

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 MULTI~~CHART~~QA-R: A BENCHMARK FOR MULTI- CHART QUESTION ANSWERING IN REAL-WORLD REASONING SCENARIOS

Anonymous authors

Paper under double-blind review

ABSTRACT

Existing benchmarks for chart analysis primarily focus on single-chart tasks, whereas multi-chart benchmarks are scarce and limited to simplistic question types, making it difficult to comprehensively evaluate the reasoning and decision-making capabilities of multimodal large language models (MLLMs) in realistic scenarios. We present **MultiChartQA-R**, a benchmark designed to evaluate multi-chart question answering capabilities, ranging from fundamental abilities to decision-making applications, with four progressively complex reasoning tasks that encompass real-world scenarios: cross-chart trend comparison, complementary data integration, anomaly and causal analysis, and strategy recommendation. The benchmark consists of versions in three major languages, each containing 695 chart–code pairs and 2,160 QA pairs, with extensibility to additional languages. We further propose a flexible multiple-choice evaluation metric that can be adjusted based on real-world scenarios, along with an extended dataset consisting of 512 charts and 1,212 QA pairs, designed to study retrieval and scaling behavior as the number of charts increases. Our evaluation of 13 representative MLLMs (4 proprietary models and 9 open-weight models) reveals significant performance gaps compared to human, especially in cross-chart visual perception, data integration, and aligning with human preferences. Additionally, our experiments reveal interesting multilingual characteristics of multi-chart question answering.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently demonstrated outstanding performance in various vision-language tasks, such as visual question answering (VQA) (Schwenk et al., 2022; Li et al., 2024b; Jia et al., 2025), chart-to-code generation (Yang et al., 2024), image captioning (Agrawal et al., 2019; Rahman et al., 2023; Kantharaj et al., 2022), and chart question answering (Masry et al., 2022; Methani et al., 2020; Wang et al., 2024; Li et al., 2025; Zeng et al., 2025). The task of chart question answering can be found everywhere in our daily work. Charts, as a powerful tool for data visualization, enable users to quickly grasp trends, patterns, and relationships within the data, thereby facilitating the formulation of strategies for subsequent actions. Many critical scenarios involve the comprehensive analysis of multiple charts. For example, in the financial sector, analysts examine several stock-related indicator charts to predict market trends; researchers compare multiple experimental data charts to discover patterns; and business managers analyze multiple charts related to departmental performance and costs to devise response strategies. However, the effectiveness of multimodal large language models in handling real-world multi-chart analysis scenarios remains insufficiently investigated.

Current chart analysis benchmarks mostly focus on single-chart tasks (Kahou et al., 2018; Kafle et al., 2018; Methani et al., 2020; Masry et al., 2022; Xu et al., 2023; Wang et al., 2024), primarily studying data extraction and multi-hop reasoning within a single chart. These benchmarks do not cover the multi-chart analysis scenarios encountered in real-world applications. The number of multi-chart benchmarks (Liu et al., 2024; Zhu et al., 2025b) is limited, and the variety of questions is insufficient. Research in this area primarily focuses on data comparison between charts and multi-hop question answering, with less emphasis on more complex cross-chart deep logical reasoning and multi-dimensional information integration. Moreover, existing benchmarks for multi-chart

analysis are predominantly English-centric, failing to meet the practical demands of multilingual chart analysis in a globalized context.

To address this, we introduce **MultiChartQAR** (fig. 1), a benchmark designed to evaluate multi-chart question answering abilities, from fundamental skills to decision-making applications. The core of multi-chart joint question answering lies in addressing questions that cannot be answered by a single chart alone, requiring the extraction, correlation, and reasoning of information across multiple charts. MultiChartQA-R is designed to reflect the practical scenarios of multi-chart question answering. It defines four tasks (section 2.1) to evaluate the capabilities of MLLMs. 1) **Cross-chart trend inference:** Emphasizes the ability of "information correlation," requiring the identification of dynamic relationships (e.g., synchronization, divergence) between indicators across different charts, which is fundamental for multi-chart analysis. 2) **Complementary data integration:** Focuses on the "data utilization" ability, emphasizing the extraction of hidden insights through logical or mathematical combinations of multi-chart data, highlighting the core value of "complementarity" in multi-chart data. 3) **Anomaly and pattern analysis:** Centers on the "deep analysis" ability, requiring the exploration of the underlying causes behind surface-level phenomena by combining chart information with external knowledge, thus reflecting the "depth" of the analysis. 4) **Strategy recommendation:** Focuses on the "practical application" ability, providing actionable decision-making suggestions based on the correlation patterns between multiple charts, thereby reflecting the "practicality" of the analysis. Together, these four tasks form a comprehensive logical chain for multi-chart analysis, progressing from "basic correlation" and "data utilization" to "deep analysis" and "practical application," covering the essential capabilities required for multi-chart joint question answering.

This paper presents the construction and expansion process of MultiChartQA-R (section 2.2), which includes the creation of chart-code pairs, QA pair generation, and multilingual expansion methods. MultiChartQA-R includes 14 chart types across 36 domains, available in three languages, with each language containing 2,160 QA pairs to ensure diverse coverage, with stringent data quality control and validation applied (section 2.3). A comprehensive comparison with existing chart QA benchmarks was performed (section 2.4), demonstrating MultiChartQA-R's significance in the field of chart-based question answering.

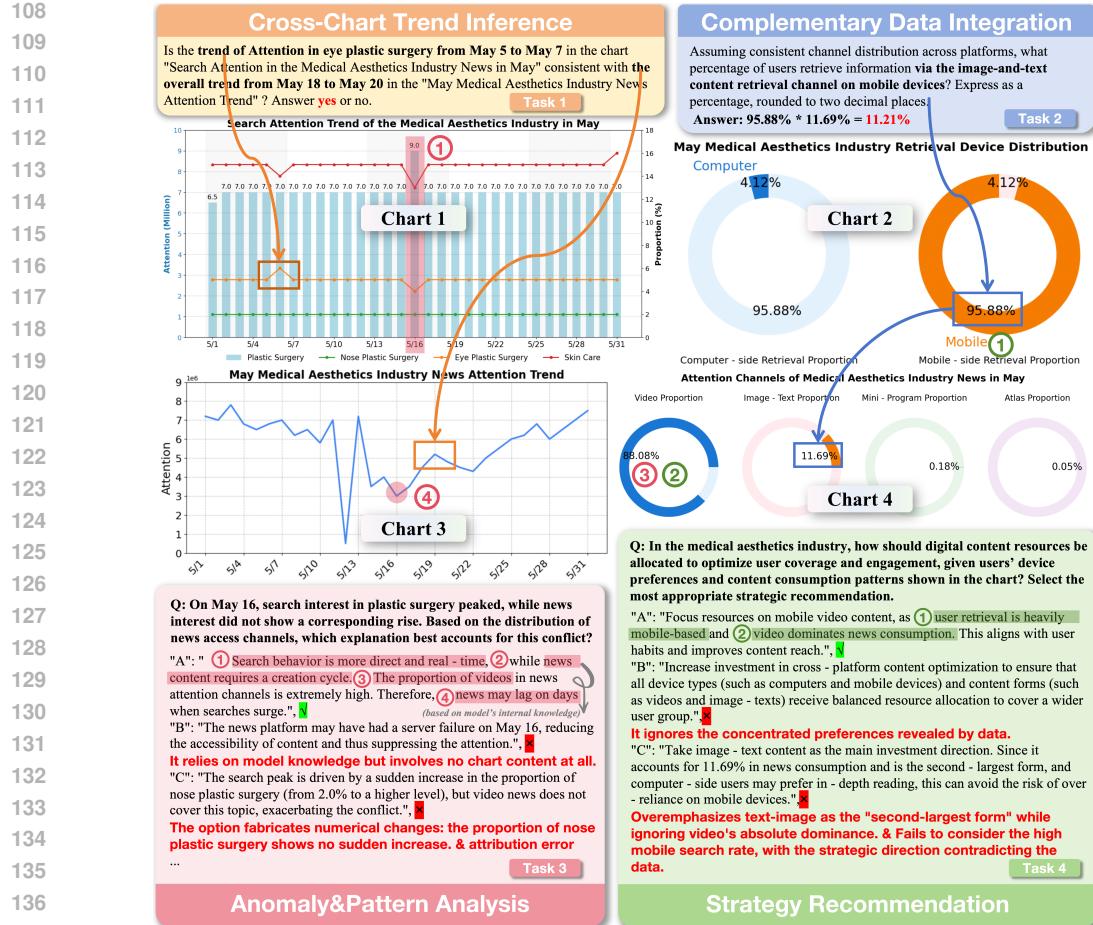
We evaluated four proprietary models and nine open-weight MLLMs on the MultiChartQA-R benchmark (section 3.1). To better evaluate model performance, we propose a flexible multiple-choice metric (appendix E.1) that balances rewards and penalties. This metric also allows for assessing the model's conservatism or aggressiveness, facilitating the training of models with different preferences. Extensive experiments show that proprietary and certain open-source models excel in analytical decision-making, but still lag behind humans in basic visual understanding and data integration (section 3.3). We conducted a comprehensive set of comparisons (section 4), exploring the performance of MLLMs in real-world multi-chart QA scenarios and examining the multilingual aspects of cross-chart question answering.

Our main contributions are as follows:

- We introduce the first scalable, multilingual benchmark for multi-chart question answering, designed to focus on real-world multi-chart task scenarios.
- We propose a flexible multiple-choice evaluation metric that balances rewards and penalties, reflecting the model's analytical decision-making ability and preferences. It can also be used for training preference-based models.
- We comprehensively evaluate existing MLLMs to provide insights into the critical capabilities required for real-world multi-chart scenarios and to assess their performance on cross-modal, multilingual multi-chart tasks.

2 MULTIChartQA-R:

In this section, we first introduce the definition of four tasks involved in MultiChartQA-R (section 2.1), and then delineate the data curation process (section 2.2). Subsequently, we conduct a quantitative analysis to demonstrate the diversity of MultiChartQA-R and validate its quality through manual evaluation methods (section 2.3). Finally, we compare MultiChartQA-R with existing related benchmarks to demonstrate its superiority and effectiveness (section 2.4).



138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

Figure 1: Examples of the four tasks in MultiChartQA-R. Task1&2 use arrows to illustrate the solution process, where the model must identify corresponding trends in the charts for comparison or find complementary data for integration. Task4&5 require identifying related information across multiple charts and reasoning based on the model's internal knowledge to derive the correct inference. Information used by the model is numbered, with erroneous inferences highlighted in red.

2.1 TASK DEFINITION

We designed four tasks based on common multi-chart question answering scenarios encountered in daily life and work. These tasks form a complete logical chain of multi-chart analysis, progressing from "basic correlation" and "data utilization" to "deep analysis" and finally "practical application," covering the core capability requirements of multi-chart question answering scenarios.

Cross-Chart Trend Inference Task 1 aims to evaluate the model's ability to analyze and judge trends across multiple charts, requiring the model to discern the relationships between the trends of various indicators. Specifically, the model needs to identify the trend directions (such as increasing, decreasing, or stable) of indicators in different charts and assess their synchronization or divergence, as shown in Task 1 of Figure 1.

Complementary Data Integration Task 2 evaluates the model's ability to integrate complementary data from different charts and derive hidden information through logical combinations or mathematical operations. These data may include proportions, totals, ratios, and other forms, and the desired result cannot be directly obtained from a single chart; instead, information from multiple charts must be combined for inference, as shown in Task 2 of Figure 1.

