Chart-RL: Policy Optimization Reinforcement Learning for Enhanced Visual Reasoning in Chart Question Answering with Vision Language Models

Yunfei Bai, Amit Dhanda, Shekhar Jain
Amazon
{byunfei, amdhanda, shekhajt}@amazon.com

Abstract

The recent advancements in Vision Language Models (VLMs) have demonstrated progress toward true intelligence requiring robust reasoning capabilities. Beyond pattern recognition, linguistic reasoning must integrate with visual comprehension, particularly for Chart Question Answering (CQA) tasks involving complex data visualizations. Current VLMs face significant limitations in CQA, including imprecise numerical extraction, difficulty interpreting implicit visual relationships, and inadequate attention mechanisms for capturing spatial relationships in charts. In this work, we address these challenges by presenting Chart-RL, a novel reinforcement learning framework that enhances VLMs' chart understanding through feedback-driven policy optimization of visual perception and logical inference. Our key innovation includes a comprehensive framework integrating Reinforcement Learning (RL) from Policy Optimization techniques along with adaptive reward functions, that demonstrates superior performance compared to baseline foundation models and competitive results against larger state-of-the-art architectures. We also integrated Parameter-Efficient Fine-Tuning through Low-Rank Adaptation (LoRA) in the RL framework that only requires single GPU configurations while preserving performance integrity. We conducted extensive benchmarking across open-source, proprietary, and state-of-the-art closed-source models utilizing the ChartQAPro dataset. The RL fine-tuned Qwen3-VL-4B-Instruct model achieved an answer accuracy of 0.634, surpassing the 0.580 accuracy of the Qwen3-VL-8B-Instruct foundation model despite utilizing half the parameter count, while simultaneously reducing inference latency from 31 seconds to 9 seconds. We also performed comprehensive comparative analysis of Chain of Thought (CoT) reasoning performance across those models, demonstrating that the Chart-RL framework significantly enhanced the visual reasoning.

1 Introduction

The recent advancements in Large Language Models (LLMs) have been towards demonstrating true intelligence that requires more than pattern recognition. True intelligence demands robust reasoning capabilities, which involve processing existing knowledge through logical deduction to arrive at novel conclusions (Prabhakar et al., 2024). The reasoning capabilities are essential for complex, multi-step tasks such as mathematical problem solving, logical inference, and scientific analysis.

The reasoning challenges are magnified in the domain of VLMs, where linguistic reasoning must be fused with visual comprehension. Visual reasoning for VLMs involves two interconnected steps: accurately recognizing key information from visual inputs and conducting complex reasoning based on that extracted data. Data visualizations, particularly charts, are high-density carriers of complex information. Interpreting them is not a trivial task as it necessitates a sophisticated sequence of visual reasoning capabilities that combine chart reading (recognition) and chart understanding (comprehension) to decipher underlying

data patterns. Thereby, Chart Question Answering (CQA) has emerged as the standard, rigorous task to probe the visual reasoning capabilities of VLMs.

There are challenges in VLMs when handling CQA tasks, including less precise numerical extraction and computation for complex visual elements such as overlapping lines or stacked bars, implicit relationships between visual elements such as trends over time or proportional relationships, and the context-dependent nature of chart interpretation. This is due to the attention mechanisms in the VLM foundation model may not be able to effectively capture spatial relationships and hierarchical structures in charts, while the sequential processing of visual information does not align well with the non-linear nature of chart interpretation. Furthermore, for complex multi-step reasoning questions that require correlation of information from different parts of the chart or performing comparative analyses across multiple data points, VLMs often lack the ability to maintain and update intermediate results, along with the combination of numerical computation and visual understanding.

To address these challenges, in this paper, we present Chart-RL, a reinforcement learning framework to enhance VLMs' chart understanding and reasoning capability by creating a feedback-driven learning loop to optimize visual perception and logical inference. Our approach allows the model to learn complex visual-textual relationships inherent in charts, including data trends, correlations, and contextual information, while simultaneously improving its ability to generate coherent explanations and draw logical conclusions from visual data representations. It enables the VLMs to explore different reasoning pathways and reinforces successful strategies for chart analysis, leading to more robust performance on tasks requiring both visual comprehension and analytical reasoning. Our key contributions include:

- We implemented Chart-RL with three advanced policy optimization techniques: Group-based Reinforcement Learning from Policy Optimization (GRPO), Direct Advantage Policy Optimization (DAPO), and Group Sequence Policy Optimization (GSPO). This comprehensive framework demonstrates enhanced performance in complex chart reasoning tasks, achieving superior accuracy compared to baseline foundation models and competitive performance relative to larger state-of-the-art architectures.
- Our reinforcement learning methodology incorporated Parameter Efficient Fine-Tuning (PEFT) through Low-Rank Adaptation (LoRA), which significantly reduces computational overhead and training costs while preserving model performance integrity. This approach enables efficient model optimization without compromising the quality of understanding and reasoning capabilities of the chart.
- We conducted benchmarking across open-source, proprietary, and state-of-the-art closed-source models utilizing the ChartQAPro dataset, demonstrating the CQA accuracy improvement and latency reduction.
- We performed a comprehensive comparative analysis of the performance of chart reasoning in open-source foundation models, proprietary foundation models, and state-of-the-art closed-source models, establishing the efficacy of policy optimization-based reinforcement learning to improve model performance.

2 Related Work

2.1 Chart question answering benchmarks

CQA tasks have evolved from synthetic, template-driven tasks toward more diverse settings. Earlier datasets for CQA, such as FigureQA (Kahou et al., 2017) and DVQA (Kafle et al., 2018), were mostly generated with synthetic data using a plotting software, where the questions are generated using a small number of templates and the answers from a fixed set of vocabulary. PlotQA (Methani et al., 2019) expanded coverage to scientific plots and emphasized real-valued answers. More recent benchmarks aim to better reflect real-world chart diversity and question styles. ChartX/ChartVLM (Xia et al., 2024) dataset was created by synthesizing data and questions using GPT-4 to generate multiple chart types and a broader suite of

chart reasoning tasks, including summarization and redrawing. CharXiv (Wang et al., 2024) highlighted gaps in multimodal LLM chart understanding on figures drawn from scientific papers. ChartMuseum (Tang et al., 2025) provides targeted evaluations of the reasoning of the LVLM chart and helps characterize persistent failure modes beyond surface-level extraction. ChartQA (Masry et al., 2022) included real-world data from sources like Our World in Data, Statista, OECD, and Pew Research, and includes both human and machine-authored questions. ChartQAPro (Masry et al., 2025) further increased difficulty and diversity with real-world charts sourced from diverse online platforms with a wide variety of human-verified question-answer pairs that are factoid, multiple-choice, hypothetical, conversational, multi-chart, and unanswerable cases.

