shot setting. Higher positive values indicate stronger agreement.

= false (Llama-3.2-11B-Vision, Llava-1.5-7b, and Gemma-3-4b-it). For open-sourced MLMs (except Pixtral-12B), we ran these models in Google Colab with A100/L4 GPU. GPT-4.1, Claude-3.7-Sonnet, Gemini-2.0-flash, and Pixtral-12B were used through API calls. BERTScore is computed using the bertscore Python package (v0.3.13) with the bert-base-uncased model and default parameters.

A.6 Results across Question Difficulty Levels on PolyChartQA

Figure 8 presents the detailed human evaluation result on human-authored QAs from PolyChartQA under the Zero-shot setting across question difficulty levels. Figures 9-11 show the L-accuracy under CoT and BERTScore under the Zero-shot and CoT settings on PolyChartQA across question difficulty levels.

A.7 Results across Question Types on PolyChartQA

Figure 12 presents the H-Accuracy result on human-authored QAs from PolyChartQA under the Zero-shot setting across question types. Figures 13-15 show the L-accuracy under CoT and BERTScore under the Zero-shot and CoT settings on PolyChartQA across question types.

A.8 Results across Chart Homogeneity on PolyChartQA

Figure 16 presents the detailed human evaluation result on human-authored QAs from PolyChartQA under the Zero-shot setting across chart homogeneity. Figures 17-20 show the L-accuracy and BERTScore under Zero-shot and CoT settings on PolyChartQA across chart homogeneity.

A.9 MLM-Performance based on the Number of Sub-charts

Figure 21 presents the detailed human evaluation result on human-authored QAs from PolyChartQA under the Zero-shot setting, grouped by the number

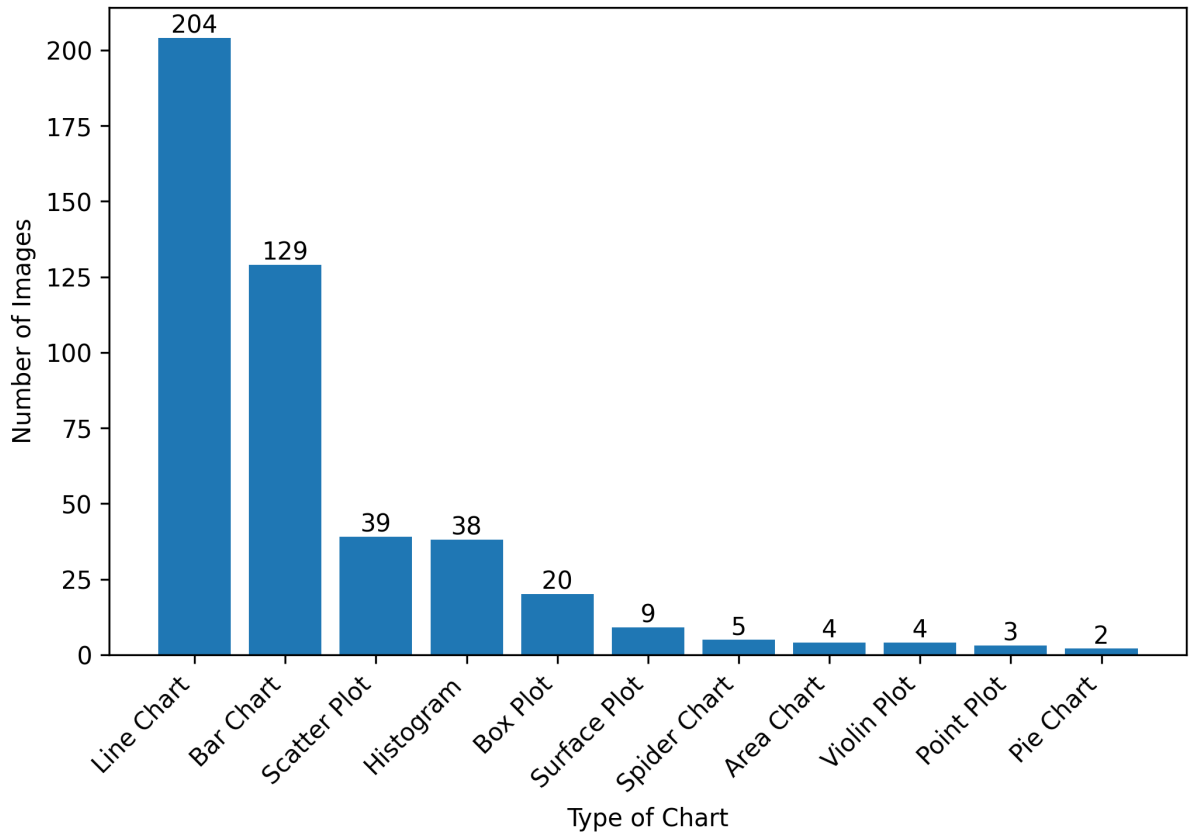


Figure 7: Distribution of chart types across homogeneous multi-chart images.

Judge	Claude-3.7-Sonnet	GPT-4.1	Gemini-2.0-flash	Gemma-3-4b-it	LLaMA-3.2-11B	LLaVA-1.5-7B	Pixtral-12B
Claude-3.7-Sonnet	93	96	95	96	95	98	95
GPT-4.1	94	89	93	94	92	99	93
Gemini-2.0-flash	90	87	90	81	84	93	92

Table 8: Percent agreement (%) between human evaluation and LLM judges across evaluated models. Columns represent the evaluated models, and rows indicate the LLM acting as the evaluation judge.

of sub-charts per image. Figures 22–25 present the L-Accuracy trends under the Zero-shot and CoT settings on PolyChartQA, grouped by the number of sub-charts per image.

A.10 Results on Human-authored vs MLM-generated PolyChartQA

Figure 26 shows BERTScore for both Zero-shot and CoT settings on human-authored and MLM-generated questions from PolyChartQA.

A.11 Performance Comparison of Zero-shot, CoT, and VDSP Prompting Strategies

Table 10 presents the BERTScore of Zero-shot, CoT, and VDSP on the human-authored questions. BERTScore decreases across all models for VDSP prompting, likely due to the longer, reasoning-heavy responses.

A.12 Performance Comparison of MLMs on Human-authored PolyChartQA with Human Evaluation

Table 11 presents H-Accuracy vs L-Accuracy on human-authored questions from PolyChartQA under the CoT setting.

A.13 Ablation Study: VDSP version 2

In this setting, the Stage 1 prompt was altered to convert each sub-chart into a structured table instead of a textual description, while Stages 2 and 3 remained unchanged. Table 12 presents the L-Accuracy Zero-shot, CoT, and VDSP version 2 on the human-authored questions. Figure 43 shows the prompt used in stage 1.

Model	Human-Authored QA		MLM-Generated QA	
	GPT-4.1 Judge	Gemini-2.0-flash Judge	GPT-4.1 Judge	Gemini-2.0-flash Judge
Claude-3.7-sonnet	0.692	0.719	0.789	0.815
GPT-4.1	0.632	0.674	0.897	0.908
Gemini-2.0-flash	0.721	0.748	0.851	0.865
Gemma-3-4b-it	0.241	0.326	0.340	0.463
LLaMA-3.2-11B	0.225	0.303	0.267	0.328
LLaVA1.5-7b	0.129	0.170	0.158	0.231
Pixtral-12B	0.534	0.563	0.692	0.720

Table 9: L-Accuracy comparison on Human-authored QA and MLM-generated QA, evaluated using GPT-4.1 and Gemini-2.0-flash judges.

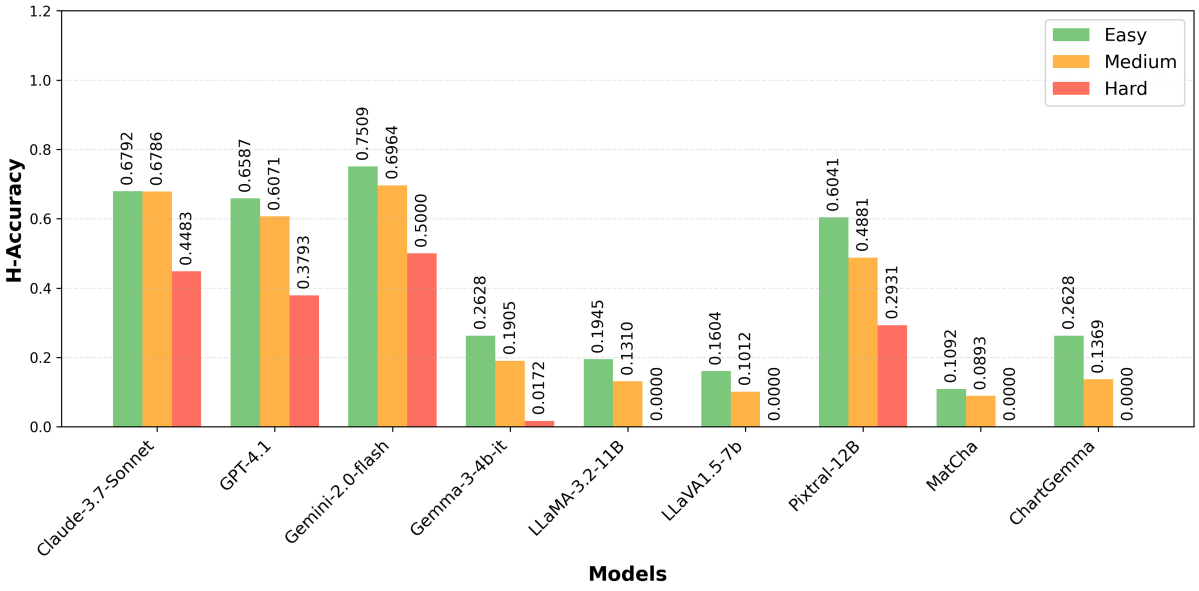


Figure 8: H-Accuracy on human-authored QAs from PolyChartQA under the Zero-shot setting across question difficulty levels.

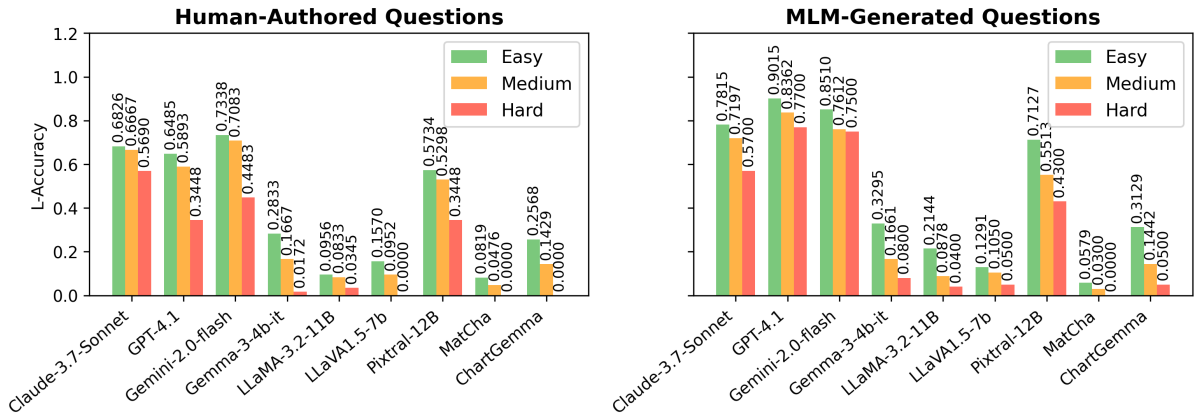


Figure 9: L-Accuracy on PolyChartQA under the chain-of-thought setting across question difficulty levels.

