Curriculum Learning based Hierarchical Scoring and Analysis Framework for Question Answering Task Evaluation

Qiong Wu[1], **Tan Yue**[1*], **Jianxin Liang**[1], **Zhen Li**[1], **Kai He**[3], **Shuai Zhao**[4], **Dongyan Zhao**[1,2*]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]National Engineering Research Center of New Electronic Publishing Technologies
[3]National University of Singapore [4]Nanyang Technological University
wu_qiong@stu.pku.edu.cn, {yuetan,liangjx,lizhen63,zhaody}@pku.edu.cn,
kai_he@nus.edu.sg, shuai.zhao@ntu.edu.sg

Abstract

The rapid progress of large language models (LLMs) has increased the demand for efficient and reliable evaluation of question answering (QA) systems. Existing evaluation methods either rely on rule-based matching with shallow semantic understanding or adopt LLM-as-a-Judge approaches that incur high cost and latency while offering limited error interpretability. Accordingly, we propose HiEval, a curriculum learning based hierarchical framework for QA task evaluation that supports both quick scoring and fine-grained error analysis. HiEval contains a quick scoring model (HiEval-QS) that predicts three-level correctness labels, and an error analysis model (HiEval-EA) that identifies incorrect responses into five error types. HiEval incorporates a class-balanced focal loss to handle label imbalance, experience replay to prevent forgetting, and contrastive unlikelihood optimization to improve error discrimination. We also construct two large-scale human-annotated evaluation datasets collected from 50 QA-related datasets, covering 8 task types and release two challenging benchmarks. Extensive experiments show that HiEval achieves state-of-the-art performance on both quick scoring and error analysis tasks, outperforming all baseline methods, including GPT-5, while being approximately 25× faster.[1]

1 Introduction

With the rapid development of large language models (LLMs) (Zhao et al., 2025a; Yue et al., 2026a), their performance has improved across a wide range of chat-based tasks (Ouyang et al., 2022; Yue et al., 2025b), increasing the need for efficient, reliable, and fine-grained question answering (QA) evaluation of generated responses (Kamalloo et al., 2023). An effective QA evaluation model that supports both accurate scoring and error analysis is essential for measuring model quality and guiding iterative improvement (Madaan et al., 2023). Existing evaluation methods can be categorized into two main types: rule-based methods (Papineni et al., 2002; Li et al., 2024b) are fast and stable but rely mainly on surface-level features, limiting their ability to assess complex reasoning and open-ended generation. In contrast, LLM-as-a-Judge approaches use LLMs to approximate human judgments, including prompt-based (Fu et al., 2023; Zhang et al., 2025), tuning-based (Chen et al., 2023a; Kim et al., 2024b), and post-processing methods (Xia et al., 2024; Yang et al., 2024).

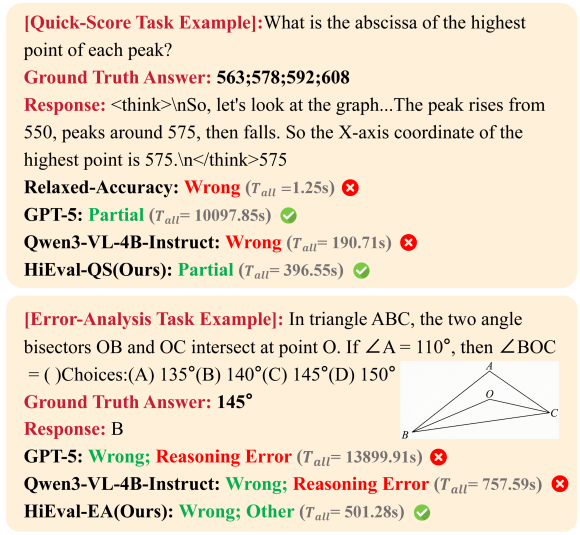


Figure 1: Comparison of complex QA evaluation. T_{all} is the total time required to complete all sample tests.

However, existing methods face several challenges in QA evaluation. First, rule-based matching often underestimates model performance due to limited semantic understanding (Wang et al., 2023). Second, most LLM-as-a-Judge approaches depend on commercial APIs (e.g., GPT-4o), which provide strong accuracy but incur high cost and latency, limiting large-scale use (Huang et al., 2025). While open-source LLMs require substantial computa-

* Corresponding author.
[1]https://github.com/wq-pku/HiEval

Findings of the Association for Computational Linguistics: ACL 2026, pages 6672–6699
July 2-7, 2026 ©2026 Association for Computational Linguistics

tion (Zhao et al., 2025b; Yue et al., 2026b; Zhao et al., 2026; Yang et al., 2026; Yue et al., 2026c), smaller models are more efficient but prone to hallucinations and misjudgments (Kalai et al., 2025). In addition, most methods lack fine-grained error analysis, focusing mainly on binary *correct/incorrect* decisions without identifying error types, which limits developers' ability to diagnose and improve models. Figure 1 shows the comparison of these methods in complex QA evaluation.

Accordingly, we propose HiEval, a curriculum learning based hierarchical framework for scoring and error analysis. The HiEval framework contains HiEval-QS model for quick scoring and HiEval-EA model for error analysis. These two models are trained in a progressive curriculum learning manner, enabling efficient large-scale evaluation while supporting fine-grained error analysis. In the first stage, to address label imbalance, we introduce a square-root–smoothed class-balanced focal loss to train the HiEval-QS, which predicts three-level scores (*Correct/ Partial/ Wrong*) given the question, ground-truth answer, and model response. In the second stage, we fine-tune HiEval-QS to build the HiEval-EA model. To preserve scoring ability while learning error discrimination, we adopt experience replay and apply contrastive unlikelihood optimization to sharpen decision boundaries and improve robustness. The final model supports both efficient scoring and classification of five error types: *Reasoning Error*, *Image Misunderstanding* (Sun et al., 2024), *Overthinking* (Fan et al., 2025), *Unanswerable* (Asai and Choi, 2021), and *Other*. Extensive experimental results demonstrate that HiEval-QS achieves 93.8% accuracy on the quick scoring task, outperforming all baselines and even the strongest GPT-5 (93%), while HiEval-EA reaches 85.5% accuracy on error analysis, substantially exceeding the best-performing GPT-5 (57.8%). Moreover, HiEval runs about 25× faster than GPT-5 with zero API cost.

Our contributions are summarized as follows: 1) We propose a hierarchical scoring and analysis framework that contains two-stage evaluation models, balancing efficient scoring with fine-grained error analysis for comprehensive assessment. 2) We construct a human-annotated high-quality evaluation dataset by sampling 50 multimodal QA datasets across eight task types, including 25,996 quick scoring instances and 10,545 error analysis instances, and also release two challenging benchmarks for quick scoring and error analysis. 3) Extensive experiments show that our models achieve state-of-the-art performance on all tasks, even surpassing GPT-5, while delivering high accuracy with zero cost and fast inference.

2 Related Works

Rule-based Methods Traditional rule-based methods mainly rely on surface-level lexical overlap, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), yet fail to capture the underlying differences. QA tasks commonly employ Exact Match (EM) and F1 scores. However, these rule-based methods are too strict and struggle to capture the semantic equivalence between complex responses and standard answers. They often underestimate the performance of the model. Although researchers have proposed "relaxed" matching strategies, such as BEM (Bulian et al., 2022) and PEDANTS (Li et al., 2024b), these methods still have poor flexibility and are difficult to handle the diversity of response forms.

LLM-as-a-Judge Methods LLM-as-a-Judge has become the mainstream paradigm for replacing manual assessment. According to the review by Li et al. (2024a), the single-LLM evaluation methods in this paradigm can be classified into three strategies: Prompt-based, Tuning-based and Post-processing methods. Prompt-based methods leverage In-Context Learning and Chain-of-Thought for training-free evaluation (Fu et al., 2023; Lin and Chen, 2023; Liu et al., 2023; Zhang et al., 2025), while recent works enhance flexibility by incorporating diverse evaluative roles and principles (Dong et al., 2024) or employing multi-turn interactions (Xu et al., 2024). Tuning-based approaches adapt LLMs via supervised fine-tuning (Chen et al., 2023a; Deshwal and Chawla, 2024) or preference alignment techniques like DPO, exemplified by Meta-Rewarding (Wu et al., 2024), JudgeLM (Zhu et al., 2023), and the PROMETHEUS series (Kim et al., 2024a,b). Post-processing methods refine outputs through probability calibration (Xia et al., 2024; Yang et al., 2024), multi-round aggregation (Sottana et al., 2023), or token-level score transformations (Ren et al., 2023) to ensure accuracy and reliability. LLM-as-a-Judge has been widely applied in various fields, such as general NLP tasks (Xiong et al., 2024), medicine (Xie et al., 2024), law (Yue et al., 2023), finance (Xie et al., 2023), education (Chiang et al., 2024), RAG system evaluation (Saad-Falcon et al., 2023), etc.