Anomaly and Pattern Analysis Task 3 requires the model to identify anomalous data phenomena or potential underlying patterns across multiple charts and provide explanations for these anomalies

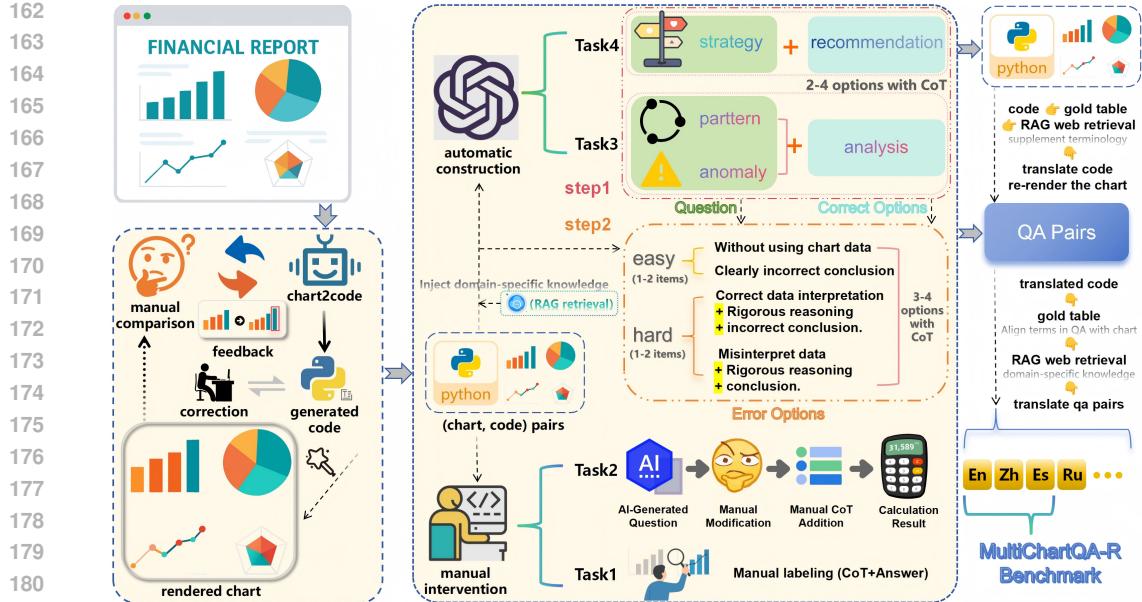


Figure 2: The Construction Pipeline of MultiChartQA-R

by combining chart information with relevant external knowledge. As shown in Task 3 of Figure 1, it involves both identifying surface-level data phenomena and investigating their underlying causes.

Strategy Recommendation Task 4 aims to evaluate the model’s ability to extract relational patterns between different indicators (such as trade-offs, threshold critical points influencing decisions, etc.) based on multi-chart analysis results, and to propose optimization strategies or decision recommendations based on these patterns. As shown in Task 4 of Figure 1, the model must synthesize key information from multiple charts and clarify the interactions between indicators. The strategies generated should align with the patterns in the charts and have practical application value.

2.2 CONSTRUCTION PROCESS

Figure 2 illustrates the construction process of MultiChartQA-R, from the collection of chart-code pairs, the construction of QA pairs, to the final multilingual expansion.

Chart-Code Pairs Collection We searched publicly available channels on the Internet for high-quality industry analysis reports and search-index dashboards, which often contain multiple inter-related charts for data analysis. Because the underlying raw data for these publicly shared charts is typically unavailable, we employed a human-in-the-loop process in which a large model reverses each chart into Python code. Through multiple rounds of manual interaction, we refined the generated code so that the reconstructed charts closely match the originals and preserve the conveyed information. This approach yields charts that reflect real-world patterns, conform to common sense, and carry greater value and significance, while human oversight ensures that the styles of the code-generated charts are attractive and diverse. We collected a total of 180 multi-chart sets and 695 chart-code pairs.

Question–Answer Pair Construction The first two tasks are manually annotated, with a standardized output format applied to the questions. For the latter two tasks, the synthesis process involves extracting a set of multi-chart gold tables from the code and using a reasoning model to generate questions, correct options, and the reasoning process for the correct options, with the task definition and gold tables serving as context in a few-shot manner. To further enhance the quality and relevance of the generated QA pairs, RAG web retrieval is employed to access domain-specific knowledge, supplementing the reasoning process with additional contextual information. The reasoning model is then used to generate 3-4 incorrect distractors, using the gold tables, correct question-answer pairs, and web-retrieved knowledge as context. These distractors are classified into two difficulty levels: “easy” and “hard.” **Easy distractors** (1 or 2) do not rely on chart content and can be easily excluded based on general industry or logical knowledge, while **hard distractors** (1 or 2) involve misinter-

216 Table 1: Statistics of MultiChartQA-R
217

Statistic	Number	Category	Number	Ans Type
Total Questions	2160	- Cross-chart Trend Inference	540(25%)	Yes/No
Unique charts	695	- Complementary Data Integration	540(25%)	Generation
Multi-chart sets	180	- Anomaly and Pattern Analysis	540(25%)	Multi-option
Average charts	3.9	- Strategy Recommendation	540(25%)	Multi-option

223 preting the chart data or using rigorous reasoning that leads to an incorrect conclusion, making them
224 harder to distinguish due to their more convincing structure and use of chart information.

225 **Multilingual Expansion** First, we construct charts in different languages by using LLMs to translate
226 the textual information in the chart rendering code into the target languages, and then execute the
227 code to render charts in those languages. During the code translation step, RAG is employed to
228 search for relevant technical terms and specialized vocabulary to ensure accurate and contextually
229 correct translation of the chart’s content. Next, we construct question-answer pairs in different
230 languages by extracting gold tables from the chart rendering code. The content of the gold tables is
231 organized into text, describing each data point, and we feed these descriptions as context to the LLM
232 during translation to ensure consistent terminology with the chart content. Additionally, RAG is
233 utilized during QA pair translation to search for related domain-specific knowledge, further ensuring
234 that the translation is consistent with the intended meaning and usage. We provide charts and QA
235 pairs in English, Chinese, and Spanish, and this approach can be extended to additional languages
236 to facilitate in-depth exploration of multilingual cross-chart tasks.

237 2.3 DATA STATISTICS & QUALITY INSPECTION

239 MultiChartQA-R includes 180 sets of charts and 695 chart-code pairs, covering three languages
240 (extendable), as detailed in Table 1. In terms of chart types, it includes 14 categories of charts,
241 spanning 36 domains, with specific details provided in Figures 4 and 3 in the appendix D.1.

242 We conducted a rigorous cross-review of all question–answer pairs in the first two tasks. Task 3 and
243 Task 4 are multi-step synthesized data. We designed a supervised scoring mechanism to perform
244 manual quality evaluation across three dimensions: question-type alignment, validity of correct
245 options, and effectiveness of distractors. We divided the experts into a review group and a problem-
246 solving group to score and solve a randomly sampled 30% of the data, respectively. The review
247 group results showed that the average scores for both task types exceeded 9.0 (out of 10), while the
248 problem-solving group achieved a 90+ MF_{β} score and over 85% inter-rater consistency, indicating
249 that the overall data quality is robust and reliable. This also reflects the rationality of the task design
250 and its research value. Detailed evaluation metrics and processes can be found in the appendix D.2.

252 2.4 COMPARISONS WITH EXISTING BENCHMARKS

254 To further distinguish the difference between MultiChartQA-R and other existing ones, we elaborate
255 the benchmark details in table 2. A comparison with other benchmarks clearly demonstrates that
256 MultiChartQA-R excels in terms of broader scope, flexibility, and real-world applicability, offering
257 superior quality and relevance for multi-chart question answering tasks.

258 Table 2: Comparison of MultiChartQA-R with existing chart-based QA benchmarks.

Benchmarks	Reflect Real Scenarios	Topic Diversity	MultiChart	Multilingual	CoT	Chart-Code Pairs	Evaluation Metric	# of Chart Types
PlotQA (Methani et al., 2020)	✗	✗	✗	✗	✗	✗	Accuracy	3
ChartQA (Masry et al., 2022)	✓	-	✗	✗	✗	✗	Accuracy	3
ChartXiv (Wang et al., 2024)	✓	✓	✗	✗	✗	✗	GPT-4 Score	18
ChartQAPro (Masry et al., 2025)	✓	✓	✓	✗	✗	✗	Accuracy + ANLS score	9+
MultiCharQA (Zhu et al., 2025b)	✓	-	✓	✗	✗	✗	Accuracy	-
MultiCharQA-R(Ours)	✓	✓	✓	✓	✓	✓	Accuracy + MF_{β} -score	14

264 3 EXPERIMENTS

266 3.1 EXPERIMENTAL SETUP

268 We conducted benchmark evaluations on 13 widely used proprietary and open-weight MLLMs in
269 the field. For proprietary models, we selected three representative models: GPT-4o (OpenAI, 2024),

270 Claude-Sonnet-4 (Anthropic, 2025), and Gemini-2.5-Pro (Team, 2025), and also included a newly
 271 discovered proprietary model, Seed1.5-VL (Guo et al., 2025) in the evaluation. For open-weight
 272 MLLMs, we selected nine competitive models, with parameter sizes ranging from 7B to 78B: In-
 273 ternVL2(26B, 76B) (Chen et al., 2024b), InternVL3-78B (Zhu et al., 2025a), Qwen2.5-VL(7B, 72B)
 274 (Qwen et al., 2025), LLaVA-OV(7B, 72B) (Li et al., 2024a), DeepSeek-VL-7B (Lu et al., 2024), and
 275 MiniCPM-V-2.6 (Yao et al., 2024). All evaluations employed the Chain-of-Thought (CoT) (Wei
 276 et al., 2022) technique, and the corresponding prompts are provided in the appendix H.

277

278

3.2 EVALUATION METRIC

279

280 **Cross-Chart Trend Inference** employs strict string matching, as the questions explicitly constrain
 281 the answer format and are predominantly Yes/No judgments.

282

283 **Complementary Data Integration** addresses the fact that large language models struggle with
 284 arithmetic. We ask the model to extract the necessary values from the charts and outline the step-
 285 by-step reasoning. We then feed those steps into DeepSeek-V3.1 to generate executable Python
 286 code, run the code to obtain the final numeric result, and use a regular expression to extract the
 287 numeric component. We evaluate the correctness of the reasoning chain with a relaxed accuracy
 288 metric(Masry et al., 2022), thus testing the model’s ability to perform long-form, multi-step inference.

289

290 **Anomaly and Pattern Analysis** and **Strategy Recommendation** involve open-ended multi-chart
 291 question-answering, where the answers are not unique, and human evaluators often make selections
 292 based on their preferences. Therefore, these tasks are designed in a multiple-choice format. To
 293 evaluate the model’s performance, we propose an evaluation metric, the Multiple-choice F_β Score
 294 (MF_β), which combines the base score, penalty, and a final composite score, comprehensively
 295 measuring the model’s ability to balance correct selections and avoid incorrect ones. The process
 296 for constructing MF_β is as follows.

297

298 *BaseScore* is used to assess the model’s ability to select the correct answers. Let the set of correct
 299 answers be denoted as A and the set of answers selected by the model as B . The BaseScore is
 300 defined as:

301
$$\text{BaseScore} = \frac{|A \cap B|}{|A|} \quad (1)$$

302

303 where $|A \cap B|$ is the number of correctly selected items by the model, and $|A|$ is the total number
 304 of correct answers. The BaseScore lies in the range $[0, 1]$, with a score of 1 for perfect correctness,
 305 a score between 0 and 1 for partial correctness, and a score of 0 for complete incorrectness.