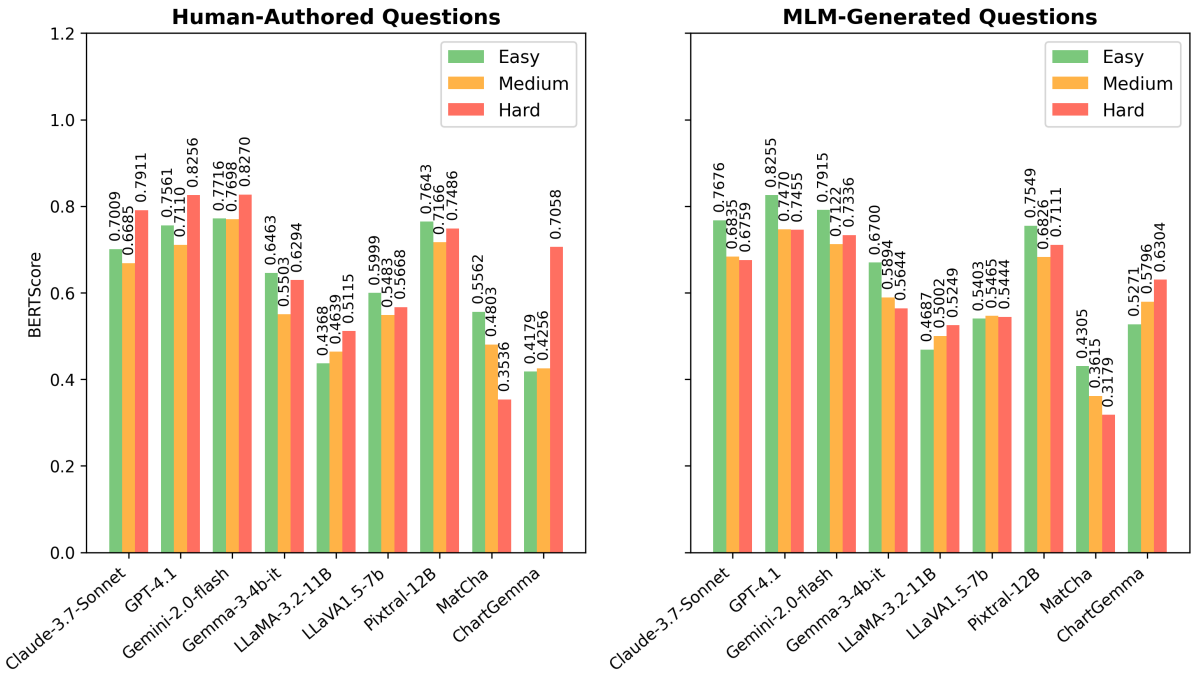


Figure 10: BERTScore on PolyChartQA under Zero-shot setting across question difficulty levels.

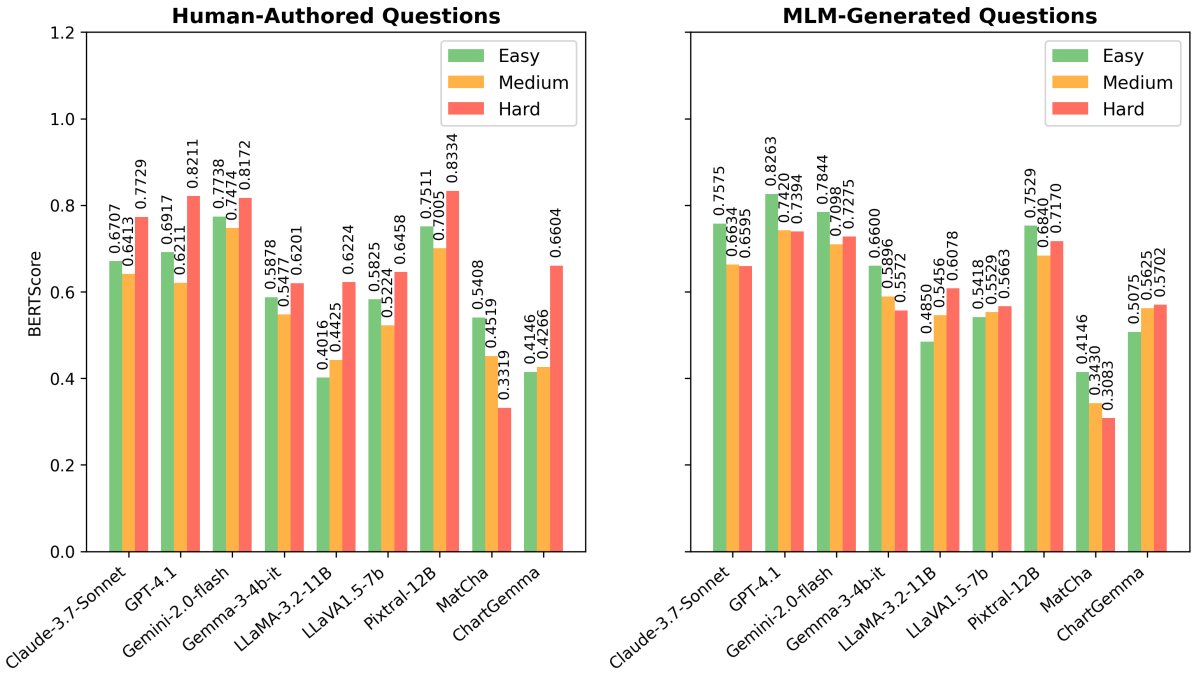


Figure 11: BERTScore on PolyChartQA under chain-of-thought setting across question difficulty levels.

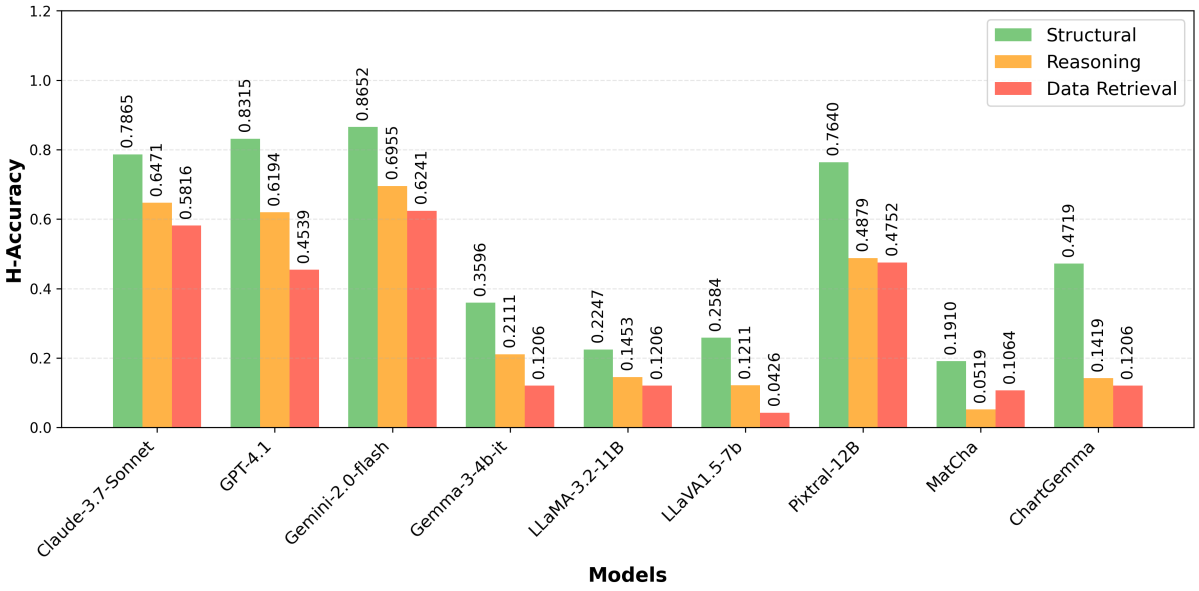


Figure 12: H-Accuracy on human-authored QAs from PolyChartQA under Zero-shot setting across question types.

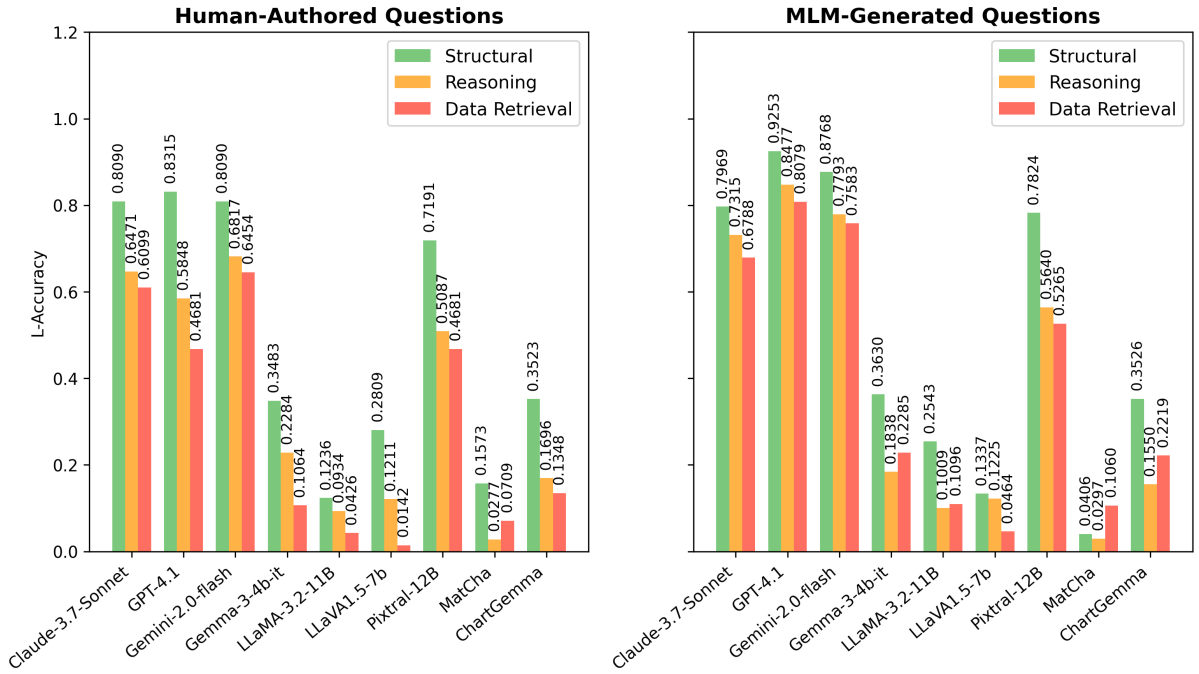


Figure 13: L-Accuracy on PolyChartQA under CoT setting across question types.