However, closed-source models (such as GPT-4) have slow evaluation speeds and high costs, while deploying large-scale open-source models requires more computing resources. Small-scale models are lightweight but difficult to ensure the accuracy of the evaluation. In addition, there are shortcomings such as hallucinations and a lack of the ability to explain errors (Li et al., 2024a). More details are in the Appendix A.

3 Methodology

This study proposes a hierarchical evaluation framework (HiEval) with two models for different use cases: a quick scoring model (HiEval-QS) for rapid assessment and an error analysis model (HiEval-EA) for scoring with detailed error analysis. Inspired by the human learning process of progressing from shallow to deep understanding (Bengio et al., 2009), we design a curriculum learning based training method. It adopts a two-stage training strategy: the first stage trains a foundational quick scoring model, while the second stage trains the error analysis model on this foundation. HiEval enables efficient scoring while providing reliable identification of major error types.

3.1 Quick Scoring Model

Quick Scoring Task Definition The model's input comprises the question, the ground truth answers, the model's response to be evaluated and the predefined scoring instruction. The output is one of the following: {*Correct*, *Partial*, *Wrong*}. Let Q denote the question, and A denote the set of ground truth answers for that question. Let R denote the set of answers within the model's response. We define the scoring logic as follows:

$$Score = \begin{cases} \textit{Correct}, & \text{if } R = A \\ \textit{Partial}, & \text{if } R \subset A \wedge R \neq \emptyset \\ \textit{Wrong}, & \text{if } R \cap A^c \neq \emptyset \vee R = \emptyset \end{cases} \quad (1)$$

Specifically, the score *Partial* is activated only when the response set R is a non-empty proper subset of the ground truth answer set A (denoted as $R \subset A$). This indicates that all elements within the response are correct, but the response fails to cover the entire solution space of the problem. For any case containing erroneous elements ($R \not\subseteq A$), our model classifies it as *Wrong*.

Optimization Objective In the quick scoring task, training data often suffers from sample imbalance, for example, there are an abundance of "Correct" and "Wrong" samples while partially correct ("Partial") samples are scarce. The traditional cross-entropy loss function tends to bias models toward classes with larger sample sizes. To address this issue, we propose a Class-Balanced Focal Loss with Square Root Smoothing.

To mitigate class imbalance, we first compute class weights based on the true label distribution in the training set. Let N denote the total number of samples, and N_c denote the number of samples in class $c \in C$ (where $C = \{\textit{Correct}, \textit{Partial}, \textit{Wrong}\}$). To prevent weight values from exploding due to extremely sparse samples in certain classes, we employ a square root smoothing strategy to compute the inverse class frequency: $w_c = \sqrt{\frac{N}{N_c}}$. Subsequently, we normalize the weights and introduce a truncation threshold T_{max} for training stability, yielding the final category weights α_c:

$$\begin{aligned} \alpha_c &= \min\left(\frac{w_c}{\bar{w}}, T_{\text{max}}\right), \\ \text{where} \quad \bar{w} &= \frac{1}{|C|}\sum_{j \in C} w_j \end{aligned} \quad (2)$$

This smoothing and clipping mechanism ensures the model focuses on minority classes ("Partial") without causing training divergence due to excessively large gradients from individual extreme samples.We incorporate the above weight α_c into the focal loss. For each generated scoring token, our final loss function $\mathcal{L}_{\text{Stage1}}$ is defined as:

$$\mathcal{L}_{\text{Stage1}} = -\alpha_c(1 - p_t)^\gamma \log(p_t) \quad (3)$$

where p_t is the model's predicted probability for the target ground truth label, α_c is the balancing weight for the current target category, and γ is the focal parameter. Through this design, during backpropagation, the loss weight for samples belonging to the majority class with high prediction confidence is significantly attenuated by $(1-p_t)^\gamma$. Conversely, for challenging samples from the minority class (weighted by α_c) with inaccurate predictions, their gradient contribution is preserved or even amplified.

3.2 Error Analysis Model

After the model acquires basic scoring capabilities, the second phase aims to extend its capabilities from "coarse-grained classification" to "fine-grained analysis". Specifically, for samples classi-

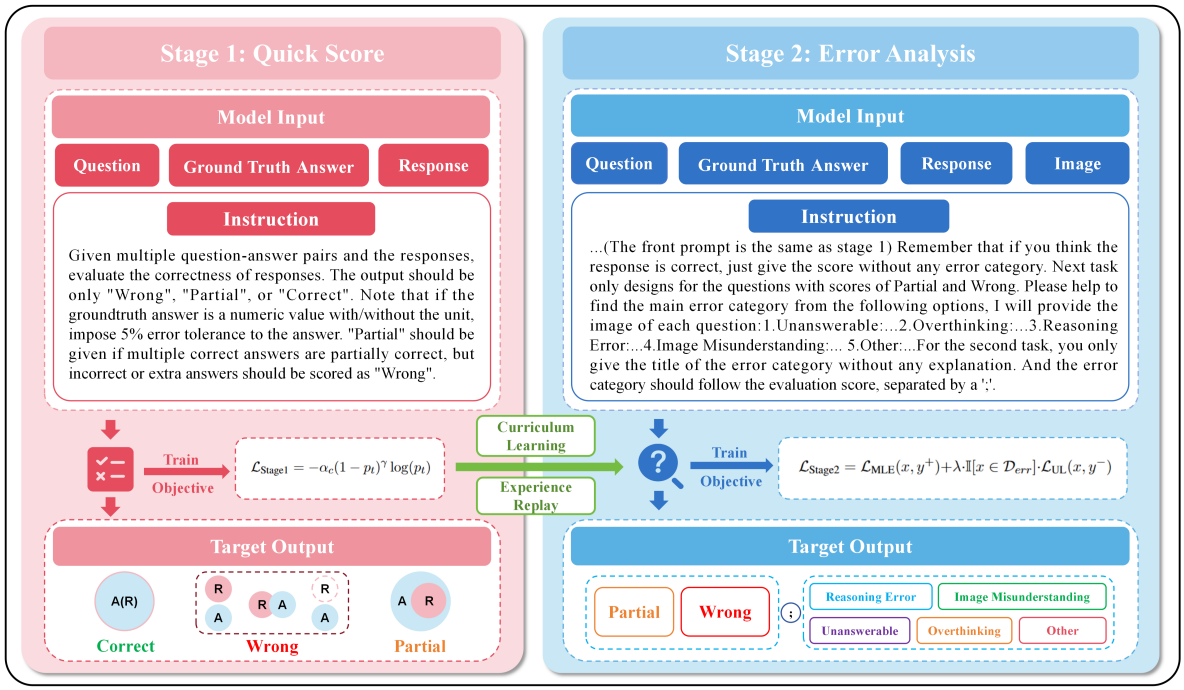


Figure 2: Training Framework. The left is the training process for the first stage of quick scoring. In the target output, A denotes the set of ground truth answers, and R denotes the set of answers within the model's response. The right is the training process for the second stage of error analysis. Details of instructions are in Appendix B.

fied as *Wrong* or *Partial*, the model must accurately identify the specific error type.

Error Analysis Task Definition Unlike the quick scoring task, the task in the second stage is defined as a multimodal analysis process. Formally, we define the input to this task as a quintuple $X = (I, Q, A, R, \mathcal{I}_{diag})$, specifically comprising: I means "Input image", Q means "Question", A means "Ground truth answer", R means "Model response to be evaluated", $\mathcal{I}_{diag}$ means Analysis Instruction, used to guide the model to perform scoring and error analysis tasks.

The model's objective is to learn the mapping function $\mathcal{F} : (I, Q, A, R, \mathcal{I}_{diag}) \rightarrow Y_{out}$. The output sequence Y_{out} comprises a scoring result $s \in \{Correct, Partial, Wrong\}$ and (optionally) an error analysis category $c \in \mathcal{C}$. The target output format is defined as:

$$Y_{out} = \begin{cases} s, & \text{if } s = Correct \\ s \oplus ";" \oplus c, & \text{if } s \in \{Wrong, Partial\} \end{cases} \quad (4)$$

where $\mathcal{C}$ denotes the error classification set defined in Section 4 (e.g. Reasoning Error), and $\oplus$ represents text concatenation. This definition clearly states that the model must generate analysis results based on the differences between A and R, guided by $\mathcal{I}_{diag}$ and incorporating I and Q.