306

307 *Penalty* measures the model’s misselection behavior, especially when the model selects incorrect
 308 or interfering options. We classify interfering items into two categories: easy-to-confuse items (set
 309 E) and hard-to-confuse items (set H). Each time the model selects an interfering item, it incurs
 310 a penalty. The penalty coefficients w_e and w_h correspond to the easy and hard interfering items,
 311 respectively, and satisfy the constraint $w_e = 2w_h$ and $w_e \cdot |E| + w_h \cdot |H| = 1$, where $|E|$ and $|H|$
 312 represent the number of easy and hard interfering items. The total penalty is then computed as:

313

314
$$\text{Penalty} = w_e \cdot |B \cap E| + w_h \cdot |B \cap H| \quad (2)$$

315

316 where $|B \cap E|$ and $|B \cap H|$ represent the numbers of easy and hard interfering items selected by the
 317 model.

318

319 If the score for each task is simply computed as $\text{Score} = \max(0, \text{BaseScore} - \text{Penalty})$, this
 320 formula ensures non-negative performance by considering both correct selections and error avoidance.
 321 However, a low score may indicate that the model is either too conservative (e.g., $|B| = 1$,
 322 $|E| = |H| = 0$) or too aggressive (e.g., $|B| = 4$, $|E| + |H| = 4$). To address this, we propose a
 323 more integrated metric that evaluates the BaseScore, Penalty, and Score simultaneously.

324

325 We introduce the F_β -score to multiple-choice tasks and construct the MF_β evaluation metric, which
 326 considers two key aspects of performance: selecting correct answers (BasicScore) and avoiding
 327 incorrect ones (EscapeScore). This approach provides a more comprehensive assessment of the
 328 model’s overall effectiveness.

329

330
$$\text{EscapeScore} = 1 - \text{Penalty} \quad (3)$$

6

$$MF_{\beta} = (1 + \beta^2) \times \frac{\text{BaseScore} \times \text{EscapeScore}}{\beta^2 \times \text{BaseScore} + \text{EscapeScore}} \quad (4)$$

where β is a tuning parameter used to control the balance between BaseScore and EscapeScore. If $\beta = 1$, the model is required to both select correctly and avoid errors equally. When $\beta > 1$, greater emphasis is placed on avoiding incorrect selections, whereas if $\beta < 1$, the focus shifts toward selecting correct items.

This refined scoring mechanism offers a balanced approach for evaluating multi-option selection tasks by considering both the accuracy of selections and the avoidance of errors. We compare MF_{β} with Com^2 (Xiong et al., 2025) in appendix E.1, where we highlight MF_{β} ’s role in model selection for specific scenarios and its feasibility for preference model training.

Table 3: The MultiChartQA-R leaderboard. The best scores are in bold.

Model	Trend Inference			Data Integration			Anomaly/Pattern Attr			Strategy Rec		
	en	zh	es	en	zh	es	en	zh	es	en	zh	es
Human	97.83			94.83			90.60			91.60		
Proprietary Models												
Claude-Sonnet-4	70.00	75.78	69.33	60.99	59.64	62.98	84.92	64.12	84.84	87.53	67.09	86.95
Gemini-2.5-Pro	75.06	79.33	75.11	65.08	69.35	65.91	81.87	83.92	82.48	83.79	84.49	83.96
Seed1.5-VL	72.44	71.78	67.33	67.66	72.81	68.64	78.87	82.46	78.69	82.22	85.38	78.81
GPT-4o	64.21	62.64	59.87	64.83	63.60	64.77	71.76	63.62	67.33	76.88	67.72	70.47
open-weight MLLMs												
InternVL3-78B	73.21	68.46	64.66	67.50	70.72	63.33	81.62	82.22	78.16	81.78	84.48	76.57
InternVL2-L3-76B	59.91	48.88	59.19	51.59	53.51	51.14	68.33	71.69	70.86	70.19	76.51	75.77
Qwen2.5-VL-72B	56.25	56.95	53.13	25.40	14.77	21.51	72.62	75.89	71.12	76.34	78.75	72.85
LLaVA-OV-72B	61.33	57.59	53.33	43.02	16.25	22.22	66.82	66.82	66.58	71.37	68.62	69.79
InternVL2-26B	54.46	58.65	50.11	31.03	28.70	21.95	56.85	62.66	53.39	64.72	67.41	58.98
Qwen2.5-VL-7B	54.44	54.91	52.22	21.04	19.64	21.14	71.38	70.93	65.08	74.55	73.77	68.21
MiniCPM-V-2.6	49.88	55.36	44.22	23.29	26.15	16.91	60.67	60.73	54.87	60.13	65.27	55.52
DeepSeek-VL-7B	49.32	47.11	40.44	7.95	5.01	6.74	49.90	48.30	48.30	56.95	52.51	49.91
LLaVA-OV-7B	48.55	34.00	46.00	21.84	8.20	12.38	52.18	52.40	56.14	58.98	62.31	61.30

3.3 MAIN RESULTS

Table 3 summarizes the evaluation results of 13 MLLMs on MultiChartQA-R. Our key observations are as follows:

The foundational visual capabilities and data integration abilities of proprietary models show a significant gap compared to humans in cross-chart scenarios, but they have demonstrated good performance in pattern summarization and logical reasoning. Among the proprietary models, Gemini-2.5-Pro shows superior trend-analysis ability across three languages. The newly released Seed1.5-VL achieves the best results on data integration task. InternVL3-78B exhibits outstanding performance, achieving state-of-the-art results among open-weight MLLMs and approaching the performance of proprietary models across all tasks.

Proprietary models continue to outperform most open-weight MLLMs by a considerable margin. Although InternVL3-78B achieves performance comparable to proprietary models, the remaining open-weight MLLMs lag substantially. This marked disparity confirms that MultiChartQA-R poses a significant challenge for current open-weight multimodal large language models. In particular, 7B and 8B open-weight MLLMs attain only near-random accuracy on trend-judgment task, revealing a lack of genuine trend-analysis capability, and their accuracy across the other three tasks also remains deficient. Nonetheless, a clear positive correlation is observed between parameter scale and performance on all four tasks. Overall, these results indicate that the open-source community still has ample scope to enhance MLLMs’ competencies in complex visual understanding, cross-modal reasoning, conflict detection and attribution, and strategy formulation.

MLLMs exhibit heterogeneous performance in multi-chart tasks across different languages. Unlike other multilingual benchmarks, we did not observe English dominance. We speculate that the parameter subspace associated with data-analytic reasoning has limited overlap with that governing multilingual processing, thereby attenuating any potential English-language advantage. We will investigate MLLMs’ performance in cross-lingual multi-chart tasks further in the Discussion section.

378

4 DISCUSSION

379

380 4.1 IMPACT OF IRRELEVANT CHARTS

381
382 During annotation, We recorded the specific charts associated with each QA pair in a “charts in-
383 volved” field. In the main experiments, only these relevant charts were provided to the MLLMs.
384 To investigate the impact of irrelevant ones on MLLMs’ visual QA performance, we designed a
385 comparative experiment in which additional unrelated charts were introduced alongside the relevant
386 ones during inference. When the number of charts in a set is too large, the issue of excessively long
387 visual tokens arises. The solution to this problem is provided in the appendix F.1.388 Table 4 demonstrates that proprietary models maintain stable performance across the four tasks,
389 even with additional chart inputs. This suggests that proprietary models possess strong capabili-
390 ties for locating and retrieving relevant chart information when the question intent is clear. Open-
391 weight MLLMs suffer performance drops across all four task categories, indicating that their chart-
392 information retrieval and localization abilities still have substantial room for improvement.393

394 4.2 PERFORMANCE ACROSS DIFFERENT LANGUAGES

395 To further assess multilingual performance in multi-chart QA, we conducted two comparative ex-
396 periments. In the first setting, we used charts in English while varying the language of the question-
397 answer pairs and prompts. In the second setting, the QA pairs and prompts remained in English,
398 whereas the charts were translated into different languages. The results show that reasoning on
399 English charts with prompts in different languages resulted in significant performance fluctuations.
400 In contrast, when reasoning with different charts but the same language, the fluctuation in results
401 across languages was smaller. This indicates that the cross-linguistic consistency of reasoning in
402 multi-chart question answering tasks for MLLMs still requires improvement. The experimental data
403 are presented in Table 6&7 of the appendix F.2.404

405 4.3 EXPLORING MLLMs’ RETRIEVAL CAPABILITIES

406 Additionally, we constructed a dedicated dataset to analyze large models’ entity-extraction perfor-
407 mance across multiple charts by extracting the numeric answers and computing relaxed accuracy.408 Details about the extended-benchmark can be found in the appendix G. A brief overview of its four
409 task types is as follows: Parallel-type question-answer pairs extract content from different charts
410 based on independent sub-questions and list them individually. Union-type question-answer pairs
411 extract content from different charts based on a single question, perform combination operations,
412 and output a single answer. “PCPC” stands for “per chart per content” meaning each chart involves
413 one piece of content. “PCMC” stands for “per chart multi-content” meaning each chart may involve
414 more than one piece of content.415 Comparing the results across the four sets in Figure 7 of appendix F.3, we observed that MLLMs’
416 performance consistently worsens as the number of charts increases and the amount of informa-
417 tion per chart grows. Interestingly, the experiments also revealed that when processing multiple
418 charts—extracting one datum per chart—a simple additional computation step to produce a cal-
419 culated result achieves a higher score than directly outputting multiple data points. This finding
420 indicates that MLLMs still need to improve their ability to process multiple queries in parallel.421

422 4.4 ERROR ANALYSIS

423 In the evaluation, GPT-4o exhibited a marked decline in accuracy in task 3 and task 4 multiple-
424 choice tasks, with accuracy rates much lower than those of Claude-Sonnet-4 and Seed1.5-VL. We
425 further discovered that this performance decline in GPT-4o is closely related to its tendency to mimic
426 the format of one-shot example answers (B, C, E) in the prompt. Although the order of the options
427 in the test questions had been shuffled, GPT-4o still frequently selected options B, C, and E. For ex-
428 ample, in task 3, the misselection rates for B, C, and E were 48.0%, 53.4%, and 51.6%, respectively,
429 significantly higher than the rates for other options (with an average misselection rate of 1.2% for
430 other options) and other models (Claude had an average misselection rate of 10.4% for B, C, and E).
431 This indicates that the model failed to reason based on the chart content and instead overly replicated

432 the structural pattern of the example answers, showing a strong dependency on example structure,
 433 which resulted in higher misselection rates and an overall performance decline.
 434

435 In contrast, Claude-Sonnet-4 maintained a higher accuracy rate while demonstrating significantly
 436 lower misselection rates for non-example options (such as A, D, F), showing stronger suppression
 437 of prompt bias (Xu et al., 2024) (where pre-trained language models may develop unreasonable
 438 preferences for labels suggested by the prompt) and a more balanced judgment of the content of
 439 each option. Seed1.5-VL showed slightly lower accuracy, but its bias toward specific options was
 440 still noticeably better than that of GPT-4o. Overall, the models’ performance on multiple-choice
 441 tasks is somewhat limited by their ability to suppress irrelevant structural information introduced by
 442 the prompt and to adapt to the actual semantic requirements of the questions. The statistical results
 443 are presented in the appendix F.4.