2.2 From chart reading to chart understanding

Chart understanding is closely related to document and infographic VQA, where OCR, layout, and grounded language understanding interact. DocVQA (Mathew et al., 2021b) focused on questions over document images with strong dependence on textual content and spatial arrangement. InfographicVQA (Mathew et al., 2021a) expanded to visually rich layouts that combine text blocks, icons, and charts. Pix2Struct (Lee et al., 2023) suggested that large-scale pretraining on screenshot parsing can provide useful inductive biases for these structured, text-heavy visual domains. ChartOCR (Luo et al., 2021) utilized chart-specific pipelines that combine detection, OCR, and structural parsing. DePlot (Liu et al., 2023) translated charts into tables that can be consumed by a language model for subsequent logical and numerical reasoning. SIMPLOT (Kim et al., 2025) improved CQA by distilling essential chart information and reducing irrelevant visual content.

2.3 Multimodal understanding and reasoning in LLMs and VLMs

Progress in CQA is tightly coupled to advances in general-purpose multimodal LLMs (MLLMs). Proprietary frontier models, such as Anthropic Claude models (Anthropic, 2025a;b), Google Gemini models (Google DeepMind, 2025), and OpenAI GPT models (OpenAI, 2025) often provide strong zero-shot performance, however are difficult to customize due to limited transparency. In parallel, open-source VLMs such as InternVL (Wang et al., 2025), Janus (Wu et al., 2024), and Qwen-VL (Alibaba, 2025; Bai et al., 2025), combine LLM with vision encoders to understand, interpret, and relate visual data, meanwhile enabling controlled adaptation. Chain-of-Thought (CoT) prompting (Wei et al., 2022) can improve performance on tasks requiring multi-hop deduction and arithmetic. ReAct (Yao et al., 2022) extends CoT by interleaving reasoning with actions to update intermediate state and recover from errors. Reinforcement Learning with Verifiable Rewards (RLVR) (Wen et al., 2025) is an effective approach that incentivizes correct and logical thought chains through rewards in verifiable domains. Among the popular RLVR methods, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) captures the inherent variability in reasoning processes and trains the model to consistently outperform its own average performance. Direct Advantage Policy Optimization (DAPO) (Yu et al., 2025) employs a higher clip ratio to encourage more diverse and exploratory thinking processes. Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025) shifts the optimization to the sequence level for more sophisticated reasoning and computational efficiency improvements.

3 Methodology

We introduce a VLM reinforcement learning framework comprising three distinct pipelines, as shown in Figure 1. The first constitutes a data pre-processing pipeline for training and inference data preparation. The second is an inference pipeline that facilitates chart understanding and reasoning tasks, formally defined as follows. Given a chart image I and a corresponding textual question Q, the objective function F seeks to derive a textual and/or numerical answer A to the question Q, accompanied by a Chain of Thought reasoning procedure R, where $A \in R$. This relationship is expressed as:

$$F : (I, Q) \rightarrow (A, R) \tag{1}$$

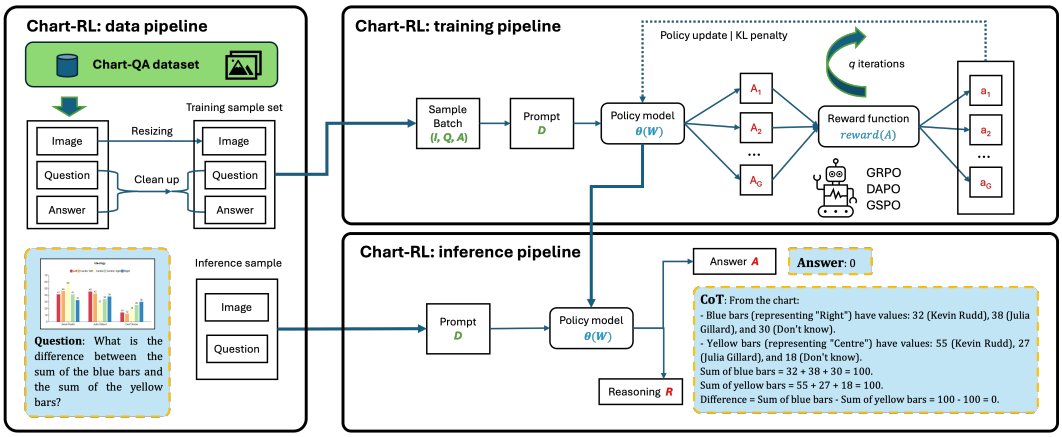


Figure 1: Overview of Chart-RL Framework. We first use data pipeline to construct training samples from pre-processed CQA dataset, then factorize policy optimization based reinforcement learning (GRPO/DAPO/GSPO) in the training pipeline. The trained policy model is then used in the inference pipeline to complete question-answering task through the enhanced visual understanding and reasoning by the VLMs.

The third pipeline encompasses the fine-tuning methodology for VLM chart reasoning optimization and generalization. While traditional approaches employ an initial stage to establish foundational reasoning paradigms through Supervised Fine-Tuning (SFT) using human-labeled or large-scale state-of-the-art MLLM-distilled datasets (A, R), our methodology directly applies fine-tuning to foundation models without requiring preliminary SFT stages. Consequently, our training pipeline employs policy optimization-based reinforcement learning to refine VLM chart interpretation capabilities and improve the accuracy and reliability of the answers for CQA tasks.

Our training pipeline incorporates GRPO, DAPO and GSPO, which eliminate the requirement for value functions. This approach enables VLMs to develop autonomous reasoning capabilities through iterative policy refinement based on reward signals derived from chart interpretation accuracy.

Formally, for each input training sample (I, Q), we infer the model policy to obtain a group of G candidate responses $\{A_1, A_2, \ldots, A_G\}$. Each answer A_i receives a reward $\text{rew}(A_i)$, and these rewards are utilized to compute a normalized relative advantage a_i for each sample:

$$a_i = \frac{\text{rew}_i - \mu(\text{rew}_i, \ldots, \text{rew}_G)}{\sigma(\text{rew}_i, \ldots, \text{rew}_G)} \tag{2}$$

Subsequently, the VLM parameters are updated to increase the probability of actions with positive advantages, while a Kullback-Leibler (KL) divergence penalty against the foundation model used as reference model ensures training stability.