Experience Replay Strategy When models transition from simple scoring tasks (Stage 1) to complex analysis tasks (Stage 2), they face the risk of catastrophic forgetting—where models may learn to analyze errors while losing foundational scoring accuracy. To address this, we introduce an experience replay mechanism. We construct a mixed dataset $\mathcal{D}_{mix}$ comprising data for task 2 ($\mathcal{D}_{task2}$) and data for task 1 ($\mathcal{D}_{task1}$) sampled proportionally at ρ:

$$\mathcal{D}_{mix} = \mathcal{D}_{task2} \cup \text{Sample}(\mathcal{D}_{task1}, \rho) \quad (5)$$

This ensures that the model, while learning new knowledge, constantly reviews old knowledge, maintaining the robustness of its scoring ability.

Contrastive Unlikelihood Training Standard supervised fine-tuning (SFT) optimizes models by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{T_{y^+}} \sum_{t=1}^{T_{y^+}} \log P(y_t^+ | x, y_{<t}^+) \quad (6)$$

The essence of this objective function is to teach the model "What is right". However, in multimodal analysis tasks, the model is confronted with the fine-grained classification problem. Training solely using MLE can easily lead to the model increasing the probability of the true label y^+ without

sufficiently suppressing the probabilities of other competing error categories, resulting in a blurred decision boundary.

To address this problem, we introduce a contrastive unlikelihood mechanism to explicitly teach the model "what is wrong". For each sample x containing an erroneous analysis, its true label is y^+ (which includes the error category c_{true}). Leveraging label mutual exclusivity, we sample a hard negative class $c_{neg} \in \mathcal{C} \setminus \{c_{true}\}$ to construct the negative sequence y^-.

We define the Normalized Sequence Probability, aiming to minimize the generation likelihood of the negative sequence:

$$P_{\text{norm}}(y^-|x) = \exp\left(\frac{1}{T_{y^-}} \sum_{t=1}^{T_{y^-}} \log P_\theta(y_t^-|x, y_{<t}^-)\right) \quad (7)$$

The corresponding non-likelihood loss function is defined as:

$$\mathcal{L}_{\text{UL}}(x, y^-) = -\log\left(1 - P_{\text{norm}}(y^-|x)\right) \quad (8)$$

Joint Training Objective The final optimization objective in Stage 2 combines universal generative capability (for all data) with discriminative constraints (for error-analysis data):

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{MLE}}(x, y^+) + \lambda \cdot \mathbb{I}[x \in \mathcal{D}_{err}] \cdot \mathcal{L}_{\text{UL}}(x, y^-) \quad (9)$$

where λ is the balancing coefficient. $\mathcal{L}_{\text{MLE}}$ ensures that the model can fluently generate scores and correct analysis, while $\mathcal{L}_{\text{UL}}$ serves as a regularization term specifically designed to penalize hallucinations produced by the model. $\mathcal{D}_{err}$ is the set of incorrect samples in error analysis task.

4 Dataset

4.1 Data Source and Diversity

We collect data from 50 mainstream open-source QA datasets (Details in Appendix C) spanning eight domains: geometric problem solving, mathematical word problems, figure-based QA, textbook and science QA, general VQA, document comprehension and OCR, biomedical QA, and general benchmarks. Responses are generated by both open-source and closed-source MLLMs, including GPT series (GPT-4o, GPT-4v, ChatGPT) (Achiam et al., 2023; Shen et al., 2023), Claude series (Anthropic, 2024), Gemini series (Team et al., 2024), Bard (Espejel et al., 2023), Qwen series (Qwen2.5-VL, Qwen2-VL) (Bai et al., 2025; Wang et al., 2024b), LLaVA series (Liu et al., 2024a), Mimo (Mimo-VL-7B-RL) (Xiaomi et al., 2025), Kimi series (Kimi-VL-A3B-Instruct/Thinking) (Team et al., 2025). The details of models are shown in Appendix D. This design captures diverse response styles, from instruction-following to chain-of-thought, and supports both multiple-choice and free-form questions. The final dataset contains 36,541 examples (shown in Table 1).

	Train	Val.	Test	Total
QS-Dataset	21,996	2,000	2,000	25,996
EA-Dataset	8,545	1,000	1,000	10,545

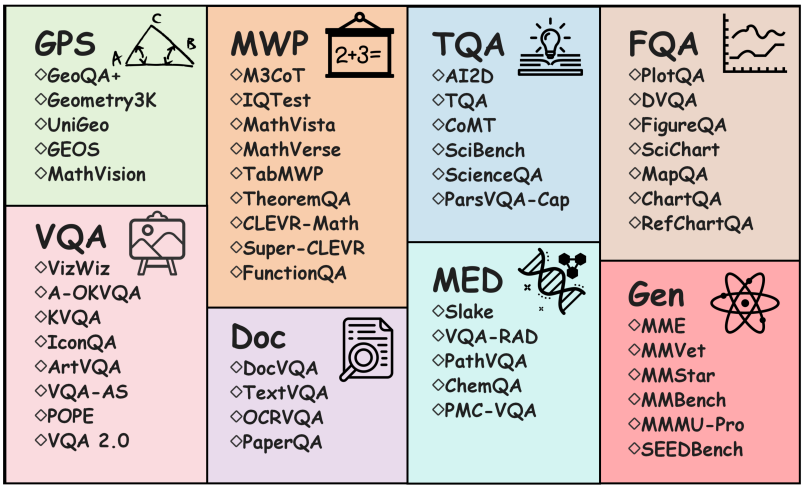
Table 1: Statistics of the constructed dataset for two stages.

4.2 Quick Scoring Dataset

The Quick Scoring Dataset (QS-Dataset) is designed to train a lightweight and efficient scoring model using a disagreement-based filtering pipeline. We first collect about 50,000 samples, each consisting of a question, ground-truth answer, model response, and an associated image (included to support Stage 2). 10 expert annotators label each sample with a ground-truth score following the criteria in Section 3.1. They use the data annotation system (in Appendix E) to label the data. In parallel, six representative models, covering both open-source and closed-source architectures, generate three-level predictions (*Correct, Partial, Wrong*) for the same data. We then apply rule-based filtering to remove trivial cases with very short responses, yielding about 40,000 samples. Next, an adversarial filtering step retains samples where at least three models disagree with the human label, targeting challenging cases. This process results in 25,996 high-quality samples, from which 2,000 are randomly selected for testing and the remainder are used for training with stratification.

4.3 Error Analysis Dataset

The goal of this stage is to enable the model to provide interpretable error analysis together with scoring. Building on Stage 1, we refine the annotation granularity by labeling error types. Based on prior studies of multimodal QA errors, we divide samples in the {*Wrong*, *Partial*} classes into five categories: **(1) Reasoning Error** (Sun et al., 2024), involving failures in logic, computation, or constraint understanding; **(2) Image Misunderstanding** (Sun et al., 2024), where image content



Statistic	Number
Total samples	36541
Domain classes	8
Sampled datasets	50
Models used to GR	12
Quick-Score Dataset	25996
Error-Analysis Dataset	10545
Avg question tokens	61.55
Avg answer tokens	1.43
Avg response tokens	262.17

"GR" = "Generate Responses"
"Avg" = "Average"

Figure 3: The datasets used for the quick scoring task and the error analysis task are composed of the QA data from these eight domains.

is misrecognized or hallucinated; **(3) Unanswerable** (Asai and Choi, 2021), where the model incorrectly judges that the question can not be answered; **(4) Overthinking** (Fan et al., 2025), mainly in reasoning-based models, characterized by redundant or circular reasoning that leads to truncation; and **(5) Other**, covering remaining cases such as overly short or poorly formatted responses. The annotation team assigns fine-grained labels to negative samples selected in Stage 1. To validate the reliability of annotations, we compute the Inter-Annotator Agreement (IAA) (in Appendix E). For training, we apply class-balanced sampling and obtain 9,545 evenly distributed instances. The test set contains 1,000 samples, with about 60% drawn from challenging cases in the Stage 1 (QS-Dataset) test set and 40% newly constructed to target specific error types, ensuring balanced coverage across all categories.

5 Experimental Setup

Baseline: Rule-based and relaxed matching approaches include Accuracy, Relaxed Accuracy, Bem, Pedant. Closed-source models include the GPT series (GPT-4o, GPT-5-mini, GPT-5) (Achiam et al., 2023; Wang et al., 2025) and the Gemini-2.0-Flash (Gemini-2.0) (Team et al., 2024). For open-source models, we select the Kimi-VL-A3B-Instruct (Kimi-VL-A3B) (Team et al., 2025), Qwen3-vl-4B-Instruct (Qwen3-vl-4B), Qwen3-4B-Instruct-2507 (Qwen3-4B) (Yang et al., 2025) and Mimo-RL (Xiaomi et al., 2025). The details of baseline models are shown in Appendix D.