444 5 RELATED WORKS

445 **Chart Question Answering Benchmarks** Early benchmarks such as FigureQA (Kahou et al.,
 446 2018), DVQA (Kafle et al., 2018), PlotQA (Methani et al., 2020), and ChartQA (Masry et al., 2022)
 447 primarily focused on basic chart types, and addressed fundamental question-answering tasks such
 448 as data extraction. These tasks were limited in scope and did not fully cover the application of charts
 449 in complex and diverse environments. In recent years, with advancements in research, new bench-
 450 marks have emerged, such as ChartBench (Xu et al., 2023), ChartLlama (Han et al., 2023), Charxiv
 451 (Wang et al., 2024), and ChartAssistant (Meng et al., 2024), which enhance the diversity of both
 452 charts and questions. Additionally, tasks like Chart-to-code (Yang et al., 2024), which involve more
 453 challenging visual understanding, have also appeared. However, research on multi-chart question
 454 answering remains relatively scarce.
 455

456 **Multilingual Chart Question Answering Benchmarks** The rapid growth of multilingual VQA
 457 benchmarks (e.g., xGQA (Pfeiffer et al., 2022), MaXM (Changpinyo et al., 2023), CVQA (Romero
 458 et al., 2025)) has addressed the English-centric bias in visual question answering. However, mul-
 459 tilingual reasoning over structured charts remains severely . POLYCHARTQA (Xu et al., 2025)
 460 spans ten languages and reveals performance deficits on non-English inputs, but its tasks empha-
 461 size shallow extraction and lack true reasoning challenges. OneChart (Chen et al., 2024a)’s ChartY
 462 benchmark covers only Chinese and English and focuses on structural extraction, lacking a sys-
 463 tematic evaluation of multilingual chart reasoning. KITAB-Bench (Heakl et al., 2025) targets En-
 464 glish–Arabic chart localization but is limited both in language coverage and task depth.
 465

466 **Multi-chart Question Answering Benchmarks** A series of multi-image question-answering
 467 benchmarks have emerged, such as Mantis-Instruct (Jiang et al., 2024), BLINK (Fu et al., 2024),
 468 and MUIRBENCH (Wang et al., 2025). However, these do not include chart-type images. MMC-
 469 Benchmark (Liu et al., 2024) contains a small subset of multi-chart data, but it only includes 52
 470 samples. ReMI (Kazemi et al., 2024) includes some multi-chart scenarios, but the question types
 471 are limited. MultiChartQA (Zhu et al., 2025b) is the first benchmark specifically designed to explore
 472 multi-chart question answering, encompassing three types of multi-chart tasks: multi-chart infor-
 473 mation extraction, cross-chart data comparison, and sequential reasoning. While it reflects some of the
 474 capabilities of MLLMs in multi-chart reasoning tasks, it does not fully capture the real-world sce-
 475 nario demands. This paper introduces more realistic tasks that better reflect the performance aspects
 476 that are of greater concern to users.
 477

478 6 CONCLUSION

479 In this paper, we introduce MultiChartQA-R, a benchmark designed to assess the multi-chart reason-
 480 ing capabilities of MLLMs through four core tasks, each reflecting a crucial aspect of the multi-chart
 481 analytical reasoning process. Additionally, it can be extended to multiple languages. We also pro-
 482 pose a flexible multiple-choice evaluation metric, MF_{β} , whose effectiveness is validated through
 483 formal reasoning and comparative experiments. Furthermore, we conduct extensive cross-chart
 484 question-answering and cross-language experiments on 13 mainstream MLLMs, revealing several
 485 intriguing phenomena. MultiChartQA-R serves as a foundation for advancing the development of
 486 more capable MLLMs in real-world scenarios.
 487

486 REFERENCES
487

488 Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra,
489 Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019
490 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8947–8956, 2019. doi:
491 10.1109/ICCV.2019.00904.

492 Anthropic. Claude opus 4 & claude sonnet 4: System card, 2025. URL <https://www.anthropic.com/claude-4-system-card>. Accessed: 2025-07-31.
493

494 Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V. Thapliyal, Idan Szpektor, Julien
495 Amelot, Xi Chen, and Radu Soricut. MaXM: Towards multilingual visual question answering. In
496 *Findings of the Association for Computational Linguistics: EMNLP*, 2023.

497 Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun,
498 Chunrui Han, and Xiangyu Zhang. Onechart: Purify the chart structural extraction via one auxil-
499 iary token, 2024a.
500

501 Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shen-
502 glong Ye, Hao Tian, Zhaoyang Liu, and et al. Internvl 2.5: Expanding performance bound-
503 aries of open-source multimodal models with model, data, and test-time scaling. <https://internvl.github.io/blog/2024-07-02-InternVL-2.0/>, 2024b. Accessed:
504 2025-07-31.
505

506 Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A.
507 Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see
508 but not perceive. In *Computer Vision – ECCV 2024: 18th European Conference, Milan,
509 Italy, September 29–October 4, 2024, Proceedings, Part XXIII*, pp. 148–166, Berlin, Heidelberg,
510 2024. Springer-Verlag. ISBN 978-3-031-73336-9. doi: 10.1007/978-3-031-73337-6_9. URL
511 https://doi.org/10.1007/978-3-031-73337-6_9.
512

513 Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang,
514 Jianyu Jiang, Jiawei Wang, and et al. Seed1.5-vl technical report. <https://arxiv.org/abs/2505.07062>, 2025. Accessed: 2025-07-31.
515

516 Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang
517 Zhang. Chartllama: A multimodal llm for chart understanding and generation, 2023.
518

519 Ahmed Heakl, Muhammad Abdullah Sohail, Mukul Ranjan, Rania Elbadry, Ghazi Shazan Ah-
520 mad, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan.
521 KITAB-bench: A comprehensive multi-domain benchmark for Arabic OCR and document under-
522 standing. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
523 (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22006–22024,
524 Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-
525 5. URL <https://aclanthology.org/2025.findings-acl.1135/>.
526

527 Mengzhao Jia, Wenhao Yu, Kaixin Ma, Tianqing Fang, Zhihan Zhang, Siru Ouyang, Hongming
528 Zhang, Dong Yu, and Meng Jiang. Leopard: A vision language model for text-rich multi-image
529 tasks, 2025. URL <https://arxiv.org/abs/2410.01744>.
530

531 Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhui Chen. Mantis:
532 Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024,
533 2024. URL <https://openreview.net/forum?id=skLtdUVaJa>.
534

535 Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visual-
536 izations via question answering. In *Proceedings of the IEEE conference on computer vision and
537 pattern recognition*, pp. 5648–5656, 2018.

538 Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and
539 Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. In *ICLR (Workshop)*.
OpenReview.net, 2018. URL <https://openreview.net/forum?id=H1mz0OyDz>.

540 Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque,
 541 and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In Smaranda
 542 Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meet-
 543 ing of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4005–4023,
 544 Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.
 545 acl-long.277. URL <https://aclanthology.org/2022.acl-long.277>.

546 Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare
 547 Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, and Ahmed Qureshi. ReMI: A dataset
 548 for reasoning with multiple images. In *The Thirty-eight Conference on Neural Information Pro-
 549 cessing Systems Datasets and Benchmarks Track*, 2024. URL [https://openreview.net/](https://openreview.net/forum?id=930e8v5ctj)
 550 [forum?id=930e8v5ctj](https://openreview.net/forum?id=930e8v5ctj).

551 Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei
 552 Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, Aug 2024a. URL
 553 <https://arxiv.org/abs/2408.03326>. arXiv:2408.03326 [cs.CV].

554 Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan.
 555 Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference
 556 on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, 2024b. doi: 10.1109/
 557 CVPR52733.2024.01263.

558 Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim,
 559 Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang
 560 Lim, and William Yang Wang. Mmsci: A dataset for graduate-level multi-discipline multimodal
 561 scientific understanding, 2025. URL <https://arxiv.org/abs/2407.04903>.

562 Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Ya-
 563 coob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruc-
 564 tion tuning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024
 565 Conference of the North American Chapter of the Association for Computational Linguistics:
 566 Human Language Technologies (Volume 1: Long Papers)*, pp. 1287–1310, Mexico City, Mexico,
 567 June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.70.
 568 URL <https://aclanthology.org/2024.naacl-long.70/>.

569 Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng
 570 Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong
 571 Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. URL <https://arxiv.org/abs/2403.05525>.

572 Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A bench-
 573 mark for question answering about charts with visual and logical reasoning. In Smaranda Mure-
 574 san, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational
 575 Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Compu-
 576 tational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL <https://aclanthology.org/2022.findings-acl.177/>.

577 Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman
 578 Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmo-
 579 hamadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartqapro:
 580 A more diverse and challenging benchmark for chart question answering, 2025. URL <https://arxiv.org/abs/2504.05506>.

581 Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo.
 582 ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and
 583 multitask instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings
 584 of the Association for Computational Linguistics: ACL 2024*, pp. 7775–7803, Bangkok, Thailand,
 585 August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.
 586 463. URL [https://aclanthology.org/2024.findings-acl.463/](https://aclanthology.org/2024.findings-acl.463).

587 Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over
 588 scientific plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*,
 589 March 2020.

594 OpenAI. Gpt-4o system card, 2024. URL <https://openai.com/index/gpt-4o-system-card/>. Accessed: 2025-07-31.
 595
 596

597 Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and
 598 Iryna Gurevych. xGQA: Cross-lingual visual question answering. In Smaranda Muresan, Preslav
 599 Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2497–2511, Dublin, Ireland, May 2022. Association for Computational Linguistics.
 600 doi: 10.18653/v1/2022.findings-acl.196. URL <https://aclanthology.org/2022.findings-acl.196/>.
 601
 602

603 An Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
 604 Dayiheng Liu, Fei Huang, Haoran Wei, and et al. Qwen2.5 technical report. <https://arxiv.org/abs/2412.15115>, January 2025. Accessed: 2025-07-31, arXiv:2412.15115 [cs].
 605
 606

607 Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md. Tahmid Rahman Laskar, Md. Hamjajul
 608 Ashmafee, and Abu Raihan Mostofa Kamal. Chartsumm: A comprehensive benchmark for
 609 automatic chart summarization of long and short summaries. *Proceedings of the Canadian
 610 Conference on Artificial Intelligence*, June 2023. doi: 10.21428/594757db.0b1f96f6. URL
 611 <http://dx.doi.org/10.21428/594757db.0b1f96f6>.
 612
 613

614 David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda
 615 Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, and et al.
 616 Cvqa: culturally-diverse multilingual visual question answering benchmark. In *Proceedings of
 617 the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red
 Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.
 618
 619

620 Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi.
 621 A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer
 622 Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022,
 623 Proceedings, Part VIII*, pp. 146–162, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-
 3-031-20073-1. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.
 624
 625

626 Gemini Team. Gemini 2.5 pro: Pushing the frontier with advanced reasoning, multimodality,
 627 long context, and next generation agentic capabilities. Technical report, DeepMind, 2025.
 628 URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf. Accessed: 2025-07-31.
 629
 630

631 Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan
 632 Xu, Wenxuan Zhou, Kai Zhang, and et al. Muirbench: A comprehensive benchmark for robust
 633 multi-image understanding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=TrVYEZtSQH>.
 634
 635

636 Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi
 637 Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding
 638 in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697,
 639 2024.
 640
 641

642 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi,
 643 Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language
 644 models. In *Proceedings of the 36th International Conference on Neural Information Processing
 645 Systems, NIPS '22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
 646
 647