3.1 Image pre-processing

To optimize computational efficiency while maintaining visual fidelity, we implemented an image preprocessing pipeline that standardizes the chart dimensions. The original chart images were resized to maintain a fixed width of 300 pixels while preserving the aspect ratio, resulting in proportionally scaled heights. This reduction significantly decreased the memory footprint and processing time compared to using full-resolution images (average original size greater than 1000 pixels in width). In this way we fit the model fine-tuning in a single GPU while achieving a 70% reduction in memory usage and inference time.

Algorithm 1 Chart-RL Training

Require: chart image I, textual question Q, initial policy model W, reward function Rew
1: Pre-process the image to reduce the image size down to fixed width of 300 pixels
2: **for** step $= 1 \ldots M$ **do**
3: Sample a batch (I_b, Q_b) from (I, Q)
4: Create prompt D
5: Infer the model policy to get a group of G candidate responses $\{A_1, A_2, \ldots, A_G\}$
6: Calculate reward for each candidate, to generate $\{\text{rew}(A_1), \text{rew}(A_2), \ldots, \text{rew}(A_G)\}$
7: Compute relative advantage a_i
8: **for** iteration $= 1, \ldots, q$ **do**
9: Update policy model W by adjusting model parameters to increase positive advantage
10: **end for**
11: **end for**
Ensure: Updated policy model W

3.2 Policy optimization techniques

We employed three RLVR methods in our framework, users can select the most appropriate method based on their tasks.

- **GRPO** generates groups of responses and evaluates each against the average score of the group, rewarding those performing above average while penalizing those below. This relative comparison mechanism creates a competitive dynamic that encourages the model to produce higher-quality reasoning.
- **DAPO** employs approximately 30% higher clip ratio than GRPO to eliminate less meaningful samples and improve overall training efficiency, applies token-level policy gradient loss to provide more granular feedback on lengthy reasoning chains, and incorporates overlong reward shaping to discourage excessively verbose responses.
- **GSPO** utilizes the sequence-level importance weights rather than the token-level weights in GRPO, to avoid high variance and unstable gradients especially for the long text outputs in the Mixture-of-Experts (MoE) model training.

3.3 Reward function design

For Reinforcement Learning, we design the reward function $\text{Rew}(A_i)$ as a composite of three types of reward: format reward for answer format, accuracy reward for answer accuracy, and reasoning reward for reasoning in response.

$$\text{rew}(A_i) = \text{rew}_{\text{format}}(A_i) + \text{rew}_{\text{accuracy}}(A_i) + \text{rew}_{\text{reasoning}}(A_i) \tag{3}$$

The format reward function $\text{rew}_{\text{format}}(A_i)$ implements a binary scoring mechanism to promote interpretability of the response and structural consistency. Responses that comply with the established template—featuring CoT reasoning and response—receive a positive reward of 1. Non-compliant responses are penalized with a reward of 0.

For evaluation of the accuracy $\text{rew}_{\text{accuracy}}(A_i)$ and reasoning $\text{rew}_{\text{reasoning}}(A_i)$ rewards, we implement a unified Large Language Model-based reward mechanism. This reward function employs gpt-oss-120B, a model from a different family than those studied in our experiments, as an automated evaluator (see Appendix B for the judge model and prompt details). The reward function employs a three-tier categorical scoring system: a reward of 1 is assigned when the predicted answer is correct AND the reasoning process can logically lead to the correct answer; a reward of 0.5 is assigned when the predicted answer is correct BUT the reasoning process cannot logically support the correct answer; and a reward of 0 is assigned when the predicted answer is incorrect, regardless of the reasoning quality.

It is important to note that although our LLM-based reward mechanism utilizes another LLM, it is fundamentally different from an unsupervised critic. The reward function

receives the ground truth answer from the ChartQAPro dataset and evaluates the predicted response against this known correct answer. The LLM judge serves as a sophisticated semantic matching function and reasoning verifier—assessing whether the predicted answer is equivalent to the ground truth while tolerating surface-level variations in phrasing, formatting, and numerical precision. It further evaluates the model's ability to generate valid reasoning for query resolution. This is analogous to how the environment provides ground truth signals in traditional RL; here the dataset's human-verified answers constitute the environment's ground truth, and the LLM judge operationalizes the comparison and validation.

3.4 PEFT/LoRA adoption

Within our Chart-RL framework, we adopt LoRA training representing a paradigmatic shift toward computationally efficient model optimization, addressing the resource constraints inherent in traditional full weight training. In our policy optimization learning process, we employ rank decomposition matrices to update only a small subset of model parameters, reducing trainable parameters to less than 0.5% of the original model parameter size while maintaining the frozen base model weights.

This configuration reduces susceptibility to catastrophic forgetting problems common in traditional fine-tuning approaches, also enhances training efficacy by enabling efficient task switching through small, swappable matrices that require minimal storage compared to full model copies. Meanwhile, this design enables our RL implementation on single GPU configurations rather than requiring distributed computing infrastructure, resulting in rapid experimentation and deployment cycles.

4 Experiments

We conducted a comprehensive experimental evaluation across multiple MLLMs and VLMs, including RL fine-tuned variants of Qwen3-VL-4B-Instruct trained via GRPO, DAPO, and GSPO methodologies. The ChartQAPro dataset was used to assess chart image understanding and question-answering capabilities. All model training and inference procedures were executed on a single GPU compute node with 24GB memory, demonstrating computational efficiency. Model performance was benchmarked using two primary metrics: answer accuracy and inference latency.

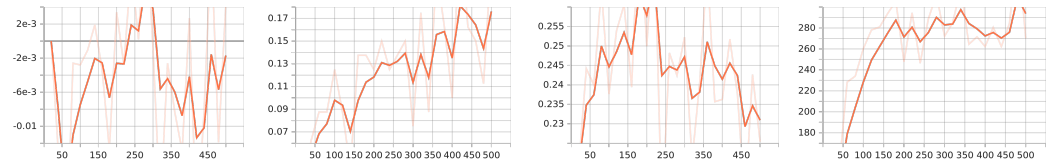
4.1 Data preparation

The ChartQAPro dataset contains diverse real-world charts with human-verified question-answer pairs. The dataset consists of 1,948 samples, with the last 500 samples (excluding unanswerable questions) used as a test set and the remaining samples for training. The question types are distributed across five categories: factoid questions (55%, n=1,081) that focus on direct data extraction and simple calculations, conversational questions (16%, n=311) that require contextual understanding, fact-checking questions (13%, n=244) that test verification capabilities, multiple-choice questions (11%, n=214) that evaluate discriminative reasoning, and hypothetical questions (5%, n=98) that assess predictive reasoning. The charts are sourced from diverse online platforms, covering various visualization types including bar charts, line charts, pie charts, and scatter plots.