Dataset and Evaluation: For Quick Scoring task, we conduct in-domain experiments using the test set from the constructed QS-Dataset and out-of-domain zero-shot experiments using the QAScore (Yue et al., 2025a) test set. For the error analysis task, since no open-source benchmark is currently available, we conduct experiments on the test set of the EA-Dataset we construct. We assess each model's performance on the benchmark using the metrics *Accuracy*, *Cost*, and *Running Time*.

Settings: We use Qwen3-VL-4B-Instruct model as the backbone. The Quick Scoring model is trained with AdamW (learning rate 5×10^{-5}) using LoRA with rank r=64, α=64, and dropout 0.05; the truncation threshold τ=5 and focus coefficient γ=1. The Error Analysis model is LoRA fine-tuned based on the Quick Scoring model with the same optimizer and LoRA settings, using a sampling ratio ρ=1 and balancing coefficient λ=0.3. Open-source models follow official implementations with default settings, while closed-source models use official APIs. Experiments run on six NVIDIA L40 (48GB) GPUs, and results are averaged over five runs. Test settings are in Appendix F.

6 Results

6.1 Quick Scoring Task

In-Domain Performance Results in Table 2 show that among rule-based methods, PEDANTS achieves the highest accuracy (61.25%), offering fast and low-cost evaluation but limited performance. Under the LLM-as-a-Judge paradigm, GPT-5 reaches 93% accuracy but incurs extremely long

	RT(s)	Cost	Acc
Rule-Based Methods			
Acc	1.31	Free	45.30
Relaxed-Acc	1.25	Free	48.10
BEM	269.27	Free	54.80
PEDANTS	16.89	Free	61.25
Closed-Source Models			
Gemini-2.0	2767.15	$0.1(Q)+$0.4(A)	71.10
GPT-4o	3048.24	$2.5(Q)+$10(A)	79.95
GPT-5-Mini	8930.65	$0.25(Q)+$2(A)	90.00
GPT-5	10097.85	$1.25(Q)+$10(A)	93.00
Open-Source Models			
Kimi-VL-A3B	701.58	Free	58.45
Mimo-VL-7B	39290.33	Free	62.90
Qwen3-4B	752.34	Free	70.95
Qwen3-VL-4B	190.71	Free	76.60
Ours			
HiEval-QS	396.55	Free	**93.80**

Table 2: Results of the evaluation on Quick Scoring Benchmark. RT=Running Time. Acc=Accuracy. $0.1(Q)+$0.4(A) means that input 1 million tokens cost $0.1 and output(response) 1 million tokens cost $0.4.

inference time (over 10,000 seconds) and high API costs. Among open-source models, Qwen3-VL-4B-Instruct is the fastest, yet reaches only 76.6% accuracy. In contrast, our HiEval-QS model (4B) achieves 93.8% accuracy with fast running time and free cost, outperforming all open-source baselines and even strong proprietary models (GPT-5, Gemini-2).

Out-of-Domain Performance We evaluate the out of domain performance of the HiEval-QS model on the QAScore test set. Results are shown in Table 3, our model achieves 91.7% accuracy and outperforms all rule-based and LLM baselines. Simultaneously, our model maintains free cost and efficient inference speed, demonstrating the HiEval-QS model's strong generalization and robustness.

6.2 Error Analysis Task

Main Performance As shown in Table 4, among closed-source models, GPT-5 gets the highest error analysis accuracy (57.80%) but suffers from extremely slow inference (13899s) and high cost. Among open-source models, Qwen3-VL-4B-Instruct performs best with 43.8% analysis accuracy, which is limited by its lower scoring accuracy (84.3%), as error analysis depends on reliable scoring. In contrast, our HiEval-EA model achieves 85.5% analysis accuracy with a scoring accuracy of 93.8%, substantially outperforming all baselines.

	RT(s)	Cost	Acc
Rule-Based Methods			
Acc	4.17	Free	47.50
Relaxed-Acc	4.36	Free	47.60
BEM	56.34	Free	63.50
PEDANTS	15.89	Free	70.20
Closed-Source Models			
Gemini-2.0	1172.85	$0.1(Q)+$0.4(A)	83.10
GPT-4o	1524.26	$2.5(Q)+$10(A)	83.70
GPT-4o-mini	1886.73	$0.15(Q)+$0.6(A)	84.50
Open-Source Models			
Kimi-VL-A3B	701.58	Free	58.45
Qwen3-4B	752.34	Free	70.95
Mimo-VL-7B	16567.37	Free	76.40
Qwen3-VL-4B	190.71	Free	83.50
Ours			
HiEval-QS	170.26	Free	**91.70**

Table 3: Results of the out-of-domain evaluation on QAScore dataset. RT=Running Time. Acc=Accuracy.

	RT(s)	Cost	A-Acc	S-Acc
Closed-Source Models				
Gemini-2.0	3367	$0.1(Q)+$0.4(A)	33.50	70.40
GPT-4o	3194	$2.5(Q)+$10(A)	33.00	81.30
GPT-5-Mini	16638	$0.25(Q)+$2(A)	45.20	85.30
GPT-5	13899	$1.25(Q)+$10(A)	57.80	90.40
Open-Source Models				
Kimi-VL-A3B	630	Free	8.00	47.90
Mimo-VL-7B	10234	Free	37.40	80.90
Qwen3-VL-4B	757	Free	43.80	84.30
Ours				
HiEval-EA	**501**	**Free**	**85.50**	**93.80**

Table 4: Results of the evaluation on Error Analysis Benchmark. RT=Running Time. S-Acc=Scoring Accuracy. A-Acc=Error Analysis Accuracy.

Breakdown analysis To examine fine-grained discrimination across error types, we evaluate accuracy on five error categories in the Error Analysis task. As shown in Table 5, the proposed model outperforms all baselines in every category. Notably, accuracy on *Overthinking* increases from 64.29% (GPT-5) to 99.05%, *Unanswerable* from 74.00% (GPT-5) to 98.00%, and *Other* from 29.70% (Kimi-VL-A3B-Instruct) to 84.85%. These results confirm the model's strong capability in fine-grained error type classification.

	RE	IM	Una.	Ove.	Other	Overall
Kimi-VL-A3B	1.15	10.19	1.00	0.00	29.70	8.00
GPT-4o	56.15	64.53	3.00	4.76	0.00	33.00
Gemini-2.0	55.77	21.89	17.00	54.29	0.61	33.50
Mimo-RL-7B	56.25	50.22	48.99	35.18	16.95	41.77
Qwen3-VL-4B	60.38	81.89	62.00	0.48	0.61	43.80
GPT-5-mini	55.00	72.83	47.00	32.38	0.61	45.20
GPT-5	73.08	67.55	74.00	64.29	0.00	57.80
Δ (vs Best)	↑0.77	0	↑24.00	↑34.76	↑55.15	↑27.70
HiEval-EA	**73.85**	**81.89**	**98.00**	**99.05**	**84.85**	**85.50**

Table 5: The results of the breakdown analysis for the error-analysis task. The metric is the analysis accuracy. RE=Reasoning Error. IM=Image Misunderstanding. Una.=Unanswerable. Ove.=Overthinking.

Variants	Quick Scoring	Error Analysis	
	Acc (%)	S-Acc (%)	A-Acc (%)
HiEval	**93.8**	**93.8**	**85.5**
w/o FL	92.5	-	-
w/o CU	-	91.0	84.0
w/o ER	-	55.2	41.3

Table 6: Ablation study on the effectiveness of different components in HiEval. "w/o" denotes removing a specific module or replacing it with a standard baseline. FL=Focal Loss. CU=Contrastive Unlikelihood. ER=Experience Replay.

6.3 Case Study

We compare HiEval-QS with GPT-5, Gemini-2.0-flash, and Qwen3-VL-4B-Instruct on the cases shown in Figure 4. In the first case, under the Quick Scoring task, both GPT-5 and HiEval-QS correctly judge the response as *Wrong*, whereas the other baseline models provide incorrect judgments. In the subsequent fine-grained error analysis, HiEval-EA categorizes the response as *Other*, since it neither provides a clear reasoning process nor descriptive information about the image, and does not exhibit overthinking or refusal behavior. In contrast, GPT-5 incorrectly classifies this case as a *Reasoning Error*. For the second case, both HiEval-QS and GPT-5 accurately classify the response as *Wrong*. However, only HiEval-EA correctly identifies the error type as *Overthinking*, as the response is clearly truncated. Other baseline models fail to make this distinction and provide incorrect classification.