648 Kai Xiong, Xiao Ding, Yixin Cao, Yuxiong Yan, Li Du, Yufei Zhang, Jinglong Gao, Jiaqian Liu,
 649 Bing Qin, and Ting Liu. Com²: A causal-guided benchmark for exploring complex commonsense
 650 reasoning in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and
 651 Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association
 652 for Computational Linguistics (Volume 1: Long Papers)*, pp. 16119–16140, Vienna, Austria, July
 653 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL <https://aclanthology.org/2025.acl-long.785/>.
 654

648 Yichen Xu, Liangyu Chen, Liang Zhang, Wenxuan Wang, and Qin Jin. Polychartqa: Benchmarking
 649 large vision-language models with multilingual chart question answering, 2025. URL <https://arxiv.org/abs/2507.11939>.
 650

651 Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A
 652 benchmark for complex visual reasoning in charts. *ArXiv*, abs/2312.15915, 2023. URL <https://api.semanticscholar.org/CorpusID:266550948>.
 653

654 Ziyang Xu, Keqin Peng, Liang Ding, Dacheng Tao, and Xiliang Lu. Take care of your prompt
 655 bias! investigating and mitigating prompt bias in factual knowledge extraction. In Nicoletta
 656 Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue
 657 (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics,
 658 Language Resources and Evaluation (LREC-COLING 2024)*, pp. 15552–15565, Torino, Italia,
 659 May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1352/>.
 660

661 Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu,
 662 Siheng Li, Yuxiang Zhang, et al. Chartmimic: Evaluating lmm’s cross-modal reasoning capability
 663 via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024.
 664

665 Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li,
 666 Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding
 667 Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong
 668 Sun. Minicpm-v: A gpt-4v level mllm on your phone, Aug 2024. URL <https://arxiv.org/abs/2408.01800> [cs.AI].
 669

670 Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. Advancing multimodal large language
 671 models in chart question answering with visualization-referenced instruction tuning. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):525–535, 2025. doi: 10.1109/TVCG.2024.3456159.
 672

673 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan,
 674 Hao Tian, Weijie Su, Jie Shao, and et al. Internvl3: Exploring advanced training and test-
 675 time recipes for open-source multimodal models, 2025a. URL <https://arxiv.org/abs/2504.10479>. Accessed: 2025-07-31.
 676

677 Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. MultiChartQA: Bench-
 678 marking vision-language models on multi-chart problems. In Luis Chiruzzo, Alan Ritter, and
 679 Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of
 680 the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long
 681 Papers)*, pp. 11341–11359, Albuquerque, New Mexico, April 2025b. Association for Compu-
 682 tational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.566. URL
 683 <https://aclanthology.org/2025.naacl-long.566/>.
 684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702	APPENDIX	
703		
704		
705	A Ethics Statement	15
706		
707	B Reproducibility Statement	15
708		
709	C LLM Usage	15
710		
711		
712	D benchmark	15
713	D.1 Statistics	15
714	D.2 Quality Inspection	15
715		
716		
717	E Experiments	16
718	E.1 Evaluation Metric	16
719		
720		
721	F Discussion	17
722	F.1 Impact of Irrelevant Charts	17
723	F.2 Performance Across Different Languages	19
724	F.3 Exploring MLLMs’ Retrieval Capabilities	20
725	F.4 Error Analysis	20
726		
727		
728		
729	G Extended Benchmark.	21
730	G.1 Task Description	21
731	G.2 Pipeline	21
732	G.3 Data Statistics	22
733		
734		
735	H QAs Prompts	23
736		
737	I Task 3&4 Synthesis Method	25
738		
739		
740		
741		
742		
743		
744		
745		
746		
747		
748		
749		
750		
751		
752		
753		
754		
755		

756 **A ETHICS STATEMENT**
757758 This work adheres to the ICLR Code of Ethics. In this study, no animal experimentation was in-
759 volved, and no personal data containing privacy or sensitive information was used.
760761 Human annotators were engaged during dataset construction and evaluation. Their tasks were
762 strictly limited to assessing the validity, logical consistency, and accuracy of chart-based question-
763 answer pairs. These activities did not involve personally identifiable information and posed no
764 privacy, safety, or psychological risks. All annotation and evaluation procedures were carried out
765 under compliant and safe conditions.
766767 We took care to avoid potential biases or discriminatory outcomes in both the dataset and the reported
768 results. The authors are committed to maintaining transparency and academic integrity throughout
769 the research process.
770771 **B REPRODUCIBILITY STATEMENT**
772773 We have made every effort to ensure that the results presented in this paper are reproducible. The
774 paper provides detailed descriptions of the data annotation process, quality control mechanisms,
775 experimental design, and evaluation methodology, enabling other researchers to understand and
776 replicate our work. All code and datasets will be released in an anonymous repository upon publica-
777 tion to facilitate replication and verification. The comparative experiments reported in this paper are
778 based on publicly available models and methods, ensuring consistent and reproducible evaluation
779 results. We believe these measures will enable other researchers to reproduce our work and further
780 advance the field.
781782 **C LLM USAGE**
783784 Large Language Models (LLMs) were partially used during this research. Specifically, in dataset
785 construction, LLMs were employed in the initial generation of a subset of questions and answer
786 options, after which human annotators conducted verification and quality checks to ensure accuracy
787 and safety. In manuscript preparation, LLMs were used to assist with language polishing, improving
788 clarity, accuracy, and overall fluency of the text.
789790 It is important to note that LLMs were not involved in research ideation, methodological design,
791 or experimental planning. All research concepts, scientific claims, and data analyses were indepen-
792 dently developed and carried out by the authors. The authors take full responsibility for the content
793 of the manuscript, including sections that involved LLM assistance, and have ensured that the use
794 of LLMs complies with academic ethical standards without contributing to plagiarism or research
795 misconduct.
796797 **D BENCHMARK**
798801 **D.1 STATISTICS**
802803 The statistical results can be found in Figure 3 and Figure 4.
804805 **D.2 QUALITY INSPECTION**
806807 For the first two task types—true/false and numerical-answer questions, both characterized by objec-
808 tively verifiable answers—we implemented a systematic cross-review of all question–answer pairs.
809 The annotation team consisted of four members and employed a cyclic peer-review mechanism.
810 Each annotator’s work was independently verified by another member, and any identified errors
811 were promptly corrected. The review criteria included:
812813 • **Question validity:** ensuring that each question conformed to its definition and was clearly
814 formulated;
815

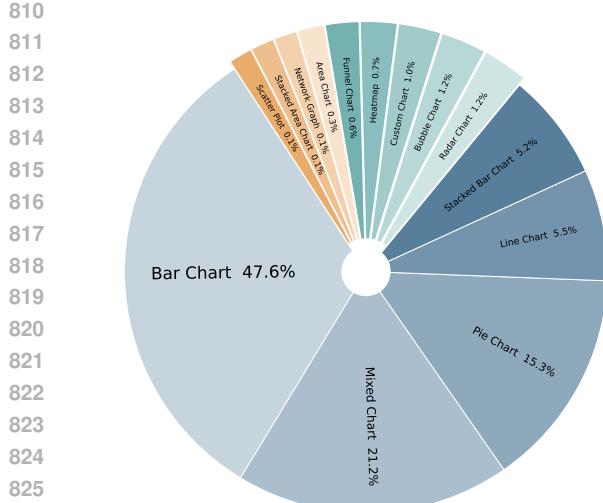


Figure 3: Distribution of Domains

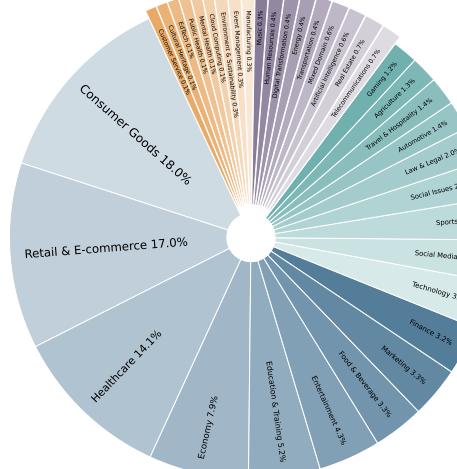


Figure 4: Distribution of Chart Types

- **Reasoning correctness:** verifying that the annotated reasoning chain was logically sound and rigorous;
- **Answer accuracy:** confirming that the final answer was fully consistent with the chart data.

Through this process, we ensured the dataset's robust quality and reliability in terms of question formulation, reasoning, and answer accuracy.

For the two more challenging task types, we randomly sampled 30% of the data for human quality evaluation. Each question-answer pair was scored by evaluators with reference to the reasoning chain used during its generation. We applied a 10-point evaluation scale based on three criteria:

- **Question-type alignment:** assessing whether the question corresponded to the intended task type and was reasonably designed;
- **Validity of correct options:** ensuring that the reasoning behind correct options was strictly grounded in the chart data, logically coherent, and led to reliable conclusions;
- **Effectiveness of distractors:** requiring simple distractors to appear superficially plausible yet independent of the chart data, and difficult distractors to superficially rely on chart reasoning while containing critical logical flaws (e.g., misinterpreting a downward trend as upward).

Evaluation results indicate that the average score for the third task type was 9.1/10, with an inter-rater agreement of 85%, while the fourth task type achieved an average score of 9.3/10 and an inter-rater agreement of 87%, demonstrating highly robust overall evaluation outcomes.

In addition, the evaluators completed all four task types, obtaining corresponding scores of 97.83, 94.83, 90.60, and 91.60, respectively. Human performance consistently exceeded that of the models, though a non-negligible error rate remained. This observation both confirms that the tasks remain challenging for current models and highlights the intrinsic difficulty of the tasks, underscoring their research value.

E EXPERIMENTS

E.1 EVALUATION METRIC

MF_β and Com^2 exhibit similar distributions across the evaluation of various models, indicating that MF_β can effectively capture the model's multiple-choice capabilities. However, Com^2 amplifies

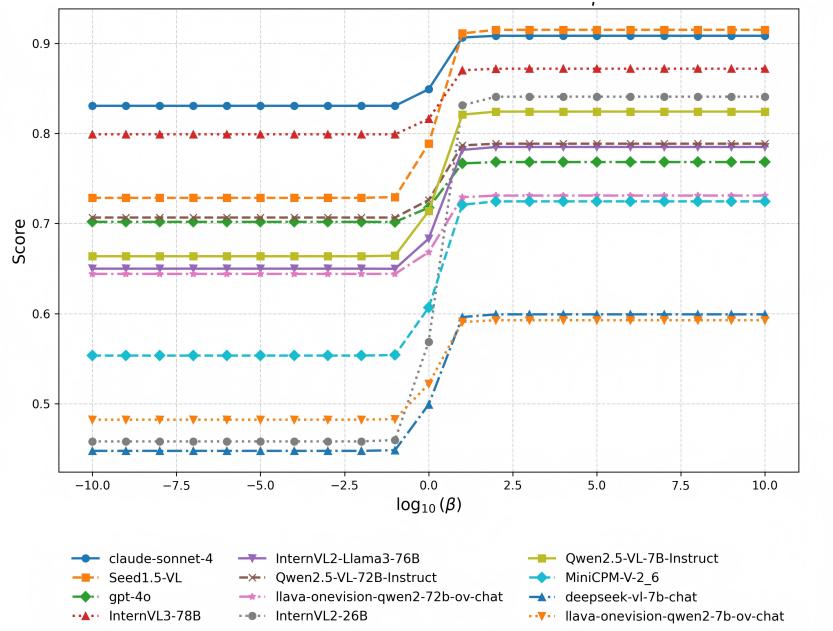
864 the impact of distractors, making its evaluation results not applicable in all scenarios. In contrast,
 865 MF_β can be more flexible and applicable to a wider range of situations by adjusting the β parameter.
 866