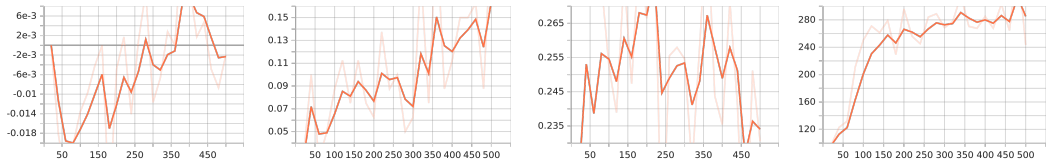
4.2 Policy optimization

For training implementation, we configured LoRA adapters with rank 256 and alpha value of 1024 for parameter-efficient fine-tuning, targeting the query and value projection layers with a dropout rate of 0.05. The GRPO training was conducted with a learning rate of 1e-5 and a batch size of 2 per device, using bf16 precision. The model was trained for one epoch with maximum prompt and completion lengths of 8192 and 1024 tokens respectively.

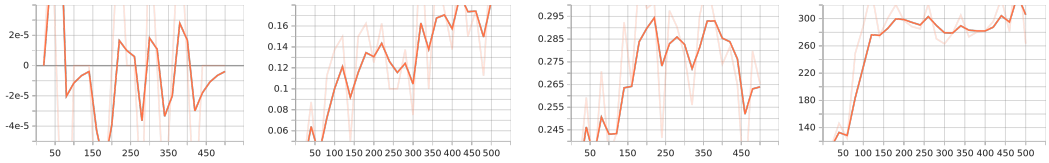
The training processes for GRPO, DAPO and GSPO on Qwen3-VL-4B-Instruct, including training loss, reward, entropy and completion mean length, are depicted in Figure 2. For



(a) GRPO: Loss (b) GRPO: Reward (c) GRPO: Entropy (d) GRPO: Length



(e) DAPO: Loss (f) DAPO: Reward (g) DAPO: Entropy (h) DAPO: Length



(i) GSPO: Loss (j) GSPO: Reward (k) GSPO: Entropy (l) GSPO: Length

Figure 2: Reinforcement Learning processes by GRPO, DAPO and GSPO. The training loss decreases as the model learns. The RL reward shows an upward trend during training. The policy entropy demonstrates a stable non-zero level indicating active exploration. The episode mean length shows rapid and stable convergence.

all three optimization approaches, the reward trajectory exhibited a consistent upward trend, while the entropy values demonstrated a monotonic decrease, indicating a progressive stabilization of the learning dynamics. The completion mean length stabilized after approximately 150 training steps, demonstrating rapid convergence.

4.3 Quantitative analysis on CQA

The experimental results demonstrate significant improvements achieved through our reinforcement learning framework, as shown in Table 1. We benchmarked the question-answer accuracy of RL fine-tuned models against two distinct baseline categories: state-of-the-art (SOTA) MLLMs and popular open-source VLMs.

In a vertical comparison, reinforcement learning fine-tuning substantially improved the accuracy of the Qwen3-VL-4B-Instruct foundation model. The GSPO-fine-tuned variant improved accuracy from baseline 0.396 to 0.622, while the GRPO and DAPO fine-tuned variants achieved accuracies of 0.627 and 0.634, respectively. Notably, despite utilizing only half the parameter count, all three fine-tuned 4B models surpassed the performance of the larger Qwen3-VL-8B-Instruct baseline (accuracy: 0.580).

The fine-tuned models exhibited marked improvements in computational efficiency, achieving inference latencies between 9.48 and 9.69 seconds—a 71% reduction relative to the Qwen3-VL-8B-Instruct model (31.59 seconds). Although closed-source SOTA models retain

Table 1: Model performance comparison. The first set are SOTA proprietary MLLMs, the second set are open-source models, and the third set are our fine-tuned Qwen models.

Model Name	**Acc (0-1)**	**Latency (s)**
SOTA MLLM		
Claude Sonnet 3.7	0.769	–
Claude Sonnet 4.5	0.750	–
Open-source VLM		
Qwen2.5-VL-7B-Instruct	0.511	6.72
Qwen3-VL-4B-Instruct	0.396	10.04
Qwen3-VL-8B-Instruct	0.580	31.59
Janus-pro-7B	0.320	18.25
InternVL3.5-8B-MPO	0.420	27.27
Llava-v1.6	0.570	6.19
Fine-tuned VLM		
Qwen3-VL-4B-Instruct-GRPO	0.627	9.84
Qwen3-VL-4B-Instruct-DAPO	0.634	9.48
Qwen3-VL-4B-Instruct-GSPO	0.622	9.69

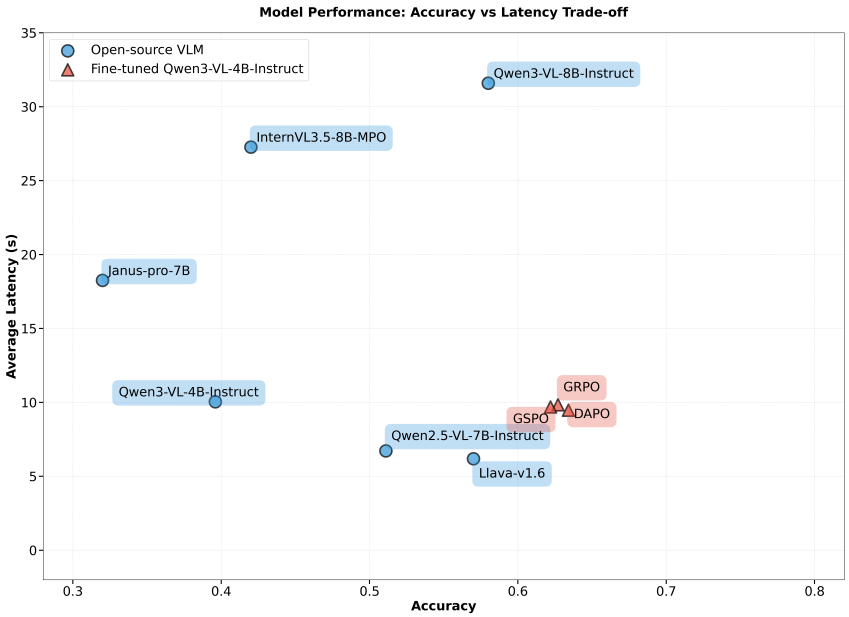


Figure 3: Latency-accuracy trade-off analysis. Our RL-enhanced models achieve an optimal balance between qualitative performance and computational efficiency.

an accuracy advantage (0.769 vs. 0.634), our methodology delivers competitive performance while substantially reducing resource requirements. These results were obtained using only a single GPU with 24GB memory. The latency-accuracy analysis presented in Figure 3 illustrates that our models cluster within the Pareto-optimal region.