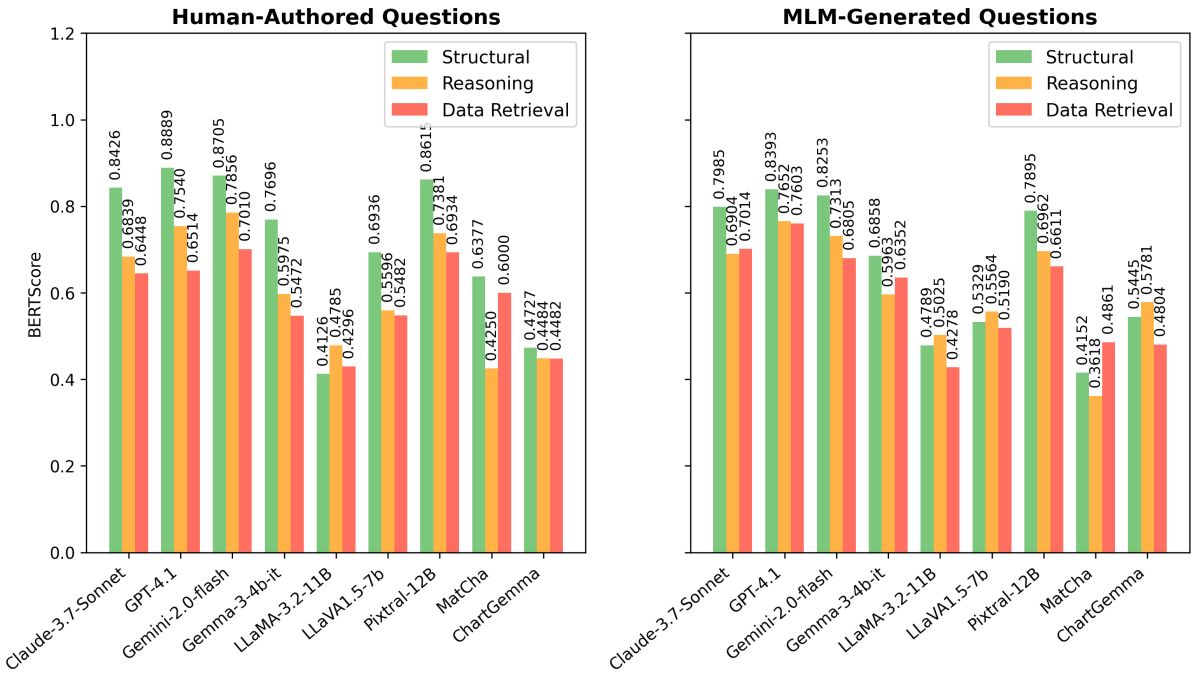


Figure 14: BERTScore on PolyChartQA under Zero-shot setting across question types.

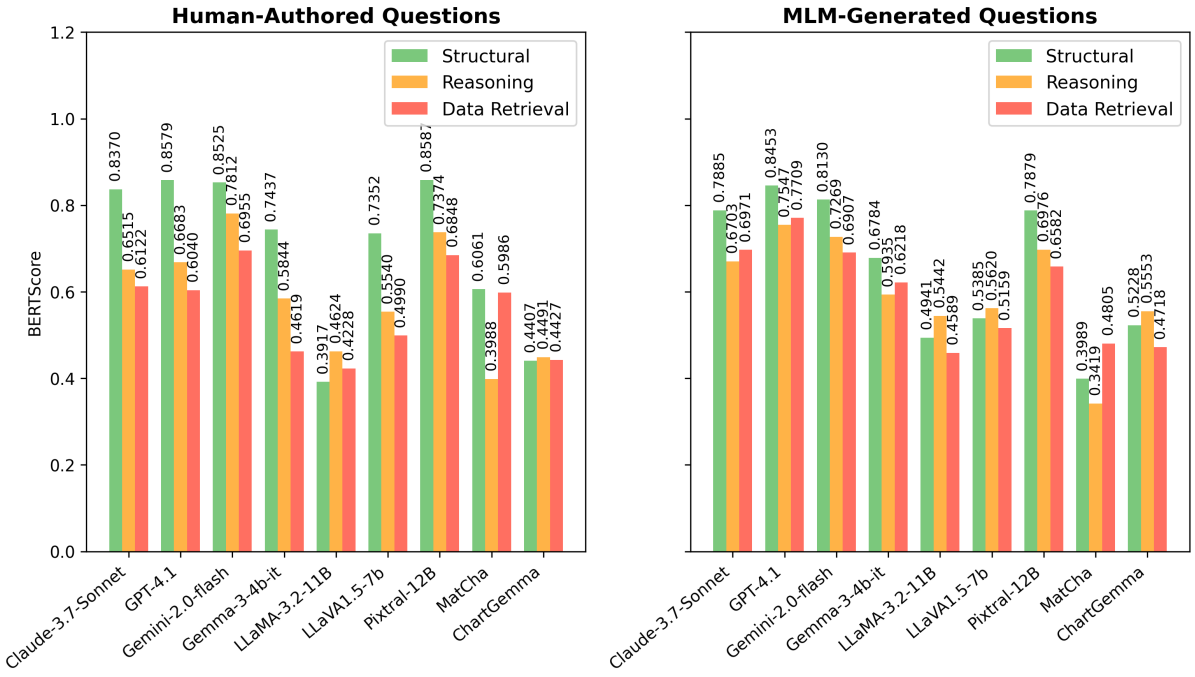


Figure 15: BERTScore on PolyChartQA under CoT setting across question types.

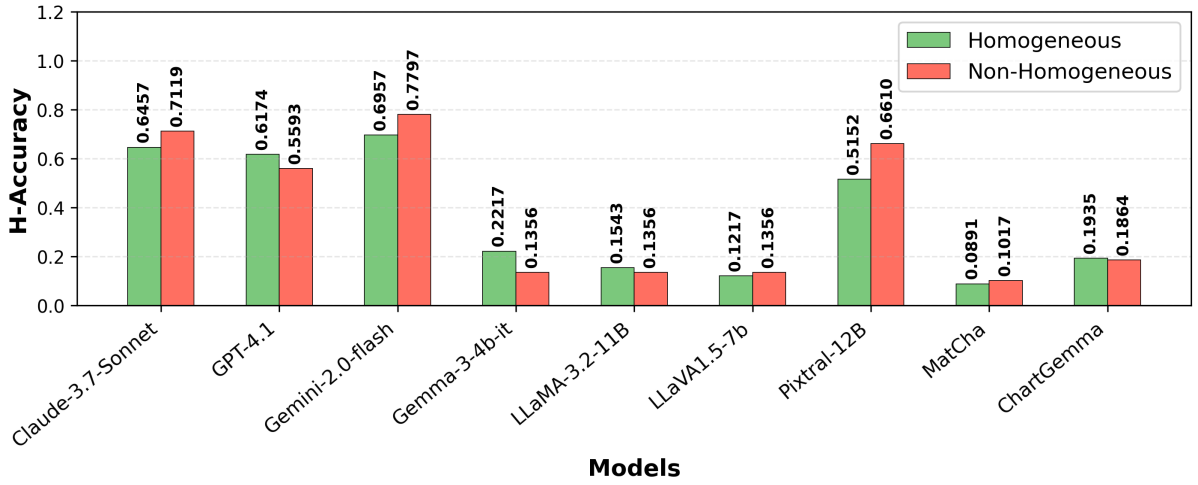


Figure 16: H-Accuracy on human-authored QAs from PolyChartQA under Zero-shot setting across chart homogeneity.

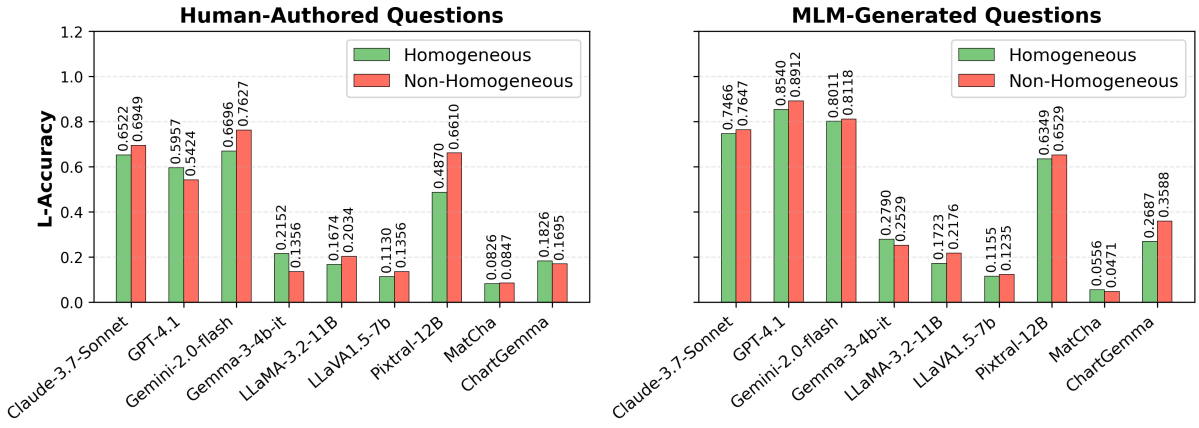


Figure 17: L-Accuracy on PolyChartQA under Zero-shot setting across chart homogeneity.

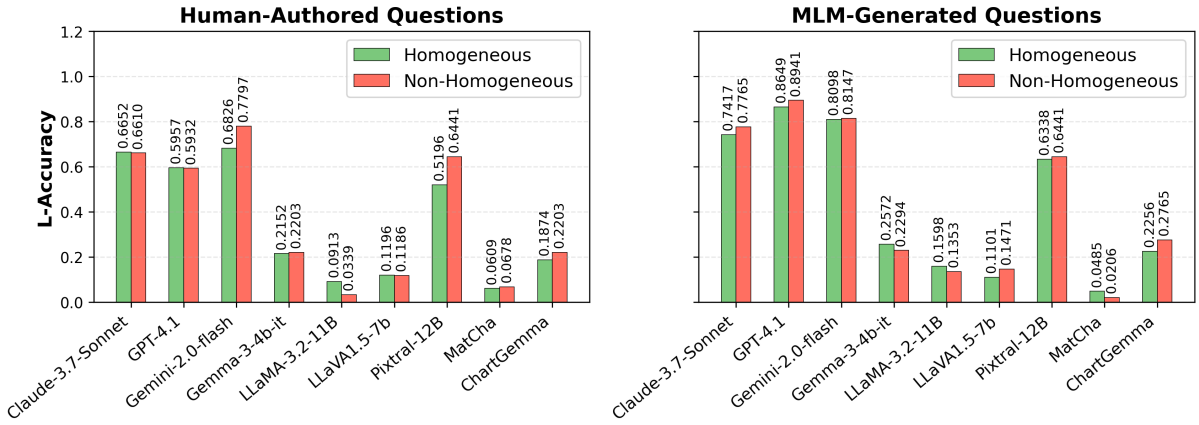


Figure 18: L-Accuracy on PolyChartQA under CoT setting across chart homogeneity.

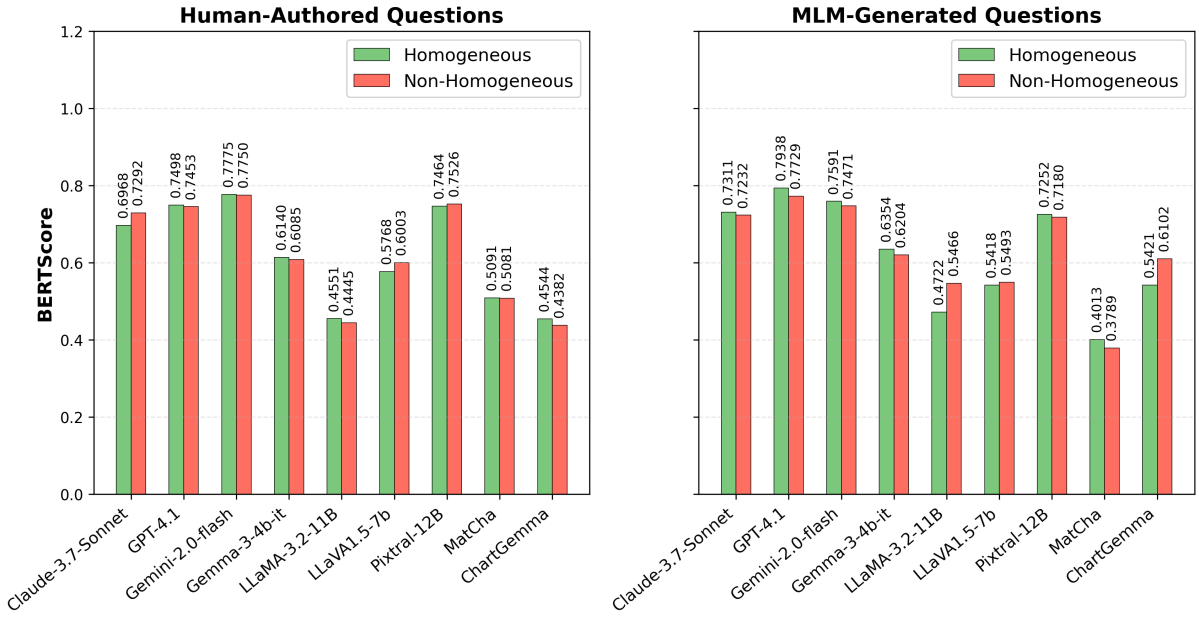


Figure 19: BERTScore on PolyChartQA under Zero-shot setting across chart homogeneity.