867 ANALYSIS OF MF_β CURVES 868

869 We plotted the MF_β curves of all models under varying values of β . Overall, the curves exhibit a
 870 monotonically increasing trend, indicating that within the current task setting, “selecting all correct
 871 options” is significantly more difficult than “avoiding incorrect options.” In other words, the models
 872 generally perform better at eliminating incorrect options than at fully covering the correct ones.
 873 We hypothesize that this phenomenon is related to the characteristics of the hard options: such
 874 options are more prone to being selected by the models, yet their penalty weight in the scoring
 875 scheme is relatively low, which to some extent “inflates” the overall scores. This effect is particularly
 876 pronounced for smaller-parameter models whose performance approaches random selection.
 877

878 **Intersection points.** The intersections between the curves carry critical implications: they reveal
 879 the trade-offs between different models in terms of “recalling correct options” versus “avoiding
 880 incorrect ones,” thereby providing guidance for model selection across application scenarios. For
 881 example, in recall-oriented tasks, models that perform better before the intersection point should
 882 be prioritized, whereas in precision-oriented tasks, models that excel after the intersection point are
 883 preferable.
 884

885 **Curve variability.** Taking InternVL2-26B as an example, its curve ranks relatively low when β
 886 is small, but improves markedly as β increases. This pronounced change highlights the model’s
 887 substantial variability across different evaluation emphases, reflecting a lack of balance—that is, an
 888 insufficiently stable ability to reconcile recall and precision.
 889



909 Figure 5: Model Performance on Task 3 under Different β
 910
 911

912 F DISCUSSION 913

914 F.1 IMPACT OF IRRELEVANT CHARTS 915

916 In these experiments, because the number of charts varies and high-quality charts with many visual
 917 tokens can cause the total input length to exceed the model’s context window, the model will begin
 918

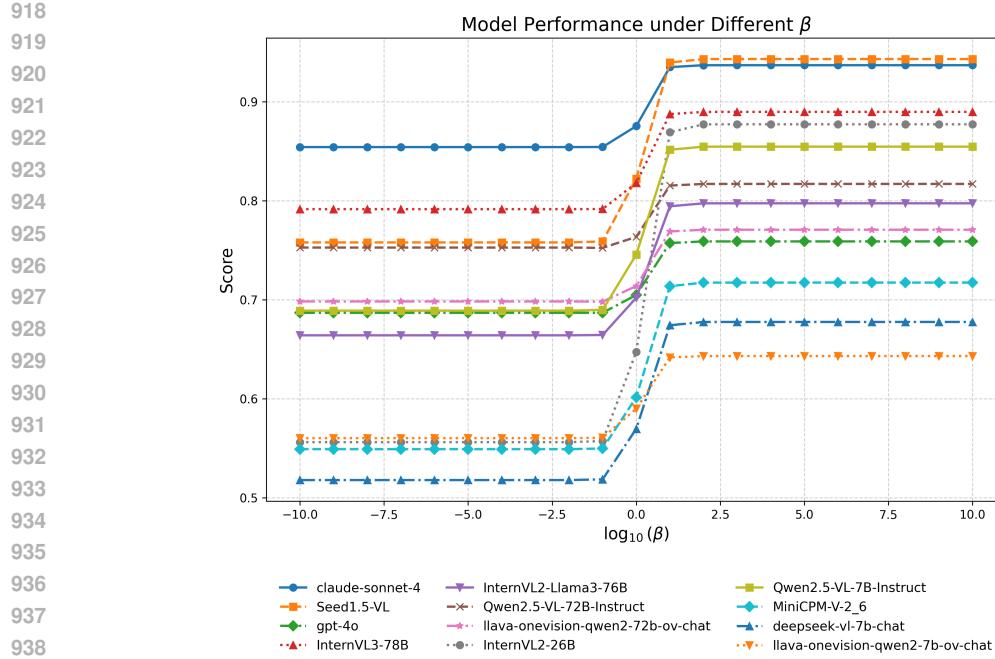


Table 4: Compare the question-answering performance when inputting all charts versus only the relevant charts.

Model	Trend Inference		Data Integration		Anomaly/Pattern Attr		Strategy Rec	
	involved	all	involved	all	involved	all	involved	all
Claude-Sonnet-4	70.00	69.11	60.99	61.38	84.92	85.32	87.53	87.85
Seed1.5-VL	72.44	72.83	67.66	66.74	78.87	79.08	82.22	81.95
GPT-4o	64.21	66.14	64.83	63.10	71.76	70.79	76.88	76.71
InternVL3-78B	73.21	60.93	67.50	64.68	81.62	81.56	81.78	80.57
InternVL2-Llama3-76B	59.91	52.46	51.59	44.36	68.33	67.23	70.19	70.04
Qwen2.5-VL-72B-Instruct	56.25	58.93	25.40	22.94	72.62	73.48	76.34	74.34
LLaVA-OV-72B	61.33	59.11	43.02	38.07	66.82	65.47	71.37	70.69
InternVL2-26B	54.46	46.33	31.03	22.90	56.85	54.53	64.72	64.37
Qwen2.5-VL-7B-Instruct	44.00	53.56	21.04	23.98	71.38	66.70	74.55	71.51
MiniCPM-V-2.6	49.88	52.89	23.29	27.48	60.67	59.85	60.13	59.44
DeepSeek-VL-7B	49.32	24.22	7.95	3.48	49.90	45.58	56.95	55.74
LLaVA-OV-7B	48.55	45.19	21.84	13.55	52.18	52.10	58.98	57.92

972 Table 5: Compare the question-answering performance when inputting all charts versus only the
973 relevant charts under the com2 metric.

Model	Trend Inference			Data Integration			Anomaly/Pattern Attr			Strategy Rec		
	involved	all	involved	all	involved	all	involved	all	involved	all	involved	all
Claude-Sonnet-4	70.00	69.11	60.99	61.38	65.04	62.56	69.67	69.15				
Seed1.5-VL	72.44	72.83	67.66	66.74	58.72	58.63	66.30	63.81				
GPT-4o	64.21	66.14	64.83	63.10	39.29	35.55	39.13	38.54				
InternVL3-78B	73.21	60.93	67.50	64.68	57.37	54.25	57.20	54.20				
InternVL2-Llama3-76B	59.91	52.46	51.59	44.36	34.24	33.99	36.95	36.18				
Qwen2.5-VL-72B-Instruct	56.25	58.93	25.40	22.94	32.93	33.19	39.44	37.26				
LLaVA-OV-72B	61.33	59.11	43.02	38.07	27.26	27.48	31.67	30.91				
InternVL2-26B	54.46	46.33	31.03	22.90	36.32	30.81	36.53	37.27				
Qwen2.5-VL-7B-Instruct	44.00	53.56	21.04	23.98	37.33	32.81	43.19	39.52				
MiniCPM-V-2_6	49.88	52.89	23.29	27.48	21.27	24.63	21.90	22.86				
DeepSeek-VL-7B	49.32	24.22	7.95	3.48	11.35	7.59	17.19	16.30				
LLaVA-OV-7B	48.55	45.19	21.84	13.55	9.20	8.97	12.03	10.13				

987
988 F.2 PERFORMANCE ACROSS DIFFERENT LANGUAGES
989990 Table 6: English Charts - Different Language QAs
991

Model	Trend Inference			Data Integration			Anomaly/Pattern Attr			Strategy Rec		
	en	zh	es	en	zh	es	en	zh	es	en	zh	es
Seed1.5-VL	72.44	71.56	67.56	67.66	68.76	67.04	78.87	81.05	79.22	82.22	85.17	79.19
InternVL3-78B	73.21	70.98	66.00	67.50	70.69	65.91	81.62	83.47	78.54	81.78	84.07	76.91
InternVL2-26B	54.46	46.09	45.07	31.03	28.57	21.72	56.85	61.36	56.91	64.72	66.25	60.77
MiniCPM-V-2_6	49.88	51.79	48.78	23.29	29.50	23.95	60.67	63.64	56.57	60.13	66.14	59.62
Qwen2.5-VL-7B	54.44	55.23	48.67	21.04	23.81	23.29	49.90	70.74	66.25	56.95	72.34	68.45
LLaVA-OV-7B	48.55	33.48	46.33	21.84	18.26	13.14	52.18	54.10	59.92	58.98	66.08	62.03

1000 Table 7: Different Language Charts - English QAs
1001

Model	Trend Inference			Data Integration			Anomaly/Pattern Attr			Strategy Rec		
	en	zh	es	en	zh	es	en	zh	es	en	zh	es
Seed1.5-VL	72.44	68.22	69.78	67.66	65.69	65.70	78.87	80.26	80.66	82.22	82.78	81.30
InternVL3-78B	73.21	68.97	69.87	67.50	67.26	58.03	81.62	80.88	80.93	81.78	82.46	81.97
InternVL2-26B	54.46	53.56	48.44	31.03	18.20	19.46	56.85	56.36	53.09	64.72	65.30	61.48
MiniCPM-V-2_6	49.88	50.00	49.56	23.29	24.53	25.29	60.67	61.81	61.81	60.13	62.21	61.29
Qwen2.5-VL-7B	54.44	52.35	52.67	21.04	14.41	20.95	49.90	70.11	72.74	56.95	72.65	74.33
LLaVA-OV-7B	48.55	41.67	46.85	21.84	5.11	14.25	52.18	52.03	52.95	58.98	55.96	57.67

1009 Table 8: The test results of English Q&A on mixed-language charts.
1010

Model	Trend Inference	Data Integration	Anomaly/Pattern Attr	Strategy Rec
Seed1.5-VL	70.38	65.11	80.45	82.83
InternVL3-78B	69.93	64.96	81.78	81.75
Qwen2.5-VL-7B-Instruct	54.12	19.38	70.76	72.71

1017 Table 9: English Charts - Different Language QAs Under the Com2 Metric
1018

Model	Trend Inference			Data Integration			Anomaly/Pattern Attr			Strategy Rec		
	en	zh	es	en	zh	es	en	zh	es	en	zh	es
Seed1.5-VL	72.44	71.56	67.56	67.66	68.76	67.04	58.72	63.57	59.39	66.30	69.63	62.52
InternVL3-78B	73.21	70.98	66.00	67.50	70.69	65.91	57.15	55.21	51.69	57.20	58.60	52.78
InternVL2-26B	54.46	46.09	45.07	31.03	28.57	21.72	36.32	26.54	26.76	36.53	31.22	29.44
MiniCPM-V-2_6	49.88	51.79	48.78	23.29	29.50	23.95	21.27	25.93	17.44	21.90	24.93	18.01
Qwen2.5-VL-7B	54.44	55.23	48.67	21.04	23.81	23.29	37.33	35.05	39.07	43.19	36.07	48.67
LLaVA-OV-7B	48.55	33.48	46.33	21.84	18.26	13.14	9.20	11.19	18.52	12.03	20.04	17.04