Additionally, we compare HiEval-QS with baseline models on cases in the Appendix G.1 and compare HiEval-EA with baseline models in the Appendix G.2. These cases further demonstrate the superior quick scoring and error analysis capability of our models.

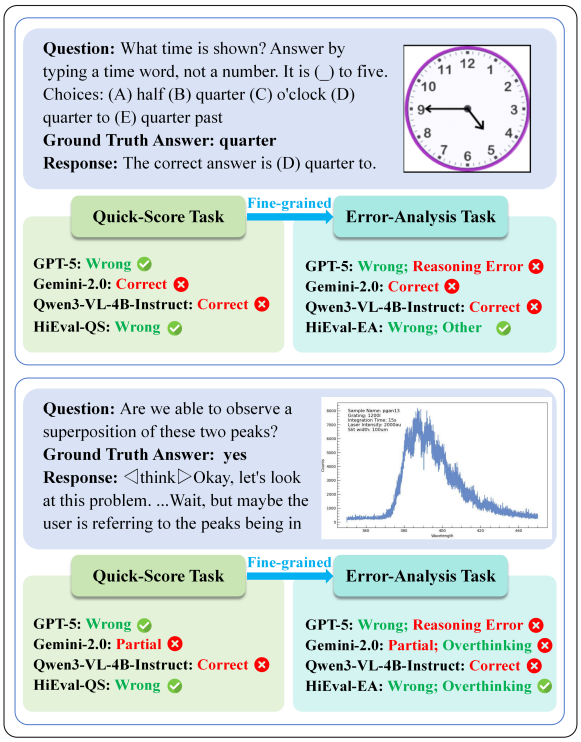


Figure 4: Case study for quick scoring and error analysis tasks.

6.4 Ablation Study

To assess the contribution of each component, we conduct an ablation study (Table 6). For the HiEval-QS model, replacing the square-root-smoothed category-balanced focal loss (w/o FL) with standard cross-entropy reduces scoring accuracy by about 1.3%, showing the benefit of the proposed loss under sample imbalance. For HiEval-EA model, removing the contrastive unlikelihood objective (w/o CU) reduces accuracy by 1.5%, indicating its importance for distinguishing error categories. Training without experience replay (w/o ER) causes a sharp drop in performance, with scoring accuracy decreasing to 55.2% and analysis accuracy to 41.3%, confirming that experience replay is essential to mitigate catastrophic forgetting and preserve reliable scoring and analysis.

7 Conclusion

In this work, we present HiEval, a hierarchical framework for QA evaluation that unifies quick scoring and fine-grained error analysis through curriculum learning. Extensive experiments show that both HiEval-QS and HiEval-EA achieve state-of-the-art performance, surpassing strong LLM-as-a-Judge baselines while operating with much lower cost and latency. In addition, we build a large-

scale, human-annotated evaluation dataset and two challenging benchmarks that support high-quality training and fair comparison of QA evaluators. The proposed HiEval provides an accurate, efficient, and interpretable solution for large-scale QA evaluation and analysis.

Limitations

The Quick Scoring model and the Error Analysis model are only applicable for accurately scoring and analyzing errors in QA tasks with clear answers. They are not suitable for open-ended generation tasks without standard answers. Additionally, the Error Analysis model can only provide the main types of errors for questions with a score of "wrong" or "partial". Currently, it can not give detailed explanations or indicate the specific locations of the errors, and further research is needed in this regard.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC, 62576016 and 62506014) and the China Postdoctoral Science Foundation under Grant Number 2025M781446.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433. IEEE.

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking qa: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1492–1504.

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and 1 others. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305.

Jie Cao and Jing Xiao. 2022. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th international conference on computational linguistics*, pages 1511–1520.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. 2023a. Adaptation with self-evaluation to improve selective prediction in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024a. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. 2023b. Theoremqa: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin.

2025. Comt: A novel benchmark for chain of multimodal thought on large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. *arXiv preprint arXiv:2407.05216*.

Mahesh Deshwal and Apoorva Chawla. 2024. Phudge: Phi-3 as scalable judge. *arXiv preprint arXiv:2405.08029*.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? *arXiv preprint arXiv:2406.11657*.

Jessica López Espejel, El Hassane Ettifouri, Mahaman Sanoussi Yahaya Alassan, El Mehdi Chouham, and Walid Dahhane. 2023. Gpt-3.5, gpt-4, or bard? evaluating llms reasoning ability in zero-shot setting and performance boosting through prompts. *Natural Language Processing Journal*, 5:100032.

Chenrui Fan, Ming Li, Lichao Sun, and Tianyi Zhou. 2025. Missing premise exacerbates overthinking: Are reasoning models losing critical thinking skill? *arXiv preprint arXiv:2504.06514*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, and 1 others. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. In *North American Chapter of the Association for Computational Linguistics*.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A dataset and baselines for visual question answering on art. In *European Conference on Computer Vision*, pages 92–108.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895.

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

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

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zhuowan Li, Xingrui Wong, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan Yuille. 2023c. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14963–14973. IEEE.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. 2024b. Pedants: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

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

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hannedige, A Ansari, Kai-Wei Lin, S andchang, and 1 others. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-chun Zhu. 2021a. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

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

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

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

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh Eetemadi. 2022. Parsvqa-caps: A benchmark for visual question answering and image captioning in persian. *people*, 101(404):1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jie Ren, Yao Zhao, Tu Vu, Peter J Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. In *Proceedings of the 40th International Conference on Machine Learning*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. 2015. Solving geometry problems: Combining text and diagram interpretation. In *EMNLP*, volume 1, page 3.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *AAAI Conference on Artificial Intelligence*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of gpt-4: reliably evaluating large language models on sequence to sequence tasks. *arXiv preprint arXiv:2310.13800*.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juanzi Li. 2024. Mm-math: Advancing multimodal math evaluation with process evaluation and fine-grained classification. In *Conference on Empirical Methods in Natural Language Processing*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Alexander Vogel, Omar Moured, Yufan Chen, Jiaming Zhang, and Rainer Stiefelhagen. 2025. Refchartqa: Grounding visual answer on chart images through instruction tuning. In *International Conference on Document Analysis and Recognition*, pages 523–537. Springer.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2023. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36:77013–77042.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

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

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024c. Scibench: evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 50622–50649.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

Tingyu Xia, Bowen Yu, Yuan Wu, Yi Chang, and Chang Zhou. 2024. Language models can evaluate themselves via probability discrepancy. *arXiv preprint arXiv:2405.10516*.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, and 1 others. 2025. Mimo:

Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. Doclens: Multi-aspect fine-grained evaluation for medical text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Shuying Xu, Junjie Hu, and Ming Jiang. 2024. Large language models are active critics in nlg evaluation. *arXiv preprint arXiv:2410.10724*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. 2024. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9073.

Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. 2026. Svagent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition 2026*. IEEE.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57730–57754.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and 1 others. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Tan Yue, Rui Mao, Xuzhao Shi, and Erik Cambria. 2026a. Interarm: Interpretable affective reasoning model for multimodal sarcasm detection. *IEEE Transactions on Affective Computing*.

Tan Yue, Rui Mao, Xuzhao Shi, Shuo Zhan, Zuhao Yang, and Dongyan Zhao. 2025a. Qaeval: Mixture of evaluators for question-answering task evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14717–14730.

Tan Yue, Xuzhao Shi, Rui Mao, Zilong Song, Zonghai Hu, and Dongyan Zhao. 2025b. Anafig: A human-aligned dataset for scientific figure analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12837–12843.

Tan Yue, Qiong Wu, and Dongyan Zhao. 2026b. Mars: Multimodal adaptive reasoning model for avoiding overthinking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34539–34547.

Tan Yue, Qiong Wu, and Dongyan Zhao. 2026c. We may not need much visual encoding of web data for question answering. In *Proceedings of the ACM Web Conference 2026*, pages 8437–8440.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Shuai Zhao, Qika Lin, Yanhao Jia, Xinyi Wu, Yuwen Li, and Luu Anh Tuan. 2026. Unifle: Uniform fusion of multiple lora experts for backdoor defense in large language models. *IEEE Transactions on Dependable and Secure Computing*.

Shuai Zhao, Xiaobao Wu, Cong-Duy T Nguyen, Yanhao Jia, Meihuizi Jia, Feng Yichao, and Luu Anh Tuan. 2025a. Unlearning backdoor attacks for llms with weak-to-strong knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4937–4952.