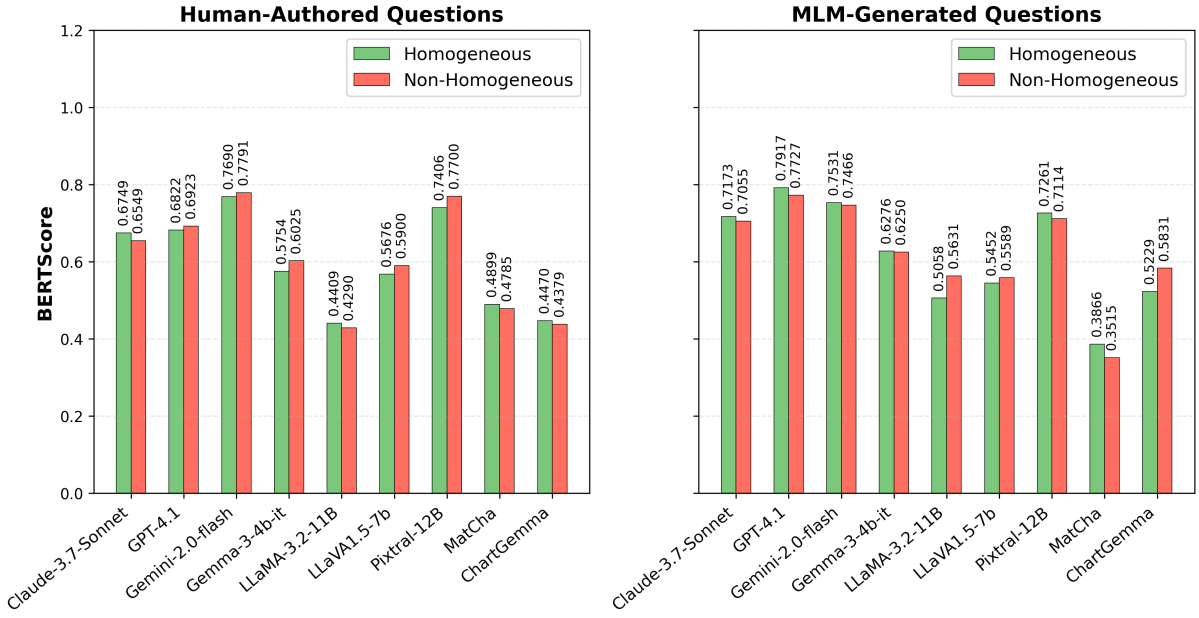


Figure 20: BERTScore on PolyChartQA under CoT setting across chart homogeneity.

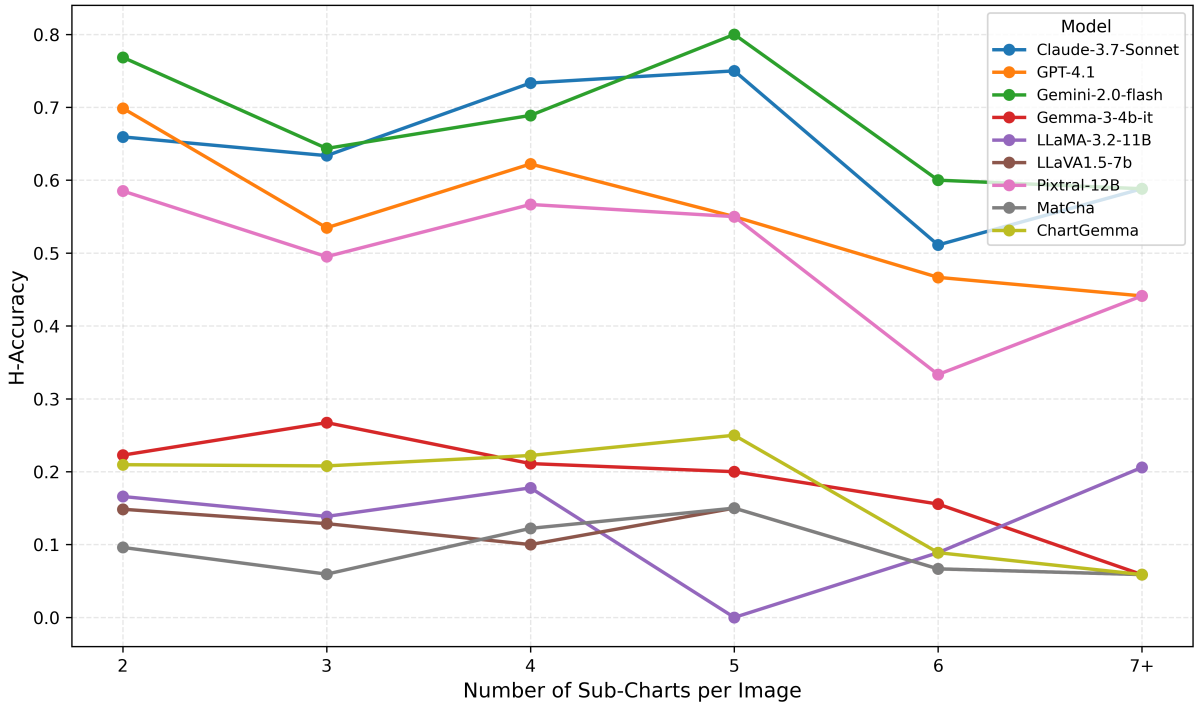


Figure 21: H-Accuracy on Human-authored PolyChartQA under Zero-shot setting based on the number of sub-charts in an image.

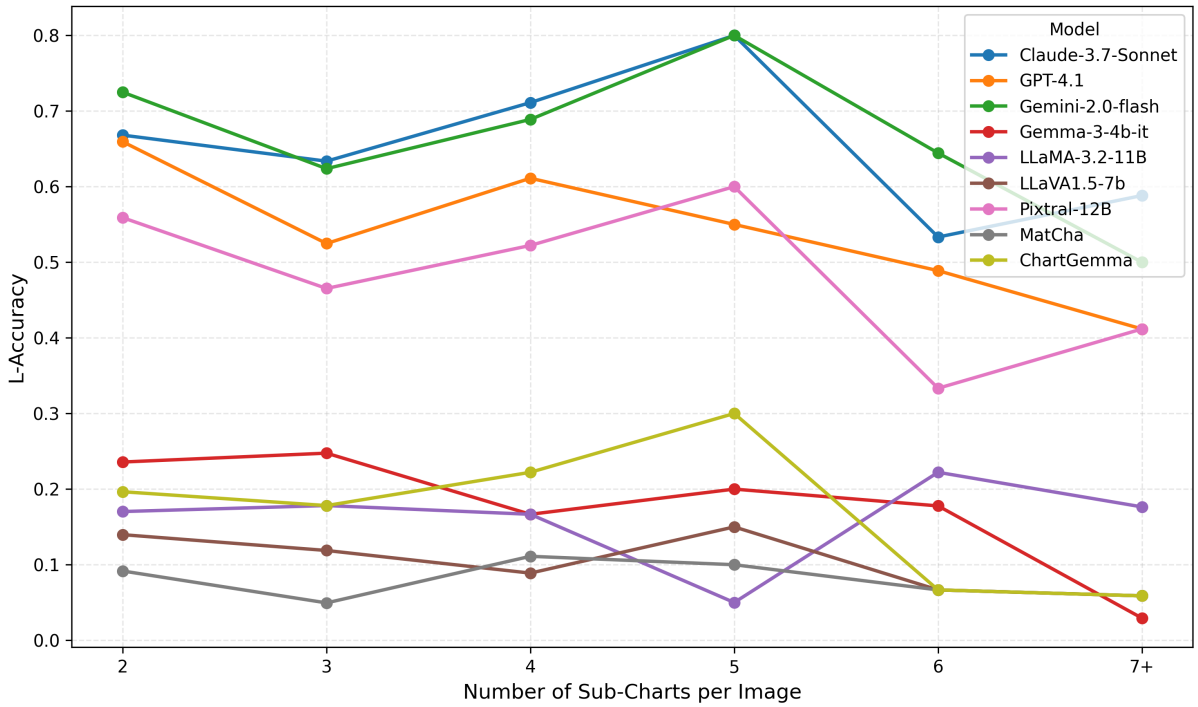


Figure 22: L-Accuracy on Human-authored PolyChartQA under Zero-shot setting based on the number of sub-charts in an image.

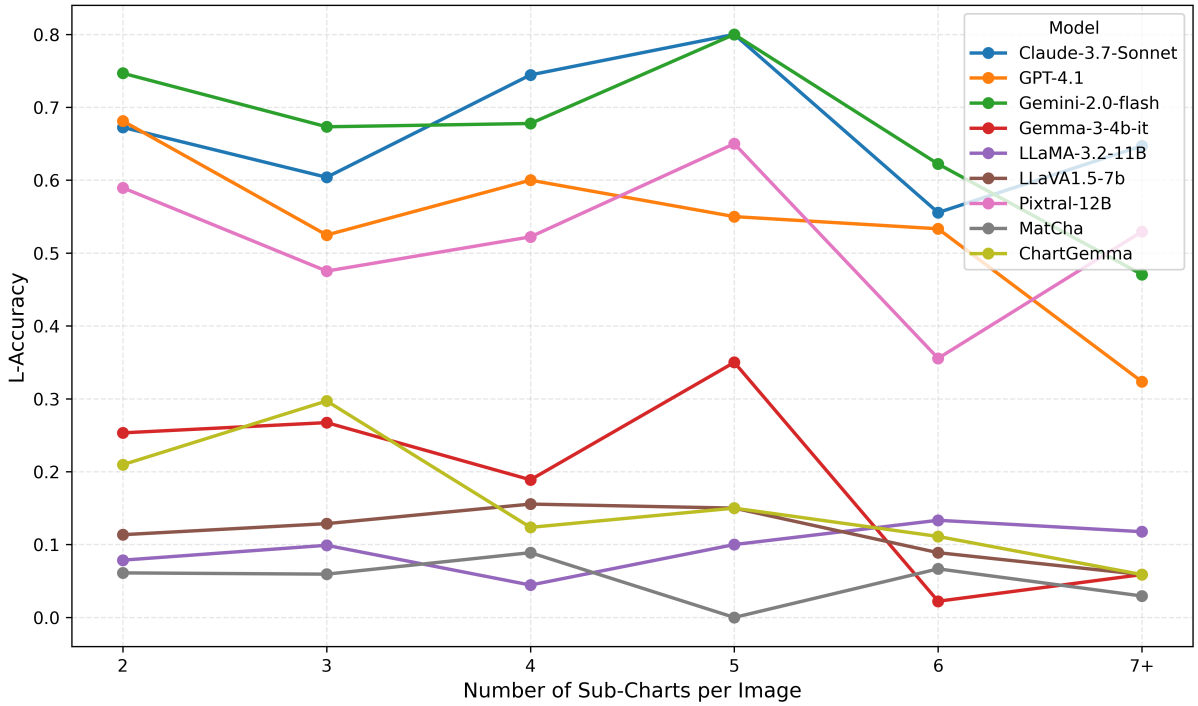


Figure 23: L-Accuracy on Human-authored PolyChartQA under CoT setting based on the number of sub-charts in an image.