Table 10: Different Language Charts - English QAs Under the Com2 Metric

Model	Trend Inference			Data Integration			Anomaly/Pattern Attr			Strategy Rec		
	en	zh	es	en	zh	es	en	zh	es	en	zh	es
Seed1.5-VL	72.44	68.22	69.78	67.66	65.69	65.70	58.72	63.06	61.43	66.30	66.00	63.52
InternVL3-78B	73.21	68.97	69.87	67.50	67.26	58.03	57.15	56.34	53.75	57.20	55.47	57.23
InternVL2-26B	54.46	53.56	48.44	31.03	18.20	19.46	36.32	35.51	30.54	36.53	41.07	35.94
MiniCPM-V-2.6	49.88	50.00	49.56	23.29	24.53	25.29	21.27	23.19	23.04	21.90	21.86	21.41
Qwen2.5-VL-7B	54.44	52.35	52.67	21.04	14.41	20.95	37.33	35.26	41.41	43.19	39.20	43.52
LLaVA-OV-7B	48.55	41.67	46.85	21.84	5.11	14.25	9.20	8.43	8.75	12.03	8.48	11.15

F.3 EXPLORING MLLMs’ RETRIEVAL CAPABILITIES

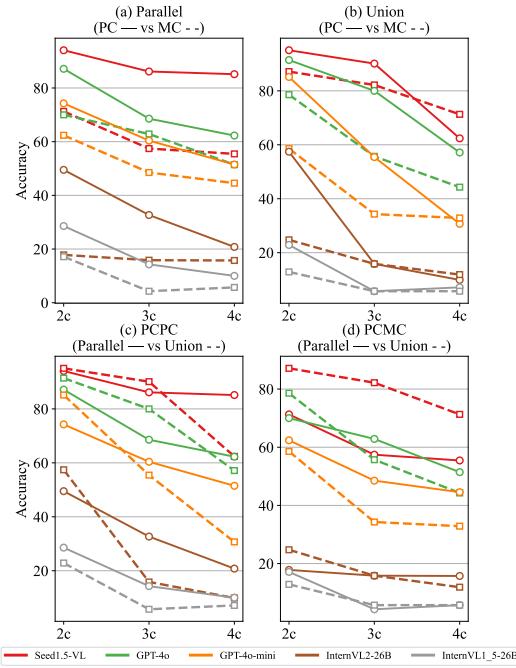


Figure 7: Compare the performance of MLLMs across four retrieval tasks. The x-axis represents the number of charts involved in the question, for example, 2c denotes two charts.

F.4 ERROR ANALYSIS

The detailed response statistics of the four proprietary models on Task 3 and Task 4 are presented in Tables 11 & 12.

Table 11: Omission and multiple-selection rates per option for each model for task 3 (%). In this context, “claude” refers to Claude-Sonnet-4, “seed” refers to Seed1.5-VL, “gpt” refers to GPT-4o, and “gemini” refers to Gemini-2.5-Pro.

Option	claude	seed	gpt	gemini
A	24.5 / 6.3	36.8 / 0.8	64.2 / 1.3	26.9 / 1.7
B	7.9 / 9.2	10.2 / 8.1	3.4 / 48.0	2.3 / 19.0
C	9.6 / 10.3	6.9 / 12.6	4.8 / 53.4	2.7 / 24.4
D	17.3 / 6.6	36.1 / 0.8	76.0 / 0.8	36.1 / 1.2
E	9.4 / 11.6	9.9 / 10.5	4.7 / 51.6	1.6 / 23.3
F	8.9 / 5.6	29.7 / 1.6	63.4 / 1.6	34.7 / 1.6
G	12.4 / 4.3	33.8 / 1.6	57.9 / 1.0	26.9 / 0.7

1080

1081 Table 12: Omission and multiple-selection rates per option for each model for task 4 (%). In this
1082 context, “claude” refers to Claude-Sonnet-4, “seed” refers to Seed1.5-VL, “gpt” refers to GPT-4o,
1083 and “gemini” refers to Gemini-2.5-Pro.

Option	claude	seed	gpt	gemini
A	11.6 / 4.4	40.7 / 1.2	62.8 / 0.4	31.6 / 0.5
B	4.3 / 9.1	2.9 / 7.9	1.9 / 49.2	3.0 / 21.1
C	2.5 / 6.8	2.5 / 10.4	1.0 / 47.4	1.2 / 30.5
D	8.9 / 2.7	32.8 / 0.0	62.0 / 1.2	25.2 / 0.9
E	4.2 / 11.6	7.8 / 12.8	2.1 / 58.1	0.0 / 32.7
F	9.1 / 4.4	25.6 / 0.7	58.0 / 2.9	26.3 / 3.8
G	7.3 / 4.0	35.8 / 0.0	58.9 / 2.3	34.2 / 0.0

1092

1093

G EXTENDED BENCHMARK.

1094

1095

G.1 TASK DESCRIPTION

1096

PARALLEL TYPE

1097

1098

• EXAMPLE

1099

1100

question: question_1 about chart_1 question_2 about chart_2 question_3 about chart_3 question_4 about chart_4

1101

answer: answer1. answer2. answer3. answer4.

1102

1103

• PCPC

1104

1105

A question contains multiple parallel sub-questions, each retrieving one piece of information from a single chart.

1106

1107

2c, 3c, and 4c refer to retrieving one piece of information from charts 2, 3, and 4, respectively, and providing separate answers.

1108

1109

• PCMC

1110

1111

A question contains multiple parallel sub-questions, each retrieving multiple pieces of information from a single chart.

1112

1113

2c, 3c, and 4c refer to retrieving at least one piece of information from charts 2, 3, and 4, respectively, and providing separate answers.

1114

UNION TYPE

1115

1116

1117

• EXAMPLE

1117

question: question about chart_1&2&3&4

1118

answer: answer.

1119

1120

• PCPC

1121

The question involves multiple charts, retrieving one piece of information from each chart, combining the information to perform a simple calculation, and outputting the final result.

1122

1123

• PCMC

1123

The question involves multiple charts, retrieving at least one piece of information from each chart, combining the information to perform a simple calculation, and outputting the final result.

1124

1125

G.2 PIPELINE

1126

1127

1. Manually select articles from Pew that contain at least four charts, and scrape both the articles and charts.
2. Extract chart information and perform manual sampling-based screening to ensure quality.
3. Automatically generate topic summaries based on the content of the article.
4. Generate question-and-answer pairs by utilizing the extracted chart information and setting different prompts based on the topics.

1134 5. Use scripts to check whether the number of charts involved in all question-and-answer
 1135 pairs and the amount of content related to each chart meet the expected criteria. If they do
 1136 not meet the criteria, regenerate the pairs and manually adjust them until all question-and-
 1137 answer pairs satisfy the conditions.
 1138 6. Use scripts to check the length of all answers, manually screen answers that exceed the
 1139 threshold, remove redundant parts, and retain concise answers.
 1140 7. Utilize the rationale of generated question-and-answer pairs to generate code-based com-
 1141 putation results, manually compare all inconsistent answers, correct the answers and ratio-
 1142 nales in the labels, and ensure the accuracy of the computation results.
 1143

1144 **G.3 DATA STATISTICS**

1145 We collected 101 articles from Pew Research Center, each containing at least four charts centered on
 1146 the same topic, spanning nine topics in total (Refer to Table 13), yielding 1,212 QA pairs. These QA
 1147 pairs span four distinct question types, comprehensively probing models’ capabilities in cross-chart
 1148 information retrieval.
 1149

1150 **Table 13: Category-wise Number of Articles**

1151 Category	1152 Number of Articles
1154 Economy & Work	11
1155 Politics & Policy	11
1156 Internet & Technology	11
1157 Family & Relationships	11
1158 Age & Generations	11
1159 Immigration & Migration	11
1160 Science	11
1161 News Habits & Media	12
1162 Other Topics	12

1188 H QAs PROMPTS

Task 1

```
1193     output_prompt = '''\nOutput according to the following json format:  
1194     {  
1195         "rationale": "Reasoning process",  
1196         "answer": "Final answer"  
1197     }\nPlease strictly output according to the json format.\n'''  
1198     prompt = image_tokens + question + output_prompt
```

Task 2

```
1203 output_prompt = ''''  
1204 Please answer according to the following steps, **all details must  
1205     be included and must not be omitted**:  
1206 1. Extract relevant information: List all data required for the  
1207     calculation and indicate which chart each data comes from.  
1208 2. Explain the association logic: Explain why these data are needed  
1209     .  
1210 3. Calculation process: Write out the detailed calculation process.  
1211 4. Conclusion summary: Answer according to the format required in  
1212     the question.  
1213 5. Output requirements: Extract the output format requirements  
1214     after the question.\n'''
```

prompt

Task 5

```
1219 text1 = "The options for the question are as follows: "
1220 hints = "(Multiple choice question) Please analyze the content of
1221     the chart and select the correct options based on the chart
1222     information. Note: Correct options should be supported by chart
1223     data; otherwise, they will be regarded as incorrect."
1224 output_prompt = '''\nAnswer the question according to the following
1225     json format:
1226     {
1227         "rationale": "Reasoning process",
1228         "answer": ["B", "C", "E"] /* Select the correct one or more
1229             options according to the actual situation */
1230     }\n'''
1231
1232 prompt = image_tokens + hints + question + text1 + answer_choices +
1233     output_prompt
```

Task 4

```
1235 text1 = "The options for the question are as follows: "
1236 hints = "(Multiple choice question) Please analyze the content of
1237     the chart and select the correct options based on the chart
1238     information. Note: Correct options should be supported by chart
1239     data; otherwise, they will be regarded as incorrect."
1240 output_prompt = '''\nAnswer the question according to the following
1241     json format:
```

```
1242
1243 {
1244     "rationale": "Reasoning process",
1245     "answer": ["B", "C", "E"] /* Select the correct one or more
1246         options according to the actual situation */
1247 }\n'''
1248
1249 prompt = image_tokens + hints + question + text1 + answer_choices +
        output_prompt
```