5 Chart image understanding and reasoning analysis

The enhanced reasoning capabilities are evident in the improved accuracy scores. The GRPO, DAPO and GSPO variant models excelled in handling complex multi-step reasoning tasks, demonstrating more coherent Chain of Thought reasoning compared to baseline models.

We present three examples demonstrating how policy optimization techniques effectively enhance both the accuracy of final answers and the quality of intermediate reasoning steps.

5.1 Example 1: Open-ended question

Question: *What is the difference between the sum of the blue bars and the sum of the yellow bars?*

Ground Truth: 0

Model Responses: A SOTA MLLM exhibited a visual perception error, misidentifying a yellow bar from the "Don't know" category by selecting the bar with value 12 instead of the correct value of 18. This misclassification propagated through the computational chain, yielding an incorrect summation of $55 + 27 + 12 = 94$ rather than the ground truth of $55 + 27 + 18 = 100$. The baseline Qwen3-VL-4B-Instruct model similarly produced an incorrect response. However, the reinforcement learning fine-tuned variants accurately identified both blue and yellow bar categories and correctly executed the requisite mathematical operations (summation and subtraction), arriving at the correct answer of 0. These results demonstrate that RL fine-tuning yields substantial improvements across two critical dimensions: visual understanding and object classification accuracy, and multi-step mathematical reasoning capability.

5.2 Example 2: Multiple choice

Question: *Assuming the growth trend from 2010 to 2014 continues, what would be the estimated pension fund assets in 2015 (in N billion)?* (a) 4,800 (b) 5,200 (c) 5,500 (d) 6,000

Ground Truth: (b) 5,200

Model Responses: A SOTA MLLM employed only the endpoint values (2010 and 2014) to construct a Compound Annual Growth Rate (CAGR) model, arriving at option (c) 5,500. The baseline Qwen3-VL-4B-Instruct model failed to generate a valid response. In contrast, the RL fine-tuned variants exhibited sophisticated multi-step reasoning: extracting asset values across all intermediate years, identifying the non-linear nature of year-over-year growth patterns, selecting the most recent growth increment (2013-2014: 700) as the optimal predictor, computing the projected 2015 value (4500 + 700 = 5200), and correctly identifying answer option (b). This demonstrates that RL fine-tuning substantially enhances both visual data extraction accuracy and temporal reasoning capabilities.

5.3 Example 3: True/false question

Question: *The chart image in top left corner indicates a positive correlation between the account value of open cases and their creation date.*

Ground Truth: TRUE

Model Responses: A SOTA MLLM mischaracterized positive correlation as monotonic increase rather than recognizing the correct definition: a tendency for higher values of one variable to co-occur with higher values of another. The model failed to identify that larger bubbles (higher account values) systematically appear at later temporal positions. The baseline Qwen3-VL-4B-Instruct model demonstrated complete task failure. In contrast, the RL-enhanced models accurately performed spatial pattern recognition, identifying the clustering of lower-value accounts at earlier dates and higher-value accounts at later dates, correctly inferring the positive correlation and selecting "True."

6 Conclusion

We present Chart-RL, a novel framework that significantly advances VLMs' capabilities in chart understanding and reasoning through policy optimization techniques. Our comprehensive approach, integrating GRPO, DAPO, and GSPO with adaptive reward functions using an LLM judge, demonstrates substantial improvements in chart reasoning performance

while maintaining computational efficiency through PEFT via LoRA. The Qwen3-VL-4B-Instruct model enhanced with reinforcement learning achieved superior accuracy (0.634) compared to larger foundation models while reducing computational overhead by 50% and inference latency by 71%. The framework's ability to achieve competitive performance with significantly reduced parameters and training costs positions it as a practical approach for real-world deployment.

Future work will focus on enhancing the reinforcement learning framework through multi-stage reward refinement, where initial RL training with an LLM judge is followed by stages incorporating human feedback or ensemble reward models, reducing potential reward hacking and expanding the ceiling of achievable performance.

References

Alibaba. Qwen3-VL: Sharper vision, deeper thought, broader action. 2025. https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afeffrom=research.latest-advancements-list.

Anthropic. Claude 3.7 sonnet and claude code, 2025a. https://www.anthropic.com/news/claude-3-7-sonnet.

Anthropic. Introducing claude sonnet 4.5, 2025b. https://www.anthropic.com/news/claude-sonnet-4-5.

S. Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. *arXiv preprint arXiv:1801.08163*, 2018.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Akos Kadar, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. In *NeurIPS Visually Grounded Interaction and Language Workshop*, 2017.

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

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pp. 18893–18912, 2023.

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

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. ChartOCR: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

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

Ahmed Masry, M. S. Islam, M. Ahmed, A. Bajaj, F. Kabir, A. Kartha, Md. T. R. Laskar, M. Rahman, S. Rahman, M. Shahmohammad, M. Thakkar, Md. R. Parvez, E. Hoque, and S. Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*, 2025.

Minesh Mathew, Viraj Bagal, Rubén Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. InfographicVQA. *arXiv preprint arXiv:2104.12756*, 2021a.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200–2209, 2021b.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. *arXiv preprint arXiv:1909.00997*, 2019.

OpenAI. Introducing GPT-5, 2025. https://openai.com/index/introducing-gpt-5/.

A. Prabhakar, T. L. Griffiths, and R. T. McCoy. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. *arXiv preprint arXiv:2407.01687*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

W. Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, S. Malladi, Xindi Wu, H. Liu, A. Chevalier, S. Arora, and D. Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. *arXiv preprint arXiv:2406.18521*, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. *arXiv preprint arXiv:2506.14245*, 2025.