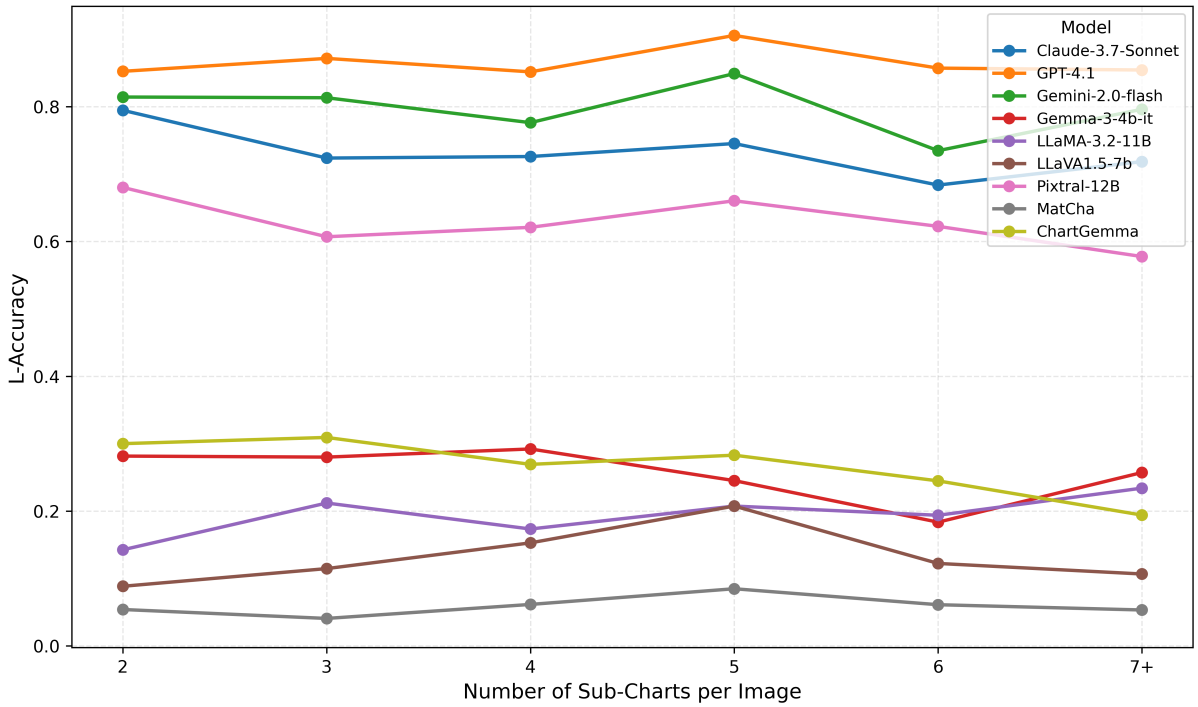


Figure 24: L-Accuracy on MLM-generated PolyChartQA under Zero-shot setting based on the number of sub-charts in an image.

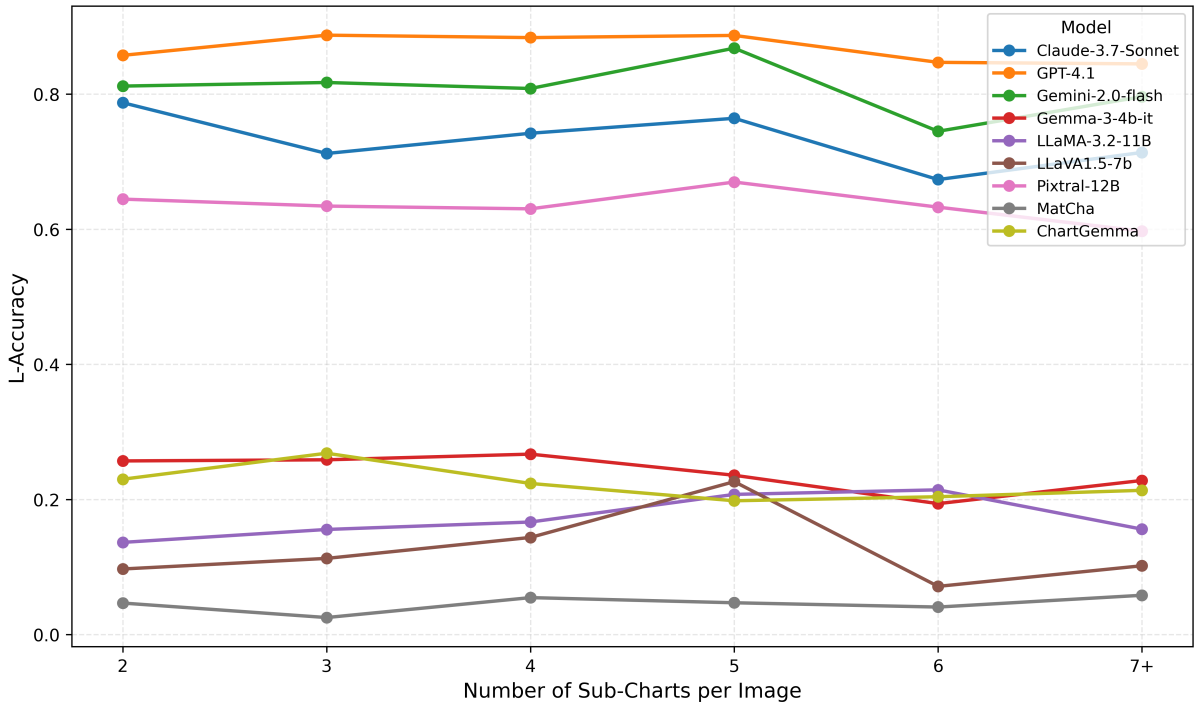


Figure 25: L-Accuracy on MLM-generated PolyChartQA under CoT setting based on the number of sub-charts in an image.

Model	Zero-shot	CoT	VDSP
Claude-3.7-Sonnet	0.7298	0.6726	0.5046
GPT-4.1	**0.7905**	0.6833	0.6033
Gemini-2.0-flash	0.7572	0.7701	0.6812
Pixtral-12B	0.7240	0.7434	0.4591

Table 10: BERTScore of Zero-shot, CoT, and VDSP prompting methods on human-authored QA from PolyChartQA.

Model	H-Accuracy	L-Accuracy	Difference
Claude-3.7-Sonnet	0.6609	0.6647	0.0038
GPT-4.1	0.5954	0.5954	0.0000
Gemini-2.0-flash	**0.7187**	**0.6936**	0.0251
Pixtral-12B	0.5376	0.5337	0.0039
LLaMA-3.2-11B-Vision	0.0809	0.0848	0.0039
LLaVA1.5-7b	0.1175	0.1195	0.0020
Gemma-3-4b-it	0.2274	0.2158	0.0116
MatCha	0.0617	0.0617	0.0000
ChartGemma	0.2351	0.1911	**0.0440**

Table 11: H-Accuracy vs L-Accuracy using CoT prompting on the human-authored PolyChartQA.

Model	Zero-shot	CoT	VDSP v2
GPT-4.1	0.5896	0.5954	0.5742
Claude-3.7-Sonnet	0.6570	0.6647	0.6686
Gemini-2.0-flash	0.6802	**0.6936**	**0.6936**
Pixtral-12B	0.5067	0.5337	0.4547

Table 12: L-Accuracy of Zero-shot, CoT, and VDSP version 2 prompting methods on the human-authored QA from PolyChartQA.

A.14 Ablation Study: VDSP version 3

In this variant, all three stages were modified, where Stage 2 followed a dual-persona design—an Analyst generating an answer and a Reviewer verifying it—while Stages 1 and 3 retained tasks similar to the original VDSP setting. Table 13 presents the L-Accuracy Zero-shot, CoT, and VDSP version 3 on the human-authored questions. Figure 44 shows the prompt used in stage 2.

Model	Zero-shot	CoT	VDSP v3
GPT-4.1	0.5896	0.5954	0.5934
Claude-3.7-Sonnet	0.6570	0.6647	0.6763
Gemini-2.0-flash	0.6802	**0.6936**	0.6801
Pixtral-12B	0.5067	0.5337	0.4798

Table 13: L-Accuracy of Zero-shot, CoT, and VDSP version 3 prompting methods on the human-authored QA from PolyChartQA.

A.15 Computational Overhead of VDSP

As a multi-stage prompting strategy, VDSP requires multiple model calls, resulting in higher token consumption and inference latency than single-pass prompting. This introduces a three-way trade-off among interpretability, cost, and accuracy. By transforming a single prediction into a traceable reasoning pipeline, VDSP is particularly valuable

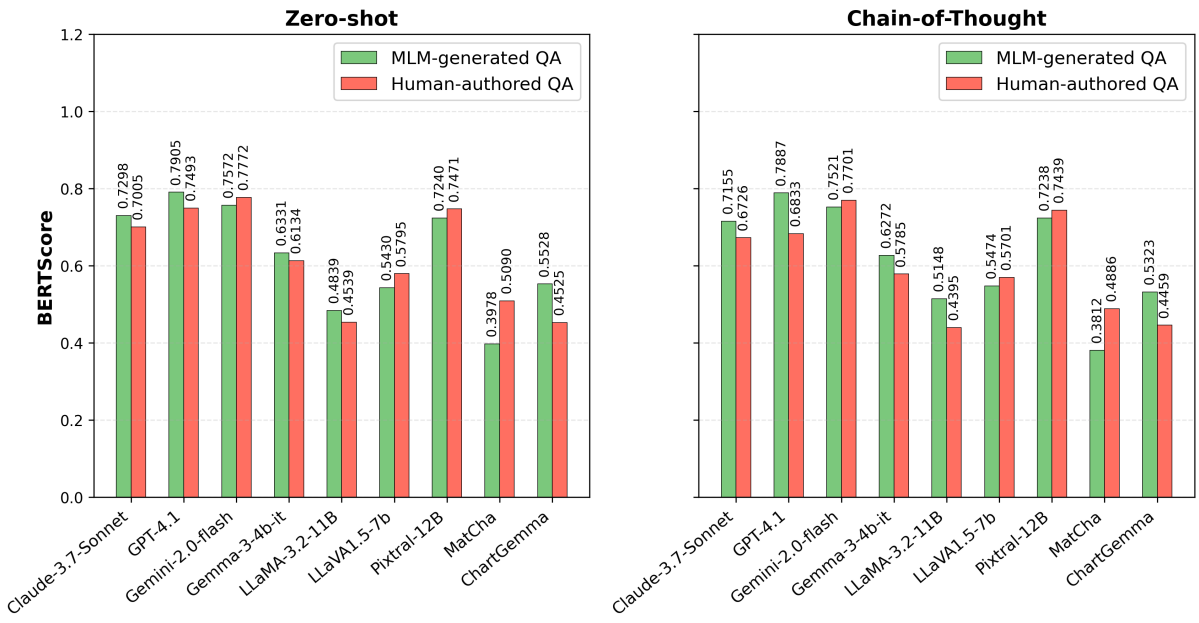


Figure 26: BERTScore on the human-authored vs MLM-generated questions from PolyChartQA under Zero-shot and CoT settings.

in analytical and high-stakes settings where understanding the model's decision process is as important as the final answer. While single-pass prompting may be preferable in latency- and cost-sensitive applications, VDSP provides structured intermediate reasoning, improving transparency and enabling error diagnosis in complex multi-chart tasks.