Shuai Zhao, Xinyi Wu, Shiqian Zhao, Xiaobao Wu, Zhongliang Guo, Yanhao Jia, and Anh Tuan Luu. 2025b. P2p: A poison-to-poison remedy for reliable backdoor defense in llms. *arXiv preprint arXiv:2510.04503*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Related Work

Existing evaluation methods are mainly divided into two categories: rule-based methods and LLM-as-a-Judge methods.

A.1 Rule-based Methods

Traditional rule-based methods primarily rely on surface-level lexical overlap, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), but they cannot capture deep nuances. The QA tasks often adopt exact matching (Exact Match, EM) and F1 score, but these rule-based methods are too strict and difficult to capture the semantic equivalence of complex answers and standard answers, often underestimating the performance of the model to be evaluated.

To address this problem, researchers propose "relaxed" matching strategies. Bulian et al. (2022) introduces the BEM framework, leveraging BERT to address the asymmetry flaw in F1 scores. Li et al. (2024b) combines expert rules from the Trivia community to develop PEDANTS, achieving efficient and interpretable relaxed matching. Although these methods have improved the flexibility of evaluation to some extent, they essentially still rely on shallow semantic features or predefined rules, lacking the ability to deeply understand complex semantics. They find it difficult to cope with the diversity of answer forms.

A.2 LLM-as-a-Judge Methods

With the exponential growth of large language models (LLMs), leveraging LLMs as evaluators (LLM-as-a-Judge) has emerged as the mainstream paradigm to replace human evaluation. According to Li et al. (2024a)'s review, this paradigm primarily encompasses three architectures: Single-LLM, Multi-LLM, and human-AI collaboration. This paper focuses on single-LLM evaluation, which can be categorized into three strategies:

Prompt-based methods primarily employ In-Context Learning and chain-of-thought(CoT) strategies. Works like GPTScore (Fu et al., 2023) and LLM-EVAL (Lin and Chen, 2023) enable zero-shot evaluation by guiding models through examples; G-EVAL (Liu et al., 2023) further employs CoT to decompose complex tasks into granular steps, enhancing reasoning capabilities. Additionally, recent research aims to enhance evaluation flexibility through incorporating diverse evaluative roles and principles (Dong et al., 2024) or incorporating multi-turn interactions (Xu et al., 2024).

Tuning-based approaches aim to adapt LLMs to specific evaluation tasks through training. Some works, such as ASPIRE (Chen et al., 2023a) and PHUDGE (Deshwal and Chawla, 2024), employ supervised fine-tuning with scores. Recent research increasingly favors alignment techniques like Direct Preference Optimization (DPO), exemplified by Meta-Rewarding (Wu et al., 2024), JudgeLM (Zhu et al., 2023), and the PROMETHEUS series (Kim et al., 2024a,b) supporting custom criteria.

Post-processing methods aim to refine model outputs for enhanced accuracy and reliability. Approaches like ProbDiff (Xia et al., 2024) and CRISPR (Yang et al., 2024) quantify differences and calibrate biases through mathematical derivation or Bayesian statistics; Sottana et al. (2023) employ multi-round evaluations to mitigate subjectivity, while Ren et al. (2023) transform open-ended generation tasks into token-level predictions for quality calibration.

LLM-as-a-Judge has been widely applied across multiple domains. In general NLP tasks, it is used to evaluate the quality of dialogues, summaries, and machine translations; in the multimodal domain, it supports comprehensive judgments on images and videos (Xiong et al., 2024); In specialized domains, it aids in evaluating medical text consistency (Xie et al., 2024), legal large-model performance (Yue et al., 2023), and financial risk (Xie et al., 2023); Furthermore, this paradigm plays a crucial role in educational assignment grading (Chiang et al., 2024) and RAG system evaluation (Saad-Falcon et al., 2023).

The approach still faces significant challenges: evaluating API-dependent commercial models (e.g., GPT-4) is slow and costly, while deploying large-scale open-source models demands substantial computational resources. Lightweight models often struggle to ensure evaluation accuracy. Furthermore, models' insufficient domain-specific knowledge and hallucination issues constrain their application in scenarios requiring high reliability. Additionally, these methods also lack the ability to explain errors (Li et al., 2024a).

B Instructions

For the quick scoring task and error analysis task, we design two versions of instructions. The first version of the instructions (Figure 5 removes the

Quick-Score-Instruction

Given multiple question-answer pairs and the responses, evaluate the correctness of responses. The output should be only "Wrong", "Partial", or "Correct". **Note that if the groundtruth answer is a numeric value with/without the unit, impose 5% error tolerance to the answer.** "Partial" should be given if multiple correct answers are partially correct, but incorrect or extra answers should be scored as "Wrong".
Now evaluate:
User: [Question]: {question} [Groundtruth answer]: {answer} [Answer]: {response}
Your score:

Error-Analysis-Instruction

You have two tasks. First, you are given multiple question-answer pairs and the corresponding predictions, evaluate the correctness of predictions. The output should be only "Wrong", "Partial", or "Correct". **Note that if the groundtruth answer is a numeric value with/without the unit, impose 5% error tolerance to the answer.** "Partial" should be given if multiple correct answers are partially correct, but incorrect or extra answers should be scored as "Wrong". Remember that if you think the prediction is correct, just give the evaluation score without any error category .You can say "Correct" and you are forbidden to say "Correct; (the type of error)". Next task only designs for the questions with scores of "Partial" and "Wrong", you must answer 'Partial; the type of error' or 'Wrong; the type of error'. Second, if you find the prediction is not correct, please help to identify the main error category from the following options, I will provide the image of each question:
1. Reasoning Error: Logical reasoning errors or irrelevant answers or calculate errors or misunderstandings of the textual conditions of the question.
2. Image Misunderstanding: Misinterpretation or overlooking of the directly presented information in the image.
3. Unanswerable: The model thinks that there are not sufficient necessary conditions for answering the question, or the given conditions are incorrect or incomplete, so it did not provide an answer.
4. Overthinking: The model stops answering abruptly within the limited word count or the model clearly has not completed its reasoning or thinking process. A common situation is the absence of an ending symbol.
5. Other: Any other errors that do not fall into the above categories. For instance, the model does not engage in thinking or reasoning; it merely provides very simple answers, thus being unable to identify the main cause of the error.
For the second task, you only give the title of the error category(Reasoning Error/Image Misunderstanding/Unanswerable/ Overthinking/Other) without any explanation. And the error category should follow the evaluation score, separated by a ';'.
Now evaluate:
User: [Question]: {question} [Groundtruth answer]: {answer} [Answer]: {response}
Your score:

Figure 5: Instructions in the quick scoring task and error analysis task.

red words) provides a detailed description of the scoring rules, the rules for error analysis, the definitions of each type of error, and specifies the format of the outputs. The second version of the Instruction adds a fault tolerance rate to the first version (Figure 5 maintains the red words). This allows for a 5% tolerance rate when answering questions where the standard answer is a numerical value.

C Data Source

GeoQA+ is a highly challenging multimodal benchmark in the field of geometric question answering. It is an improvement based on the original GeoQA and aims to address the issue of monotonous question types and low reasoning difficulty in previous datasets (Cao and Xiao, 2022).

Geometry3K is a large-scale benchmark for geometric problem-solving and symbolic reasoning, with its questions selected from multiple-choice questions in high school textbooks. Geometry3K covers various geometric shapes such as basic lines, triangles, and irregular quadrilaterals, as well as complex solution targets (Lu et al., 2021a).

UniGeo is a large-scale multimodal benchmark aimed at unifying geometric computation and proof tasks. It includes both computational and proof questions, covering five major geometric reasoning sub-tasks such as parallelism, triangles, congruence, similarity, and quadrilaterals (Chen et al.).

GeoS is a dataset used for automatically solving mathematical problems. It is a dataset of SAT plane geometry problems, where each problem is accompanied by an English description, along with diagrams and multiple-choice options. The questions and answers come from the official SAT exams and practice tests provided by the College

Board. We have labeled the logical forms of all the questions in the dataset with accuracy (Seo et al., 2015).

MathVision is a comprehensive benchmark designed to evaluate the visual mathematical reasoning capabilities of MLLMs comprehensively. This dataset collates high-quality problems sourced from real mathematics competitions, covering a wide range of mathematical disciplines such as analytic geometry, graph theory, topology, and various difficulty levels. The purpose of this dataset is to address the shortcomings of existing benchmarks in terms of diversity and difficulty, and through complex visual contexts, deeply assess the visual mathematical reasoning capabilities of MLLMs (Wang et al., 2024a).