C. Wu et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

R. Xia, B. Zhang, H. Ye, X. Yan, Q. Liu, H. Zhou, Z. Chen, P. Ye, M. Dou, B. Shi, J. Yan, and Y. Qiao. ChartX & ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izzeddin Gur, Shixiang Shane Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Full responses from all models in the examples

A.1 Example 1: Open-ended question

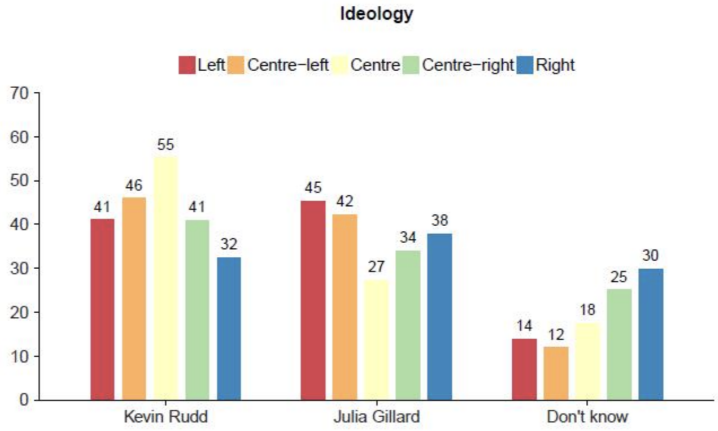


Figure 4: Chart for Example 1 (open-ended question).

Claude Sonnet 3.7

<think>

To find the difference between the sum of the blue bars and the sum of the yellow bars, I need to:

1. Identify which bars are blue and which are yellow in the chart
2. Sum up all the blue bars
3. Sum up all the yellow bars
4. Calculate the difference between these sums

Looking at the chart:

- The blue bars appear to be labeled as “Right” in the legend
- The yellow bars appear to be labeled as “Centre” in the legend

Blue bars (Right) values:

- Kevin Rudd: 32
- Julia Gillard: 38
- Don’t know: 30

Sum of blue bars = 32 + 38 + 30 = 100

Yellow bars (Centre) values:

- Kevin Rudd: 55
- Julia Gillard: 27
- Don’t know: 12

Sum of yellow bars = 55 + 27 + 12 = 94

Difference between sum of blue bars and sum of yellow bars = 100 - 94 = 6

</think>

<answer>

The sum of the blue bars (Right) is 32 + 38 + 30 = 100.

The sum of the yellow bars (Centre) is 55 + 27 + 12 = 94.

The difference between the sum of the blue bars and the sum of the yellow bars is 100 - 94 = 6.

</answer>

Qwen3-VL-4B-Instruct

<answer> 12 </answer>

Qwen3-VL-4B-Instruct-GRPO

<think>

To calculate the difference between the sum of the blue bars and the sum of the yellow bars, I need to:

1. Identify all the blue bars and sum their values.
2. Identify all the yellow bars and sum their values.
3. Subtract the sum of the yellow bars from the sum of the blue bars.

From the chart:

- Blue bars (representing "Right") have values: 32 (Kevin Rudd), 38 (Julia Gillard), and 30 (Don't know).
- Yellow bars (representing "Centre") have values: 55 (Kevin Rudd), 27 (Julia Gillard), and 18 (Don't know).

Sum of blue bars = 32 + 38 + 30 = 100.

Sum of yellow bars = 55 + 27 + 18 = 100.

Difference = Sum of blue bars - Sum of yellow bars = 100 - 100 = 0.

</think>

<answer> 0 </answer>

Qwen3-VL-4B-Instruct-DAPO

<think>

To calculate the difference between the sum of the blue bars and the sum of the yellow bars, I need to:

1. Identify all the blue bars and sum their values.
2. Identify all the yellow bars and sum their values.
3. Subtract the sum of the yellow bars from the sum of the blue bars.

From the chart:

- Blue bars (representing "Right") have values: 32 (Kevin Rudd), 38 (Julia Gillard), and 30 (Don't know).
- Yellow bars (representing "Centre") have values: 55 (Kevin Rudd), 27 (Julia Gillard), and 18 (Don't know).

Sum of blue bars = 32 + 38 + 30 = 100

Sum of yellow bars = 55 + 27 + 18 = 100

Difference = Sum of blue bars - Sum of yellow bars = 100 - 100 = 0

</think>

<answer> 0 </answer>

Qwen3-VL-4B-Instruct-GSPO

<think>

To calculate the difference between the sum of the blue bars and the sum of the yellow bars, we need to:

1. Identify the values of all blue bars (representing "Right").
2. Identify the values of all yellow bars (representing "Centre").
3. Sum the values of the blue bars.
4. Sum the values of the yellow bars.
5. Subtract the sum of the yellow bars from the sum of the blue bars.

From the chart:

Blue bars (Right):

- Kevin Rudd: 32
- Julia Gillard: 38
- Don't know: 30
- Sum of blue bars = 32 + 38 + 30 = 100

Yellow bars (Centre):

- Kevin Rudd: 55
- Julia Gillard: 27
- Don't know: 18
- Sum of yellow bars = 55 + 27 + 18 = 100

Difference = Sum of blue bars - Sum of yellow bars = 100 - 100 = 0

</think>

<answer> 0 </answer>

A.2 Example 2: Multiple choice

Claude Sonnet 3.7

<think>

To solve this problem, I need to:

1. Identify the pension fund assets values from 2010 to 2014
2. Calculate the growth trend during this period
3. Apply this trend to estimate the 2015 value
4. Compare the result with the given options

Looking at the chart, I can approximate the values:

- 2010: ∼1,800 N billion
- 2011: ∼2,400 N billion

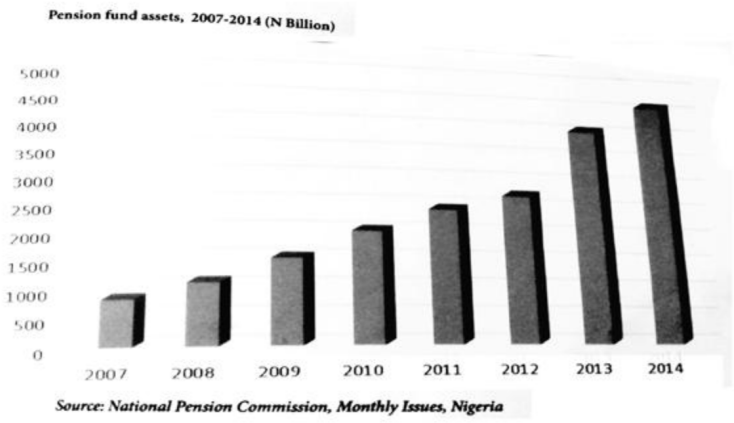


Figure 5: Chart for Example 2 (multiple choice question).