A.16 Human-Authored Question-Answer Examples

Figures 27 - 29 show examples of human-authored question-answer pairs from PolyChartQA.

A.17 MLM-generated Question-Answer Examples

Figures 30-34 show examples of MLM-generated question-answer pairs from PolyChartQA.

A.18 Prompt Examples

Figures 35-42 show the prompts used in this experiment.

A.19 Example of VDSP's Intermediate Steps Output

Figures 45–46 present the multi-chart image, the question, and the intermediate outputs along with the final answer generated by GPT-4.1 using VDSP.

A.20 Error Examples

Figure 47 shows example question where all the models were wrong.

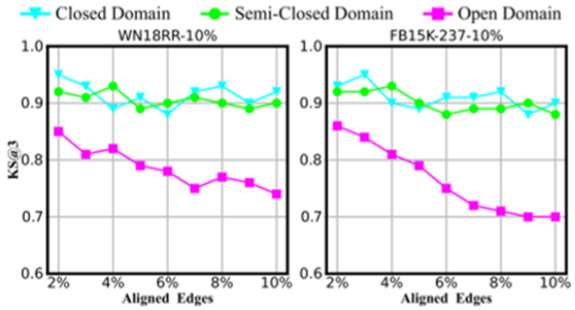


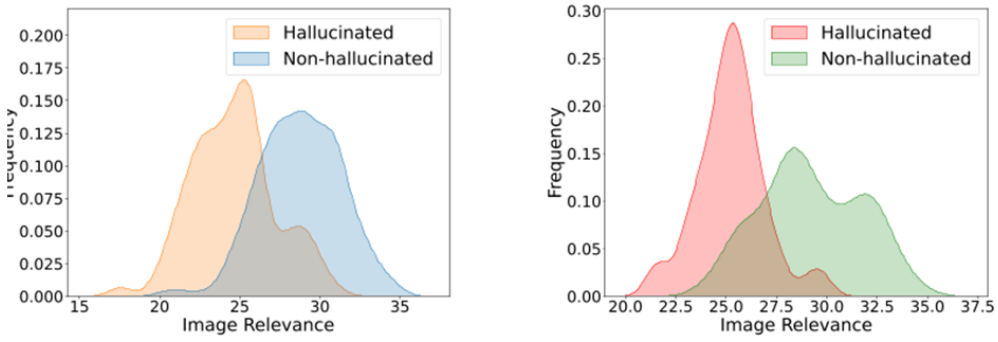
Figure 3: Impacts of the number of aligned edges on the stability of the three knowledge alignment strategies.

Easy-Data Retrieval: Q: What is the KS@3 score of the Open-Domain strategy at 10% aligned edges on the FB15K-237 dataset? Ans: 0.7

Medium-Data Retrieval: Q: What is the difference between the KS@3 score of the Semi-Closed Domain strategy on the WN18RR dataset and the Open-Domain strategy on the FB15K-237 dataset at 10% aligned edges? Ans: 0.2

Easy-Reasoning: Q: How does the trend of the KS@3 score for the Open-Domain strategy change with the increase in aligned edges on the FB15K-237 dataset? Ans: The KS@3 score for the Open-Domain strategy decreases with the increase in aligned edges on the FB15K-237 dataset.

Figure 27: Example human-authored Easy-Data Retrieval, Medium-Data Retrieval, and Easy-Reasoning QA pairs from PolyChartQA. The multi-chart is collected from Chen et al. (2024).



(a) MiniGPT4 (b) LLaVA

Figure 2: Comparison of hallucinated and non-hallucinated objects generated by MiniGPT4 (a) and LLaVA (b) on image-relevance scores.

Hard-Structural: Q: Which type of object is represented by the orange color in objects generated by MiniGPT4, and which color is used to represent the same object in objects generated by LLaVa? Ans: Hallucinated object is represented by the orange color in objects generated by MiniGPT4, and red color is used to represent it in objects generated by LLaVa.

Figure 28: Example human-authored Hard-Structural QA pairs from PolyChartQA. The multi-chart is from Xing et al. (2024).

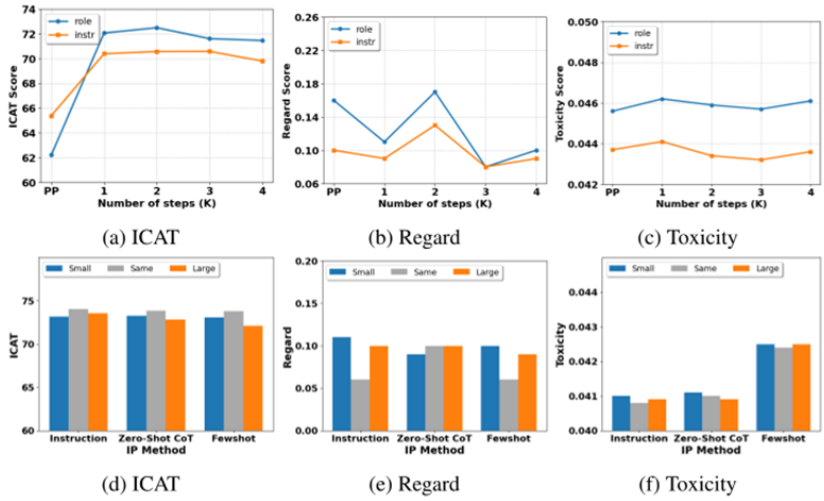


Figure 1: Fig. (a), (b), and (c) show performance upon varying number of refinement steps on ICAT, Regard and Toxicity. Fig. (d), (e), (f) show performance upon varying the size of the implication generation model.

Hard-Data Retrieval: Q: Which condition has a toxicity score of 0.044 at 1 step, and what is its regard score at PP? Ans: instr has a toxicity score of 0.044 at 1 step and its regard score at PP is 0.10.

Figure 29: Example human-authored Hard-Data Retrieval QA pairs from PolyChartQA. The multi-chart is from Furniturewala et al. (2024).

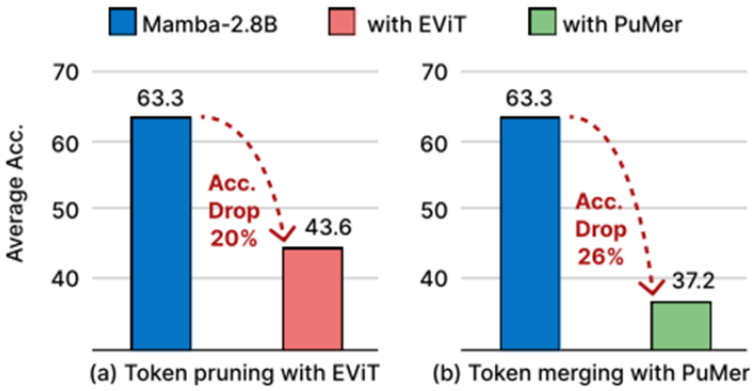


Figure 1: Performance of applying token pruning (EViT) and merging (PuMer) methods on Mamba-2.8B, showcasing significant drop in accuracy.

Easy-Structural: Q: What is the color used to represent the Mamba-2.8B baseline in both subplots? Ans: Blue

Easy-Data Retrieval: Q: What is the average accuracy value for token pruning with EViT? Ans: 43.6

Medium-Data Retrieval: Q: By how much does the average accuracy decrease when using token merging with PuMer compared to token pruning with EViT? Ans: The average accuracy decreases by 6.4 (from 43.6 to 37.2).

Medium Reasoning: Q: Which method, EViT or PuMer, results in a larger percentage drop in accuracy compared to the Mamba-2.8B baseline? Ans: PuMer results in a larger percentage drop (26%) compared to EViT (20%).

Figure 30: Example MLM-generated Easy-Structural, Easy-Data Retrieval, Medium-Data Retrieval, and Medium-Reasoning QA pairs from PolyChartQA. The multi-chart is from Zhan et al. (2024).

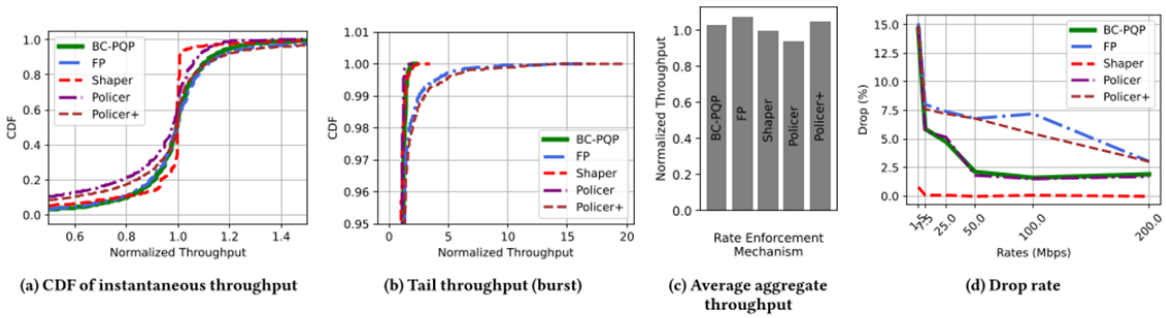
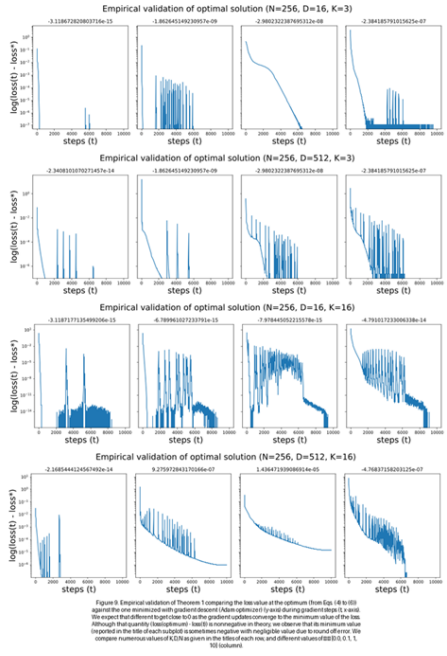


Figure 6: Aggregate rate enforced by BC-PQP and other baselines, 6a and 6b show distribution of aggregate throughput measured over 250 ms windows normalized by enforced rate, 6c shows normalized average aggregate throughput and 6d shows drop rate at different enforced rates

Easy-Reasoning : Q: Compare the average aggregate throughput of FP and Policer+ mechanisms. Which one is higher? Ans: FP has a higher average aggregate throughput than Policer+.

Figure 31: Example MLM-generated Easy-Reasoning QA pair from PolyChartQA. The multi-chart is from Tahir et al. (2024).



Medium-Structural: Q: Which row of subplots corresponds to the case where D = 512? Ans: The second and fourth rows

Figure 32: Example MLM-generated Medium-Structural QA pair from PolyChartQA. The multi-chart is from Balestriero and Lecun (2024).