VizWiz is a visual question answering (VQA) dataset derived from assistive applications for the visually impaired. It consists of real-life scene images captured by blind users along with corresponding spoken questions. The content of the dataset covers visual question-answering tasks in various daily life scenarios, mainly including text recognition (such as reading food labels or menus), color discrimination (such as distinguishing colors of clothes), object recognition (such as differentiating currency denominations), and environmental description (Bigham et al., 2010).

A-OKVQA is a visual question answering dataset that focuses on commonsense reasoning and world knowledge. It requires models to understand the content of the images and combine extensive external knowledge to answer challenging real-world scenarios. The dataset uniquely includes rationales (reasoning basis) for each question, aiming to assist models in learning reasoning logic and improving interpretability (Schwenk et al., 2022).

KVQA is a dataset specifically designed for the Knowledge-Driven Visual Question Answering task. It requires the model to answer questions by leveraging external world knowledge after identifying the named entities in the images (Shah et al., 2019).

IconQA aims to emphasize the significance of abstract chart understanding and comprehensive cognitive reasoning in real-world chart problems. This dataset consists of three sub-tasks: multiple image selection, multiple text selection, and filling in blanks. IconQA not only requires perceptual skills such as object recognition and text understanding, but also various cognitive reasoning skills such as geometric reasoning, commonsense reasoning, and arithmetic reasoning (Lu et al., 2021b).

ArtVQA is divided into two types of questions: visual and knowledge. Its purpose is to test the model's ability to understand the visual content and background knowledge of artistic works. The application fields of the dataset include art understanding, visual recognition, and natural language processing, aiming to solve visual question answering problems in the art field (Garcia et al., 2020).

VQA-AS is an abstract scene subset constructed from clip art in the VQA v1.0 dataset, aiming to eliminate the interference of lighting and object recognition in real images. It focuses on deeply evaluating the model's ability to understand spatial relationships and perform high-level logical reasoning while simplifying the difficulty of visual perception (Antol et al., 2015).

VQA 2.0 is designed to address the issue in the early datasets where the models tended to ignore the direct information in the images and guess the answers randomly. It achieves this by introducing the "complementary pairing" mechanism (where the same question corresponds to different images, leading to different answers), which helps to balance the data. This design forces the model to truly rely on the visual content for reasoning rather than answering based solely on language statistical patterns (Goyal et al., 2017).

POPE is a benchmark specifically designed for quantitatively evaluating object hallucinations in multimodal large models. It simplifies the assessment into a binary classification question of "Yes/No", directly exploring whether the model claims to have seen objects that do not exist in the image (Li et al., 2023b).

M3CoT is a multimodal Chain-of-Thought benchmark that covers three major fields: science, mathematics, and common sense. It is specifically designed to evaluate the capabilities of large models in long text and multi-step logical reasoning scenarios (Chen et al., 2024b).

MathVista is a comprehensive multimodal mathematical reasoning benchmark that covers various visual backgrounds ranging from elementary mathematics to advanced mathematics (such as geometric figures, function graphs, tables, puzzles, etc.). It integrates 28 existing multimodal datasets and adds three new sub-datasets (**IQTest, FunctionQA, and PaperQA**), and is currently one of the mainstream standards for evaluating

the mathematical capabilities of large multimodal models (Lu et al., 2024).

MathVerse is designed to reveal whether MLLMs truly "understand" mathematical charts, rather than merely relying on redundant information in the text prompts for reasoning. It creates a visually robust testing environment by removing the textual clues from the questions (only retaining the visual charts and the core questions), forcing the model to truly understand the content of the images in order to provide answers (Zhang et al., 2024).

TabMWP contains 38,431 open-domain level questions that require mathematical reasoning on text and table data. Each question in TabMWP is aligned with the table context, which is presented in the form of images, semi-structured text, and structured tables. There are two types of questions: free text and multiple choice. Each question is annotated with the gold solution to reveal the multi-step reasoning process (Lu et al., 2023).

TheoremQA consists of 800 high-quality university-level mathematical and physical questions, covering fields such as mathematics, physics, electronic engineering and finance, and each question is driven by a specific scientific theorem. It is designed to assess the ability of large models to apply professional theorems to solve complex scientific problems, rather than simple calculations (Chen et al., 2023b).

CLEVR-Math is a multimodal mathematical application dataset. It not only requires the model to identify the geometric objects in the images, but also demands the model to infer the state of the scene before or after the operations described in the text. This benchmark includes addition and subtraction, reverse counting, adversarial problems, and multi-hop reasoning tasks, aiming to evaluate the model's ability to "imagine" state changes by combining visual perception and logical reasoning (Lindström and Abraham, 2022).

Super-CLEVR is a benchmark designed to assess the robustness of visual question answering models. It decomposes the complex domain shifts into four independent factors that can be studied separately: visual complexity, question redundancy, concept distribution, and concept compositionality (Li et al., 2023c).

DocVQA is a large-scale dataset for visual question answering on document images. The images in the dataset cover various document types such as forms, tables, and handwritten texts, aiming to promote the model's ability to combine text content with document layout structure to answer natural language questions (Mathew et al., 2021).

TextVQA is a dataset for benchmarking visual reasoning based on the text in images. TextVQA requires models to read and reason about the text in the images to answer questions related to them (Singh et al., 2019).

OCR-VQA aims to promote research on answering visual questions by reading the text in images. Its questions cover information such as the title, author, type, year and version of books (Mishra et al., 2019).

AI2D is a dataset containing over 5,000 primary school science charts, with over 150,000 detailed annotations regarding the constituent elements of the charts and their syntactic relationships. The dataset also comes with over 15,000 corresponding multiple-choice questions, specifically designed to evaluate the model's ability in chart syntactic parsing and semantic reasoning (Kembhavi et al., 2016).

TQA is a multimodal reading comprehension dataset derived from the junior high school science curriculum (covering life science, earth science and physics). This dataset is designed to test the ability of MLLMs to reason in complex contexts that combine textual descriptions, charts and natural images, and requires the model to answer questions through cross-modal information integration rather than simple text retrieval (Kembhavi et al., 2017).

CoMT is a novel multimodal CoT benchmark, designed to evaluate the capabilities of MLLMs in complex visual operations and concise expression. The dataset covers four categories: visual creation, visual deletion, visual update, and visual selection. The dataset aims to address the issues of the absence of visual operations and ambiguous expression in traditional multimodal CoT benchmarks (Cheng et al., 2025).

SciBench is a novel benchmark for college-level scientific problems consisting of 695 problems sourced from textbooks. The benchmark is designed to evaluate the complex reasoning capabilities, strong domain knowledge, and advanced calculation skills of LLMs (Wang et al., 2024c).

ScienceQA is a multimodal dataset that covers three major fields: natural science, social science, and language science. The notable feature of this dataset is that it provides relevant background lectures and detailed explanations for most ques-

tions, aiming to support AI models in conducting multi-step reasoning and answering through the Chain of Thought (Lu et al., 2022).

ParsVQA-Caps is the first benchmark specifically designed for Persian visual question answering and image description tasks, aiming to fill the gap of non-English resources and address the widespread cultural bias towards Europe and America in existing datasets (Mobasher et al., 2022).

SLAKE is a medical visual question answering dataset, consisting of 642 medical images and 14,028 question-answer pairs. It covers various modalities such as CT, MRI and X-ray, as well as multiple human body parts including the head, neck and chest. It includes both healthy and unhealthy samples, involving a total of 12 diseases and 39 organs (Liu et al., 2021).

VQA-RAD is the first radiology visual question answering dataset manually constructed by clinical doctors, which includes 315 balanced image samples covering the head, chest, and abdomen from the MedPix database, as well as 3,515 pairs of questions and answers (Lau et al., 2018).

PathVQA is a dataset specifically designed for the visual question answering task in pathology. Its significance lies in that it provides practical question-and-answer cases for the automated analysis and understanding of pathological images, which is crucial for the development of AI-driven medical diagnostic systems (He et al., 2020).

ChemQA is a multimodal question-answering dataset focused on chemical reasoning, consisting of 5 tasks: counting the number of carbon and hydrogen atoms in organic molecules, calculating the molecular weight of organic molecules, converting from SMILES to IUPAC names, generating molecular titles and edits, and reverse synthetic planning.

PMC-VQA is a large-scale medical VQA dataset. The images and question-answer pairs cover various medical modalities, such as X-ray films, MRI, and CT, as well as various diseases. Its aim is to achieve medical visual understanding through the combination of visual information and language models (Zhang et al., 2023).

PlotQA not only covers various types of charts such as bar charts, line graphs and scatter plots, but also introduces a large number of complex reasoning questions whose answers exceed the fixed vocabulary list. This forces the model to handle real floating-point numbers and conduct deep visual and logical reasoning (Methani et al., 2020).