1253 The prompt for generating python calculation code based on the reasoning process in task 2
1254 is as follows:

```
1256 gen_code_prompt = '''
1257 The above is the reasoning process. Since large models are not good
1258 at calculations, ignore the calculated results in the reasoning
1259 process above, and generate executable Python code for the
1260 reasoning process.
1261 Please note:
1262 1. Strictly generate Python code based on the "rationale" process
1263 above.
1264 2. Ensure the code correctly reflects the reasoning of the "
1265 rationale" by using variables to replace the intermediate
1266 calculation results in the rationale, as the code execution is
1267 more accurate and avoids cumulative errors caused by using
1268 intermediate calculation results.
1269 3. The final output should strictly follow the required format,
1270 only printing the last output from the final 'print' statement,
1271 without any descriptive print statements. Below are some
1272 examples of the final print outputs for reference:
1273 a. Output format: "The answer should be presented as an integer
1274 .
1275 Incorrect final output 1: print(f"The final answer is: {answer}.")
1276 Incorrect final output 2: print(f"**{answer}**")
1277 Correct final output: print(f"{answer}")
1278 b. Output in percentage, rounded to 4 significant digits.
1279 Incorrect final output 1: print(f"In mobile search, the
1280 proportion of images is **{answer}%**.")
1281 Correct final output: print(f"{answer}%")
1282 c. Output in dollars, rounded to 3 decimal places. Answer: xx
1283 dollars.
1284 Incorrect final output 1: print(f"In Q1 2023, the in-app
1285 purchase revenue per download for mobile games on the
1286 App Store in Japan is {answer} USD.")
1287 Correct final output: print(f"{answer} USD")
1288 d. Output "Yes" or "No".
1289 Incorrect final output 1: print("True")
1290 Correct final output: print("Yes")
1291 e. Answer: xx times.
1292 Incorrect final output 1: print(f"{answer}")
1293 Correct final output: print(f"{answer} times")
1294 4. Be sure to distinguish between "significant digits" and "decimal
1295 places." "Significant digits" refer to all digits from the
1296 first non-zero digit to the last digit. "Decimal places" refer
1297 to the number of digits after the decimal point.
1298 5. If floating-point calculations are involved, please use the '
1299 decimal' library to avoid precision loss from floating-point
1300 operations.
1301 ...
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
21010
21011
21012
21013
21014
21015
21016
21017
21018
21019
21020
21021
21022
21023
21024
21025
21026
21027
21028
21029
21030
21031
21032
21033
21034
21035
21036
21037
21038
21039
210310
210311
210312
210313
210314
210315
210316
210317
210318
210319
210320
210321
210322
210323
210324
210325
210326
210327
210328
210329
210330
210331
210332
210333
210334
210335
210336
210337
210338
210339
210340
210341
210342
210343
210344
210345
210346
210347
210348
210349
210350
210351
210352
210353
210354
210355
210356
210357
210358
210359
210360
210361
210362
210363
210364
210365
210366
210367
210368
210369
210370
210371
210372
210373
210374
210375
210376
210377
210378
210379
210380
210381
210382
210383
210384
210385
210386
210387
210388
210389
210390
210391
210392
210393
210394
210395
210396
210397
210398
210399
210399
210400
210401
210402
210403
210404
210405
210406
210407
210408
210409
210410
210411
210412
210413
210414
210415
210416
210417
210418
210419
210420
210421
210422
210423
210424
210425
210426
210427
210428
210429
210430
210431
210432
210433
210434
210435
210436
210437
210438
210439
210440
210441
210442
210443
210444
210445
210446
210447
210448
210449
210450
210451
210452
210453
210454
210455
210456
210457
210458
210459
210460
210461
210462
210463
210464
210465
210466
210467
210468
210469
210470
210471
210472
210473
210474
210475
210476
210477
210478
210479
210480
210481
210482
210483
210484
210485
210486
210487
210488
210489
210490
210491
210492
210493
210494
210495
210496
210497
210498
210499
210499
210500
210501
210502
210503
210504
210505
210506
210507
210508
210509
210510
210511
210512
210513
210514
210515
210516
210517
210518
210519
210520
210521
210522
210523
210524
210525
210526
210527
210528
210529
210530
210531
210532
210533
210534
210535
210536
210537
210538
210539
210540
210541
210542
210543
210544
210545
210546
210547
210548
210549
210550
210551
210552
210553
210554
210555
210556
210557
210558
210559
210560
210561
210562
210563
210564
210565
210566
210567
210568
210569
210570
210571
210572
210573
210574
210575
210576
210577
210578
210579
210580
210581
210582
210583
210584
210585
210586
210587
210588
210589
210589
210590
210591
210592
210593
210594
210595
210596
210597
210598
210599
210599
210600
210601
210602
210603
210604
210605
210606
210607
210608
210609
210610
210611
210612
210613
210614
210615
210616
210617
210618
210619
210620
210621
210622
210623
210624
210625
210626
210627
210628
210629
210630
210631
210632
210633
210634
210635
210636
210637
210638
210639
210640
210641
210642
210643
210644
210645
210646
210647
210648
210649
210650
210651
210652
210653
210654
210655
210656
210657
210658
210659
210660
210661
210662
210663
210664
210665
210666
210667
210668
210669
210670
210671
210672
210673
210674
210675
210676
210677
210678
210679
210680
210681
210682
210683
210684
210685
210686
210687
210688
210689
210689
210690
210691
210692
210693
210694
210695
210696
210697
210698
210699
210699
210700
210701
210702
210703
210704
210705
210706
210707
210708
210709
210710
210711
210712
210713
210714
210715
210716
210717
210718
210719
210719
210720
210721
210722
210723
210724
210725
210726
210727
210728
210729
210729
210730
210731
210732
210733
210734
210735
210736
210737
210738
210739
210739
210740
210741
210742
210743
210744
210745
210746
210747
210748
210749
210749
210750
210751
210752
210753
210754
210755
210756
210757
210758
210759
210759
210760
210761
210762
210763
210764
210765
210766
210767
210768
210769
210769
210770
210771
210772
210773
210774
210775
210776
210777
210778
210778
210779
210779
210780
210781
210782
210783
210784
210785
210786
210787
210788
210789
210789
210790
210791
210792
210793
210794
210795
210796
210797
210798
210799
210799
210800
210801
210802
210803
210804
210805
210806
210807
210808
210809
210809
210810
210811
210812
210813
210814
210815
210816
210817
210818
210819
210819
210820
210821
210822
210823
210824
210825
210826
210827
210828
210829
210829
210830
210831
210832
210833
210834
210835
210836
210837
210838
210839
210839
210840
210841
210842
210843
210844
210845
210846
210847
210848
210849
210849
210850
210851
210852
210853
210854
210855
210856
210857
210858
210859
210859
210860
210861
210862
210863
210864
210865
210866
210867
210868
210869
210869
210870
210871
210872
210873
210874
210875
210876
210877
210878
210878
210879
210879
210880
210881
210882
210883
210884
210885
210886
210887
210888
210889
210889
210890
210891
210892
210893
210894
210895
210896
210897
210898
210898
210899
210899
210900
210901
210902
210903
210904
210905
210906
210907
210908
210909
210909
210910
210911
210912
210913
210914
210915
210916
210917
210918
210919
210919
210920
210921
210922
210923
210924
210925
210926
210927
210928
210929
210929
210930
210931
210932
210933
210934
210935
210936
210937
210938
210939
210939
210940
210941
210942
210943
210944
210945
210946
210947
210948
210949
210949
210950
210951
210952
210953
210954
210955
210956
210957
210958
210959
210959
210960
210961
210962
210963
210964
210965
210966
210967
210968
210969
210969
210970
210971
210972
210973
210974
210975
210976
210977
210978
210978
210979
210979
210980
210981
210982
210983
210984
210985
210986
210987
210988
210989
210989
210990
210991
210992
210993
210994
210995
210996
210997
210998
210998
210999
210999
211000
211001
211002
211003
211004
211005
211006
211007
211008
211009
211009
211010
211011
211012
211013
211014
211015
211016
211017
211018
211019
211019
211020
211021
211022
211023
211024
211025
211026
211027
211028
211029
211029
211030
211031
211032
211033
211034
211035
211036
211037
211038
211039
211039
211040
211041
211042
211043
211044
211045
211046
211047
211048
211049
211049
211050
211051
211052
211053
211054
211055
211056
211057
211058
211059
211059
211060
211061
211062
211063
211064
211065
211066
211067
211068
211069
211069
211070
211071
211072
211073
211074
211075
211076
211077
211078
211078
211079
211079
211080
211081
211082
211083
211084
211085
211086
211087
211088
211089
211089
211090
211091
211092
211093
211094
211095
211096
211097
211098
211098
211099
211099
211100
211101
211102
211103
211104
211105
211106
211107
211108
211109
211109
211110
211111
211112
211113
211114
211115
211116
211117
211118
211119
211119
211120
211121
211122
211123
211124
211125
211126
211127
211128
211129
211129
211130
211131
211132
211133
211134
211135
211136
211137
211138
211139
211139
211140
211141
211142
211143
211144
211145
211146
211147
211148
211149
211149
211150
211151
211152
211153
211154
211155
211156
211157
211158
211159
211159
211160
211161
211162
211163
211164
211165
211166
211167
211168
211169
211169
211170
211171
211172
211173
211174
211175
211176
211177
211178
211178
211179
211179
211180
211181
211182
211183
211184
211185
211186
211187
211188
211189
211189
211190
211191
211192
211193
211194
211195
211196
211197
211198
211198
211199
211199
211200
211201
211202
211203
211204
211205
211206
211207
211208
211209
211209
211210
211211
211212
211213
211214
211215
211216
211217
211218
211219
211219
211220
211221
211222
211223
211224
211225
211226
211227
211228
211229
211229
211230
211231
211232
211233
211234
211235
211236
211237
211238
211239
211239
211240
211241
211242
211243
211244
211245
211246
211247
211248
211249
211249
211250
211251
211252
211253
211254
211255
211256
211257
211258
211259
211259
211260
211261
211262
211263
211264
211265
211266
211267
211268
211269
211269
211270
211271
211272
211273
211274
211275
211276
211277
211278
211279
211279
211280
211281
211282
211283
211284
211285
211286
211287
211288
211289
211289
211290
211291
211292
211293
211294
211295
211296
211297
211298
211298
211299
211299
211300
2113
```

1296 I TASK 3&4 SYNTHESIS METHOD
12971298 Algorithm 1: QA_Pair_Synthesis
1299

```

1300
1301 Input:
1302     C <= Original code rendering scripts for multiple charts
1303     Q_defs <= Definitions of question types (templates + few-shot
1304         examples)
1305     M <= Reasoning model with chain-of-thought capabilities
1306 Output:
1307     QA_dataset <= Complete set of question-answer pairs (including
1308         correct options and distractors)
1309
1310     1. // Step 1: Extract Gold Tables
1311     2. gold_tables <= ExtractTablesFromCode(C)
1312
1313     3. // Step 2: Generate questions and correct answers
1314     4. for each table_set in gold_tables:
1315         5.     context <= Concatenate(Q_defs, table_set)
1316         6.     prompt_correct <= BuildPrompt(context, task="generate
1317             question + correct answer + reasoning")
1318         7.     (Q, A_correct, Reasoning) <= M.generate(prompt_correct)
1319         8.     store CorrectPair = (Q, A_correct, Reasoning)
1320
1321     9. // Step 3: Generate Easy and Hard distractors
1322     10. for each CorrectPair in dataset:
1323         11.     // Easy distractors
1324         12.     prompt_easy <= BuildPrompt(
1325             context=(table_set, CorrectPair),
1326             instructions="generate 1 or 2 easy distractors;
1327             responses that do not include information from the table, only
1328             based on internal knowledge"
1329         13.     )
1330         14.     D_easy <= M.generate(prompt_easy, count=2)
1331
1332         15.     // Hard distractors
1333         16.     prompt_hard <= BuildPrompt(
1334             context=(table_set, CorrectPair),
1335             instructions="generate 1 or 2 hard distractors;
1336             logically or numerically incorrect but superficially dependent
1337             on the table"
1338         17.     )
1339         18.     D_hard <= M.generate(prompt_hard, count=2)
1340
1341         19.     // Merge and shuffle all options
1342         20.     all_choices <= [A_correct] U D_easy U D_hard
1343         21.     random.shuffle(all_choices) // Shuffle all options
1344             randomly
1345
1346         22.     // Label the correct option after shuffling
1347         23.     correct_option <= all_choices.index(A_correct) // Find the
1348             position of the correct option after shuffling
1349
1350         24.     // Add the question, shuffled options, and correct option
1351             to the dataset
1352         25.     QA_dataset.add( Q, all_choices, correct_option )
1353
1354         26.     return QA_dataset
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
3336
3337
3338
3339
3340
3341
3342
3343
3344
3345
3346
3347
3348
3349
3350
3351
3352
3353
3354
3355
3356
3357
3358
3359
3360
3361
3362
3363
3364
3365
3366
3367
3368
3369
3370
3371
3372
3373
3374
3375
3376
3377
3378
3379
3380
3381
3382
3383
3384
3385
3386
3387
3388
3389
3390
3391
3392
3393
3394
3395
3396
3397
3398
3399
3400
3401
3402
3403
3404
3405
3406
3407
3408
3409
3410
3411
3412
3413
3414
3415
3416
3417
3418
3419
3420
3421
3422
3423
3424
3425
3426
3427
3428
3429
3430
3431
3432
3433
3434
3435

```