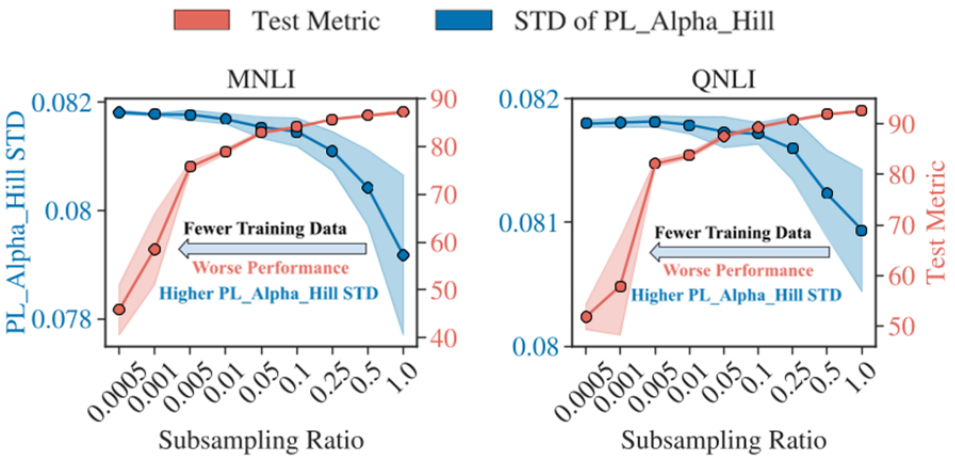


Figure 2: Test performance and STD of PL_Alpha_Hill across all layers of RoBERTa-base model trained on MNLI (Accuracy⇧) and QNLI (Accuracy⇧) under different subsampling ratios.

Hard-Data Retrieval: Q: Identify the subsampling ratio at which both datasets achieve their highest Test Metric, and describe how the PL_Alpha_Hill STD behaves at that point in each dataset.
Ans: Both datasets achieve their highest Test Metric at a subsampling ratio of 1.0. At this point, PL_Alpha_Hill STD is at its lowest value in both datasets.

Figure 33: Example MLM-generated Hard-Data Retrieval QA pairs from PolyChartQA. The multi-chart is from Liu et al. (2024c).

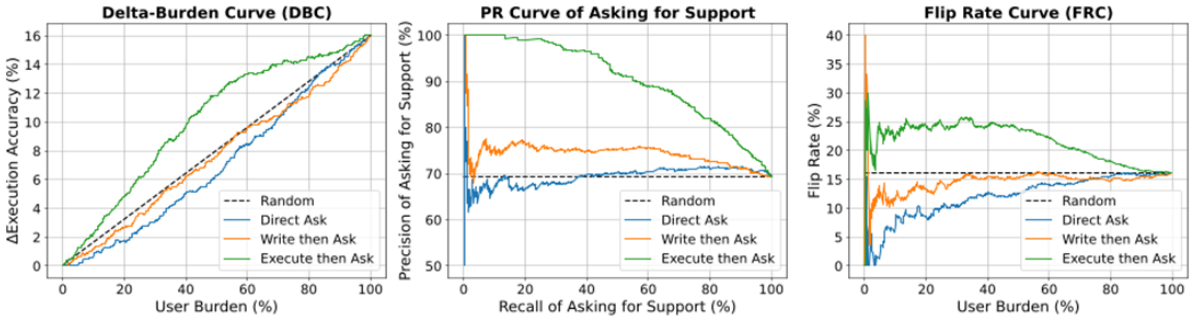


Figure 6: Performance curves of gpt-3.5-turbo-0125.

Hard-Reasoning: Q: Which method shows the largest improvement in execution accuracy over the random baseline as user burden increases, and how is this reflected in its precision of asking for support?
Ans: 'Execute then Ask' shows the largest improvement in execution accuracy over the random baseline, and this is reflected by its high and sustained precision of asking for support across most of the recall range.

Figure 34: Example MLM-generated Hard-Reasoning QA pairs from PolyChartQA. The multi-chart is from Wu et al. (2024a).

```
You are given an image of a multi-chart figure, typically extracted from a scientific paper. This
figure may contain several sub-figures arranged in a composite layout.
            Your task is to analyze the visual content and perform the following steps:
            1. Count the total number of sub-figures that are charts or plots.
            2. Classify each chart into one of the following chart types:
               - Bar Chart, Line Chart, Scatter Plot, Histogram, Box Plot, Pie Chart, Dot Plot, Pareto
Chart, Violin Plot, Area Chart, Hexbin Plot, Spider Chart.
            3. If a chart does not fall under the types listed above but is still a chart or plot,
classify it using an appropriate chart type.
            4. If any sub-figure is not a chart or plot (e.g., a diagram, architecture image,
illustration, or table), classify it as "No Chart-plot".
            5. Determine if all the charts/plots are of the same type. If yes, mark Homogeneous as
"Yes", otherwise "No".

            Output your answer strictly in the following JSON format:
            {{
            "Count": "<number of sub-figures that are charts/plots>",
            "Homogeneous": "<Yes or No>",
            "Chart types": "<Comma-separated list of unique chart types>"
            }}
            Be concise and accurate. Do not include explanations or reasoning—only return the JSON
object.
<Image>
```

Figure 35: Prompt used to annotate multi-chart images.

```
You are given an image containing multiple charts/plots, typically extracted from scientific papers.
Your task is to generate high-quality question-answer pairs based on the visual content of the multi-
chart figure.
There are three types of questions you must generate:
- Structural: Focuses on visual or structural elements (e.g., labels, colors, layout, axis names,
symbols).
- Data Retrieval: Focuses on retrieving explicit numeric or categorical values (e.g., reading data
points, extracting counts).
- Reasoning: Requires inference, comparison, aggregation, trend analysis, or contradiction detection
across charts.
Additionally, categorize each question into one of three difficulty levels based on how many charts and
what type of reasoning is involved:
- Easy:
  - Requires inspection of a **single chart only**.
  - The value or visual feature is **directly visible**.
  - Simple trend or comparison **within one chart**.
  - No need to look at any other chart to answer.
- Medium:
  - Involves **multiple charts**.
  - Requires **comparing**, **aggregating**, or **combining information** from two or more charts.
  - Information is spread across charts but reasoning is moderately direct.
- Hard:
  - Requires **multi-step reasoning** across multiple charts.
  - This can include:
    - **Multi-step Structural tasks** (e.g., extract a color label from one chart, use it to find a
symbol in another chart).
    - **Multi-step Data Retrieval tasks** (e.g., retrieve a value from one chart, then use it to infer
or retrieve another value from another chart).
    - **Complex Reasoning tasks** (e.g., contradiction detection, trend reversal, anomaly detection
across figures).
Important:
- If answering a question **requires looking at more than one chart**, it must be classified as **at
least Medium or Hard**.
- Easy questions must be fully answerable using **only one chart**.
---
For each question you generate, provide the following information:
- Question: <Write the question text>
- Answer: <Write the correct answer>
- Difficulty: <Easy/Medium/Hard>
- Question Type: <Structural/Data Retrieval/Reasoning>
- Relevant Charts: <Comma-separated list of relevant sub-chart identifiers needed to answer the
question (e.g., Chart_1, Chart_2)>
---
Output your result strictly in the following JSON format:

{{
  "QA1": {{
    "Question": "<Question>",
    "Answer": "<Answer>",
    "Difficulty": "Easy/Medium/Hard",
    "Question Type": "Structural/Data Retrieval/Reasoning",
    "Relevant Charts": "<Relevant chart identifiers>"
  }},
  "QA2": {{
    "Question": "<Question>",
    "Answer": "<Answer>",
    "Difficulty": "Easy/Medium/Hard",
    "Question Type": "Structural/Data Retrieval/Reasoning",
    "Relevant Charts": "<Relevant chart identifiers>"
  }},
  ...
}}
---
Additional Instructions:
- Generate a **balanced set** of questions across difficulty levels: approximately **one-third Easy,
one-third Medium, and one-third Hard**.
- Ensure coverage across all three question types (Structural, Data Retrieval, Reasoning).
- Prefer diversity in cognitive demands: include simple retrieval, multi-step reasoning, and cross-
subchart synthesis.
- Avoid yes/no questions.
- Do not mention specific chart names or chart numbers in the question (the model answering must infer
relevant chart(s) from question content).
- Ensure the answer is grounded in the visual evidence, not speculative.
- Keep answers concise but complete. Use short factual phrases or full sentences as needed.
- Only output the JSON object exactly as specified, with no extra explanations.
<Image>
```

Figure 36: Prompt used to generate question-answers using MLM.

```
Given the following image and question, provide the answer based only on the visual content of the
image.
            Question: {question}
            Output the answer strictly in this JSON format:

            {{
              "Answer": "<Answer>"
            }}
<Image>
```

Figure 37: Zero-shot prompt used to evaluate MLM on multi-chart images.

```
Given the following image and question, provide the answer based only on the visual content of the
image.
            Question: {question}
            Answer: Let's think step by step.
            Output the answer strictly in this JSON format:

            {{
              "Answer": "<Answer>"
            }}
<Image>
```

Figure 38: CoT prompt used to evaluate MLMs on multi-chart images.

```
You are given an image that contains multiple charts, and a question about that image.
        Question: {question}
        Task:
            1. Identify and briefly describe each individual chart.
                1.1. State its chart type (e.g., bar chart, line chart, pie chart, box plot, spider
chart, etc.).
                1.2. Describe key features like x/y-axis labels, units, data encodings (e.g., color,
shape, marker type, legend).
            2. Then, decompose the question into smaller reasoning steps needed to answer it.
                2.1 Mention which sub-chart(s) and which visual elements (e.g., bars, lines, box, axes)
are relevant to each step.
        Output the chart descriptions and the reasoning plan in plain text strictly in this JSON
format:

            {{
              "Answer": "<Plain Text>"
            }}
<Image>
```

Figure 39: VDSP Stage 1 prompt used to evaluate MLM performance on multi-chart images.

```
You are given:
            – An image with multiple charts
            – A natural-language question about the image
            – A set of chart descriptions and a step-by-step reasoning plan
        Question: {question}
        Chart Description and Step by Step Reasoning Plan : {description}
        Task:
        Using that information, solve the problem step by step.
            - For each step, clearly state:
            - what you are trying to do
            - what values you extract from which chart
            - how you use those values (e.g., comparisons, calculations, ordering)
        Finally, combine all step outputs to produce a complete and concise answer to the original
question.
        Return your full reasoning and final answer in plain text.
        Output strictly in this JSON format:

            {{
              "Answer": "<plain text>"
            }}
<Image>
```

Figure 40: VDSP Stage 2 prompt used to evaluate MLM performance on multi-chart images.

```
You are given:
            – An image with multiple charts
            – A natural-language question
            – A complete reasoning process and a proposed answer
        Question: {question}
        Reasoning process and a proposed answer: {reasoning_and_ans}
        Task:
        Carefully re-evaluate the entire reasoning process. Ask yourself:
        - Did I select the correct chart(s) for each step?
        - Did I misread any axis, bar height, legend label, or unit?
        - Did I miss any relevant information?
        If the answer is incorrect, revise it concisely.
    Output the final answer strictly in this JSON format:

    {{
        "Answer": "<Answer>"
    }}
<Image>
```

Figure 41: VDSP Stage 3 prompt used to evaluate MLM performance on multi-chart images.

```
You are given a question , ground - truth answer , and a candidate answer.
            Question : {question}
            Ground - truth answer : {ground_truth}
            Candidate answer : {generated_answer}
            Is the semantic meaning of the ground - truth and candidate answers
            similar ? Answer in one word - Yes or No .
```