- 2012: ∼2,700 N billion
- 2013: ∼3,800 N billion
- 2014: ∼4,500 N billion

I can calculate the year-over-year growth rates:

- 2010 to 2011: (2400-1800)/1800 = 33%
- 2011 to 2012: (2700-2400)/2400 = 12.5%
- 2012 to 2013: (3800-2700)/2700 = 41%
- 2013 to 2014: (4500-3800)/3800 = 18%

I can calculate the compound annual growth rate (CAGR) from 2010 to 2014:

CAGR = $(4500/1800)^{(1/4)}$ - 1 = $(2.5)^{0.25}$ - 1 = 0.26 or 26%

Using this growth rate to estimate 2015:

2015 estimate = 4500 $\times$ 1.26 = 5670 N billion

This is closest to option c) 5,500 N billion.

</think>

<answer>

Based on the chart showing pension fund assets from 2007-2014, I need to estimate the 2015 value by analyzing the growth trend from 2010-2014.

Looking at the approximate values from the chart:

- 2010: ∼1,800 N billion
- 2011: ∼2,400 N billion
- 2012: ∼2,700 N billion
- 2013: ∼3,800 N billion
- 2014: ∼4,500 N billion

To find the overall growth trend, I'll calculate the compound annual growth rate (CAGR) from 2010 to 2014:

CAGR = $(4500/1800)^{(1/4)}$ - 1 = $(2.5)^{0.25}$ - 1 = 0.26 or 26%

Applying this growth rate to estimate 2015:

2015 estimate = 4500 $\times$ 1.26 = 5670 N billion

This is closest to option c) 5,500 N billion. Therefore, the answer is c) 5,500 N billion.

</answer>

Qwen3-VL-4B-Instruct

<answer> uncertain </answer>

Qwen3-VL-4B-Instruct-GRPO

<think>

To estimate the pension fund assets in 2015 assuming the growth trend from 2010 to 2014 continues, we need to analyze the growth pattern in the bar chart.

Step 1: Observe the growth from 2010 to 2014.

- 2010: ∼2,200 N billion
- 2011: ∼2,500 N billion
- 2012: ∼2,800 N billion
- 2013: ∼3,800 N billion
- 2014: ∼4,500 N billion

Step 2: Calculate the growth from 2013 to 2014.

- Growth = 4,500 - 3,800 = 700 N billion

Step 3: Assume the growth continues at a similar rate. The growth from 2013 to 2014 is 700 N billion. If we assume the same growth rate, then 2015 would be approximately 4,500 + 700 = 5,200 N billion.

Step 4: Compare with the options.

- a) 4,800 — too low
- b) 5,200 — matches our estimate
- c) 5,500 — higher than estimated
- d) 6,000 — too high

Therefore, the estimated pension fund assets in 2015 would be 5,200 N billion.

</think>

<answer> b) 5,200 </answer>

Qwen3-VL-4B-Instruct-DAPO

<think>

To estimate the pension fund assets in 2015 assuming the growth trend from 2010 to 2014 continues, we need to analyze the growth pattern in the bar chart.

Step 1: Observe the growth from 2010 to 2014.

- 2010: ∼2,200 N billion
- 2011: ∼2,500 N billion
- 2012: ∼2,800 N billion
- 2013: ∼3,800 N billion
- 2014: ∼4,500 N billion

Step 2: Calculate the growth from 2013 to 2014.

- Growth = 4,500 - 3,800 = 700 N billion

Step 3: Assume the growth continues at a similar rate. If we assume the same growth rate, then 2015 would be approximately:

- 4,500 + 700 = 5,200 N billion

Step 4: Compare with the options.

- a) 4,800 — too low
- b) 5,200 — matches our estimate
- c) 5,500 — higher than estimated
- d) 6,000 — too high

Therefore, the estimated pension fund assets in 2015 would be 5,200 N billion.

</think>

<answer> b) 5,200 </answer>

Qwen3-VL-4B-Instruct-GSPO

<think>

To estimate the pension fund assets in 2015 assuming the growth trend from 2010 to 2014 continues, we need to analyze the growth pattern in the bar chart.

Step 1: Observe the values from 2010 to 2014.

- 2010: ∼2,200 N billion
- 2011: ∼2,500 N billion
- 2012: ∼2,800 N billion
- 2013: ∼3,800 N billion
- 2014: ∼4,500 N billion

Step 2: Calculate the growth from 2013 to 2014.

- Growth = 4,500 - 3,800 = 700 N billion

Step 3: Assume the growth continues at a similar rate. The growth from 2012 to 2013 was 1,000 N billion (3,800 - 2,800), and from 2011 to 2012 was 300 N billion (2,800 - 2,500). The growth is not linear, but the most recent growth (2013 to 2014) is 700 N billion.

Step 4: Estimate the growth from 2014 to 2015.

- If we assume the growth continues at the rate of the last year (700 N billion), then:
- 2015 ≈ 4,500 + 700 = 5,200 N billion

Alternatively, if we look at the overall trend from 2010 to 2014, the growth is accelerating. From 2013 to 2014, the growth was 700 N billion. If we assume a similar or slightly higher growth rate, 5,200 is a reasonable estimate.

Thus, the estimated pension fund assets in 2015 would be approximately 5,200 N billion.

</think>

<answer> b) 5,200 </answer>

A.3 Example 3: True/false question

Claude Sonnet 3.7

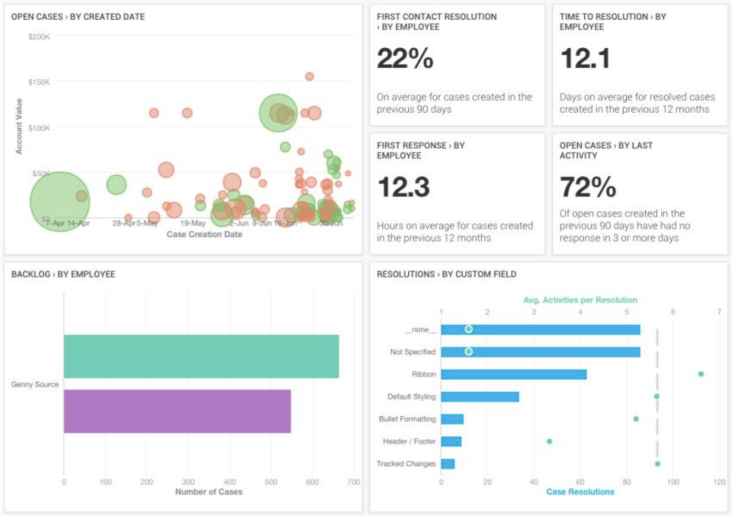


Figure 6: Chart for Example 3 (true/false question).