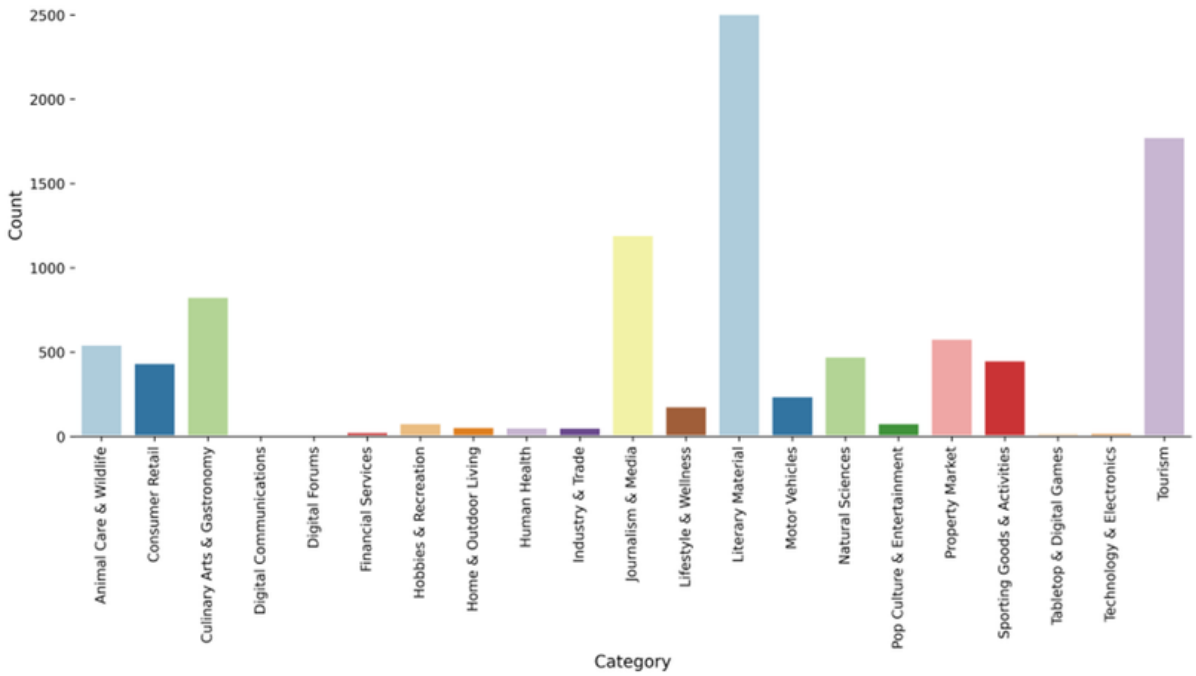
Figure 42: Prompt to compare the ground truth and an MLM response to derive L-Accuracy.

```
You are given an image that contains multiple charts, and a natural-language question.
        Question: {question}
        **Task**:
        1. For each sub-chart:
           - Assign an ID (C1, C2, ...).
           - Identify chart type (bar, line, pie, etc.).
           - Convert visible data into a structured **table** (rows for categories, columns for values,
include units).
           - If question involves encodings (color, shape, marker), add those fields.
        2. Decompose the question into a step-by-step **reasoning plan** that references chart IDs and
table fields.

        Output strictly in this JSON format:

        {{
          "charts": [
            {{
              "id": "C1",
              "type": "bar",
              "x_axis": "Year",
              "y_axis": "Population (millions)",
              "table": [
                {{"year": "2018", "population": 50}},
                {{"year": "2019", "population": 55}}
              ]
            }}
          ],
          "reasoning_plan": [
            "Step 1 ...",
            "Step 2 ..."
          ]
        }}
```

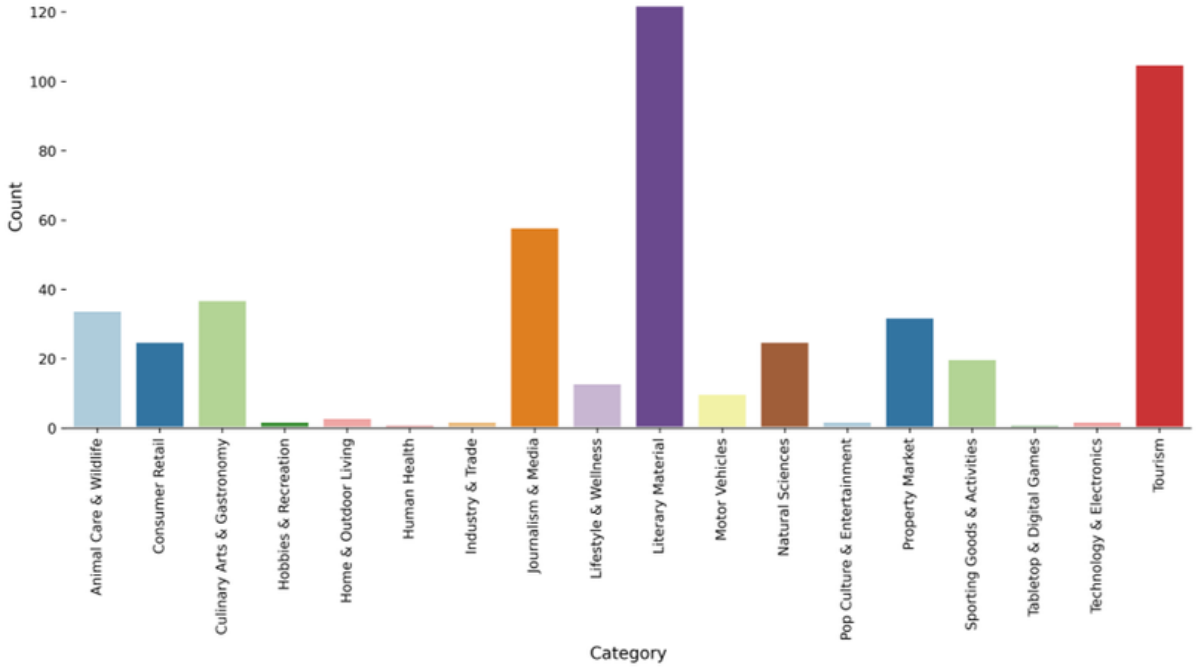
Figure 43: VDSP version 2 Stage 1 prompt used to evaluate MLM performance on multi-chart images.

```
You are given:
                - An image with multiple charts
                - A natural-language question
                Question: {question}
                - A descriptive summary from Step 1 (soft tables + meta-reasoning):
                {description}

                **Task**:
                Act in TWO roles:

                **Analyst**:
                - Follow the reasoning plan step by step.
                - Use the descriptive soft tables to extract relevant values or relationships.
                - Perform necessary comparisons or calculations.
                - Propose an answer.

                **Reviewer**:
                - Double-check the Analyst's reasoning.
                - Verify if the right chart(s) were used, if values were extracted correctly, if units
and logic are consistent.
                - If Analyst is correct, confirm. If errors exist, revise the answer concisely.

                Output strictly in this JSON format:
                {{
                  "Answer": "<Plain Text Reasoning Trace + Final Proposed Answer>"
                }}
```

Figure 44: VDSP version 3 Stage 2 prompt used to evaluate MLM performance on multi-chart images.



(a) IIW-Train Set Image Category Distribution



(b) IIW-Eval Set Image Category Distribution

Figure 7: Image Category Distribution for the IIW Dataset's Train and Eval Splits.

Figure 45: The multi-chart is from Garg et al. (2024).

```
Question: What is the count of Animal care & Wildlife in train set images?

Stage 1. Chart Structuring and Question Decomposition Output:
1. Chart Descriptions:
 1.1. The first chart (top) is a vertical bar chart titled 'IIW-Train Set Image Category
 Distribution'.
  - X-axis: 'Category' (various categories such as Animal Care & Wildlife, Consumer Retail, etc.)

  - Y-axis: 'Count' (number of images, unitless)

  - Data encoding: Each bar represents the count of images in a specific category. Bars are colored
  differently for each category. No explicit legend, but categories are labeled on the x-axis.

 1.2. The second chart (bottom) is a vertical bar chart titled 'IIW-Eval Set Image Category
 Distribution'.
  - X-axis: 'Category' (same categories as above)

  - Y-axis: 'Count' (number of images, unitless)

  - Data encoding: Each bar represents the count of images in a specific category. Bars are colored
  differently for each category. No explicit legend, but categories are labeled on the x-axis.

2. Reasoning Plan:
 2.1. Identify which chart corresponds to the 'train set' (the first/top chart labeled 'IIW-Train
 Set Image Category Distribution').
  - Relevant sub-chart: First/top bar chart.
 2.2. Locate the bar corresponding to 'Animal Care & Wildlife' on the x-axis of the train set chart.
  - Relevant visual element: The first bar from the left in the top chart.

 2.3. Read the height of this bar to determine the count value on the y-axis.
  - Relevant visual element: The height of the 'Animal Care & Wildlife' bar and its alignment with
  the y-axis tick marks.

 2.4. Report the count as the answer.
Summary: To answer the question, focus on the first/top bar chart, find the 'Animal Care & Wildlife'
bar, and read its height to get the count.

Stage 2. Structured Reasoning Execution Output:

Step 1: Identify which chart corresponds to the 'train set'. The first (top) chart is labeled
'IIW-Train Set Image Category Distribution', so this is the relevant chart.

Step 2: Locate the bar corresponding to 'Animal Care & Wildlife' on the x-axis of the train set
chart. This is the first bar from the left in the top chart.

Step 3: Read the height of this bar to determine the count value on the y-axis. The height of the
'Animal Care & Wildlife' bar aligns with the 500 mark on the y-axis.

Step 4: Report the count as the answer.
Final Answer: The count of Animal Care & Wildlife in train set images is 500.

Stage 3. Self-Verification and Final Answer Generation Output:
The count of Animal Care & Wildlife in train set images is approximately 500. The reasoning process
correctly identified the relevant chart (the top one), the correct bar, and accurately read the
y-axis value.
```

Figure 46: Question and model outputs under VDSP. The input question and the corresponding intermediate reasoning traces and final answer generated by GPT-4.1.

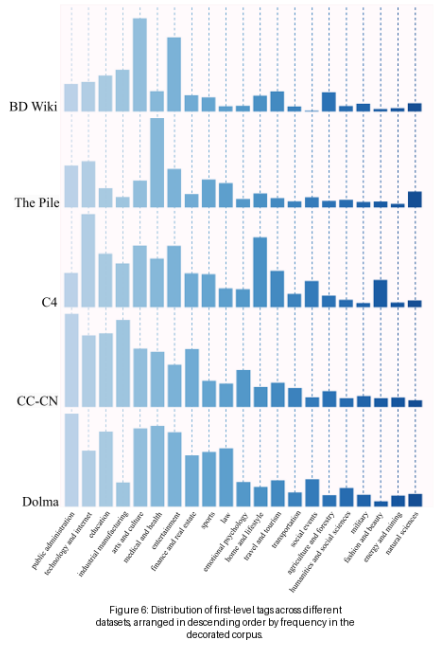


Figure 6: Distribution of first-level tags across different datasets, arranged in descending order by frequency in the decorated corpus.

Question: Which tag has the lowest frequency on The Pile dataset?
Asnwer: energy and mining

Claude-3.7-Sonnet: system_message
GPT-4.1: arts and culture
Gemini-2.0-flash: natural sciences
Gemma-3-4b-it: decorated sciences
LLaVA1.5-7b: Dolma
Pixtral-12B: causality
MATCHA: Dolma
ChartGemma: The answer can be provided based on the visual content of the image. The tag with the lowest frequency on The Pile dataset is 'science and society', which has the lowest bar on the right side of the image.
LLaMA-3.2-11B-Vision: A: military
B: natural science
C: social science
D: social

Figure 47: Examples of wrong answers by MLMs from PolyChartQA. The multi-chart is from Zhao et al. (2024).