<think>

Let me analyze the chart in the top left corner titled "OPEN CASES: BY CREATED DATE" to determine if there is a positive correlation between account value and creation date.

The chart shows:

1. X-axis: Case Creation Date (ranging from approximately April to August)
2. Y-axis: Account Value (ranging from $0 to approximately $150K)
3. Bubbles of different sizes and colors (green and red) representing open cases

To determine if there's a positive correlation, I need to see if there's a trend where account values increase as the creation date becomes more recent.

Looking at the scatter plot:

- The data points are scattered across the entire date range
- There are high-value accounts (around $100K-$150K) distributed throughout the time period
- There are low-value accounts distributed throughout the time period
- There doesn't appear to be a clear upward trend from left to right

Based on visual inspection, I don't see a clear positive correlation between account value and creation date.

</think>

<answer>

The statement is incorrect. The chart in the top left corner titled "OPEN CASES: BY CREATED DATE" does not show a positive correlation between account value and creation date.

Looking at the scatter plot, the account values (Y-axis ranging from $0 to approximately $150K) appear to be distributed relatively randomly across the entire date range (X-axis from April to August). High-value accounts (around $100K-$150K) and low-value accounts are scattered throughout the time period without showing a clear upward trend from earlier to later dates.

If there were a positive correlation, we would expect to see account values consistently increasing as we move from left (earlier dates) to right (later dates) on the chart, which is not evident in this visualization.

</answer>

Qwen3-VL-4B-Instruct

<answer> uncertain </answer>

Qwen3-VL-4B-Instruct-GRPO

<think>

To evaluate whether the chart in the top-left corner indicates a positive correlation between account value and case creation date, I need to analyze the scatter plot.

Step 1: Understand the chart's axes.

- The x-axis represents "Case Creation Date" (with dates like 7-Apr, 28-Apr, etc.).
- The y-axis represents "Account Value" (in $K, ranging from $0K to $200K).

Step 2: Observe the pattern of the data points.

- The green circles (larger bubbles) are clustered toward the right side of the chart (later dates) and also at higher account values (e.g., $100K, $200K).
- The red circles (smaller bubbles) are clustered toward the left side (earlier dates) and at lower account values (e.g., $50K, $100K).

Step 3: Determine if there's a positive correlation.

- A positive correlation would mean that as the case creation date moves forward (to the right), account value also increases.
- In this chart, the larger bubbles (representing higher account values) are indeed located toward the right (later dates), while smaller bubbles (lower account values) are on the left (earlier dates).

Step 4: Conclusion.

- The visual pattern suggests that cases created later (later dates) tend to have higher account values, which indicates a positive correlation.

</think>

<answer> True </answer>

Qwen3-VL-4B-Instruct-DAPO

<think>

To evaluate whether the chart in the top-left corner indicates a positive correlation between account value and case creation date, I need to analyze the scatter plot.

Step 1: Understand the chart.

- The x-axis represents "Case Creation Date" (with dates like 7-Apr, 28-Apr, etc.).
- The y-axis represents "Account Value" (in $K, so $100K, $200K, etc.).
- The dots represent individual open cases, with size possibly indicating case volume or importance.

Step 2: Look for a trend.

- A positive correlation would mean that as the case creation date moves forward (to the right on the x-axis), the account value (y-axis) also increases.
- I need to visually inspect if dots tend to cluster in a way that higher account values are associated with later dates.

Step 3: Analyze the visual pattern.

- The chart shows a cluster of dots on the left (earlier dates) with relatively lower account values (e.g., around $50K).
- As we move to the right (later dates), there are dots with higher account values (e.g., up to $200K).
- This suggests that cases created later tend to have higher account values.

Step 4: Conclusion.

- The visual trend supports a positive correlation: later case creation dates are associated with higher account values.

</think>

<answer> True </answer>

Qwen3-VL-4B-Instruct-GSPO

<think>

To evaluate whether the chart in the top-left corner indicates a positive correlation between account value and case creation date, let's break this down step by step:

1. Understand the chart: The top-left chart is titled "OPEN CASES: BY CREATED DATE." It plots "Account Value" (y-axis) against "Case Creation Date" (x-axis). The data points are represented by circles, with size possibly indicating case volume or importance.

2. What does "positive correlation" mean? A positive correlation implies that as one variable increases, the other variable also tends to increase. In this case, if account value increases as case creation date moves forward (e.g., from early to late months), then there's a positive correlation.

3. Analyze the chart visually:

- The x-axis represents case creation dates, spanning from early April to late June.
- The y-axis represents account value, ranging from $0K to $200K.
- The data points (circles) are clustered in different regions. Notice that the largest circles (representing higher account values) appear toward the right side of the chart (later dates), while smaller circles (lower account values) are more clustered on the left (earlier dates).

4. Interpretation:

- The trend suggests that cases created later in the period (right side of the x-axis) tend to have higher account values (top of the y-axis).
- This indicates that as the case creation date progresses, account value tends to increase — which is consistent with a positive correlation.

5. Conclusion:

- The visual pattern supports the claim of a positive correlation between account value and case creation date.

</think>

<answer> True </answer>

B LLM Judge for Reward and Evaluation

Both the reward computation during training and the answer-accuracy metric during evaluation are operationalized by an LLM judge. We employ **gpt-oss-120B** as the judge model. Importantly, the judge comes from a different model family than any of the models under study in our experiments (Qwen and Claude), which mitigates the risk of a model favorably scoring outputs from its own family and provides an independent basis for comparison.

The judge receives the question, the ground-truth answer, and the model's predicted answer. It tolerates surface-level variations in phrasing, formatting, and numerical precision, and applies a 10% error tolerance to numeric predictions. As described in Section 3.3 (Reward function design), the accuracy and reasoning rewards are combined into the three-tier score (1 / 0.5 / 0); the prompt below is the answer-correctness component of that judgment, which is also used directly as the answer-accuracy metric during evaluation.

```
Task: Given a question-answer pair and the corresponding predicted answer,
evaluate the correctness of the predicted answer.
The output should be only 1 or 0, 1 means correct and 0 means incorrect.
Note that the ground truth answer can be a numeric value with or without
unit or any format.
Impose 10% error tolerance to the predicted answer.
```