DVQA is a synthetic dataset composed of image-question pairs, covering three tasks: structure understanding, data retrieval, and reasoning. It is specifically designed to test the algorithm's ability to understand bar charts (Kafle et al., 2018).

FigureQA is a visual reasoning corpus consisting of over 1 million pairs of questions and answers based on more than 100,000 images. These images come from five types of synthetic scientific style graphics: line graphs, point-line graphs, vertical and horizontal bar graphs, and pie charts (Kahou et al., 2017).

SciChart is a multi-modal scientific chart question-answer dataset, which mainly consists of 5 types of questions: identifying the number of peaks, identifying the peak values, identifying the positions of peaks, determining the shape of the peaks, and calculating the half-width of the peaks.

MapQA is designed to test the different levels of understanding that a model can have of maps, ranging from simple recognition of map styles to complex problems that require reasoning about underlying data (Chang et al., 2022).

ChartQA is a benchmark for chart-based questioning, aiming to answer questions about charts through visual and logical reasoning. It includes a large number of complex reasoning problems that involve multiple logical and arithmetic operations, and often involve the visual features of the charts (Masry et al., 2022).

RefChartQA is a large-scale benchmark for visual grounding in chart-based question answering. It extends the ChartQA and TinyChart-PoT datasets by adding explicit bounding box annotations that link each answer to supporting visual elements in the chart (Vogel et al., 2025).

MME is a comprehensive evaluation benchmark for MLLMs, covering two major categories of tasks: perception and cognition. It includes tasks such as commonsense reasoning, numerical calculation, text translation, and code reasoning, and mainly adopts the binary questioning format of "yes-no" (Fu et al., 2024).

MMVet is a benchmark designed to evaluate the ability of large multi-modal models (LMMs) to handle complex integrated tasks. It defines and examines six core visual-language capabilities, including recognition, OCR, knowledge, language generation, spatial perception, and mathematics, as well as their derived 16 integrated capabilities (Yu et al., 2024).

MMStar is a multimodal benchmark. Its samples cover various tasks, such as multiple-choice, question answering, and visual question answering, and it supports both image and text modalities (Chen et al., 2024a).

MMBench is a system-constructed multimodal evaluation benchmark. Its data content covers two major categories: Perception and Reasoning. It is further divided into 20 fine-grained capability dimensions (such as object localization, social relationship reasoning, etc.), aiming to conduct a comprehensive and hierarchical evaluation of MLLMs (Liu et al., 2024b).

MMMU-Pro is an enhanced and selected version of the large-scale multi-disciplinary multi-modal understanding (MMMU) benchmark, aiming to provide a more rigorous and challenging evaluation standard by eliminating samples that can be answered solely based on text without the need to view the images. This dataset is specifically designed to examine the model's ability to conduct in-depth reasoning by truly integrating visual information in various interdisciplinary fields, in order to more accurately reflect the understanding level of MLLMs (Yue et al., 2024).

SEED-Bench is a large-scale multimodal benchmark test consisting of 19,000 manually verified multiple-choice questions. It covers 12 evaluation dimensions in image (spatial) and video (temporal) understanding. This dataset is constructed by combining the automatic generation of large models with an advanced pipeline of manual screening, aiming to objectively and comprehensively evaluate the generative understanding capabilities of MLLMs (Li et al., 2023a).

D Details of the models used to generate responses and serve as baselines

GPT-4V (Achiam et al., 2023) is a significant multimodal extension of the GPT-4 series, incorporating a visual encoder to process interleaved image and text inputs. It is widely used for tasks such as chart, document, and scene graph understanding, visual question answering (VQA), and visual reasoning.

GPT-4o (Achiam et al., 2023) is an end-to-end trained omni-modal model capable of native processing across text, audio, and vision. Designed for low-latency real-time interaction, it maintains GPT-4 level capabilities while offering faster speeds and more natural voice and visual interactions. Typical applications include real-time dialogue, image understanding, voice assistants, and multimodal agent interactions.

ChatGPT (Shen et al., 2023) serves as a conversational interface powered by the GPT series models, integrating capabilities such as web browsing, advanced data analysis, and image generation to function as a comprehensive AI assistant.

GPT-5 (Wang et al., 2025) is described as a significant leap forward compared to previous GPT series, covering capabilities in coding, mathematics, writing, and visual understanding. Officially emphasized as a "unified system," it can adaptively switch between "fast responses" and "longer thinking" processes to enhance the quality of solutions for complex problems.

GPT-5-mini (Wang et al., 2025) is a lightweight variant of the GPT-5 architecture, optimized for inference cost and high throughput while retaining a 400K token long context window for extensive document processing tasks.

LLaVA Series (Liu et al., 2024a) introduces visual instruction tuning techniques to connect a vision encoder with a large language model, demonstrating that a simple linear projection layer can effectively align visual features with the language space.

Claude Series (Anthropic, 2024) focuses on safety and controllability, with the Claude 3.5 Sonnet and Claude 4 models exhibiting exceptional performance in code generation, complex instruction following, and visual data extraction, supported by a context window exceeding 200K tokens.

Qwen2-VL (Wang et al., 2024b) features "any resolution perception," proposing mechanisms like dynamic resolution processing and Multimodal Rotary Positional Embeddings (M-RoPE) to unify image/video and text position alignment. It is commonly used for multimodal reasoning, document and chart understanding, and video comprehension tasks.

Qwen2.5-VL (Bai et al., 2025) builds upon the Qwen2 architecture with enhanced visual encoders, delivering improved grounding capabilities and robust performance across multi-language environments and dense text scenarios.

Qwen3-VL-4B-Instruct (Yang et al., 2025) is officially positioned as the next-generation upgrade to the Qwen visual language series, emphasizing stronger visual perception and reasoning, as well as longer context and enhanced agent interactions. The 4B-Instruct version is a lightweight, instruction-aligned model suitable for deployment in resource-constrained environments for multi-

modal evaluation and application prototyping.

Qwen3-4B-Instruct-2507 (Yang et al., 2025) serves as a specialized text-only instruction-following model optimized for speed and precision in multilingual dialogue. By removing the overhead of deep reasoning processes, it is well-suited for high-frequency interaction scenarios.

Gemini-2.0-Flash (Team et al., 2024) is a highly efficient multimodal model capable of processing native audio and video inputs. With a 1-million-token context window, it is specifically optimized for low-latency agentic applications.

Bard was an early conversational AI service powered by the LaMDA and subsequently PaLM architectures, laying the groundwork for the subsequent transition to the unified Gemini model family.

Mimo-VL-7B-RL (Xiaomi et al., 2025) consists of a "native resolution ViT encoder + MLP projection + MiMo-7B language model," emphasizing fine-grained visual details and efficient cross-modal alignment. The RL version (7B-RL) is described in technical reports as performing strongly on multiple multimodal/reasoning tasks, with a specific emphasis on GUI grounding capabilities.

Kimi-VL-A3B-Instruct (Team et al., 2025) is described as an efficient Mixture-of-Experts (MoE) visual language model, focusing on multimodal reasoning, long context, and agent capabilities. While maintaining a small active parameter size (A3B) on the language decoding side, the Instruct version demonstrates strong instruction-following abilities.

Kimi-VL-A3B-Thinking (Team et al., 2025) extends the base Kimi-VL architecture by incorporating internal reasoning chain mechanisms, allowing the model to perform "slow thinking" for complex geometric and logical visual problems before generating a response.

E Data Annotation

A total of 10 annotators are recruited for the data annotation task in this study. All the annotators are native Chinese speakers from top universities and have passed the College English Test Band 6 (CET-6). They possess excellent bilingual skills and have extensive professional experience in NLP and question answering (QA) fields. We have formulated detailed annotation rules and operation guidelines for each dataset. The average hourly wage of the annotators is 130 CNY, which is far above the local minimum wage standard. Additionally, the annotators need to take a 10-minute break every 30 minutes of work.

Category	**Cohen's κ**
Score Category	
Wrong	0.9397
Partial	0.8665
Correct	0.9520
Overall	**0.9424**
Error Category	
Image Misunderstanding	0.8083
Other	0.8910
Overthinking	0.9158
Reasoning Error	0.7121
Unanswerable	0.9661
Global Average	**0.8352**

Table 7: Inter-annotator agreement (Cohen's Kappa) for different scores and error categories. The results indicate a high level of consistency across both evaluation dimensions.

We have designed a data annotation system (Figure 6) to facilitate the annotation process for the annotators. In this system, the annotation progress of the Quick Scoring task and the Error Analysis task (displayed in the upper right corner), the ID of the question, the domain, the question, the image, the response, and the ground truth answer will be shown. When the annotators determine that a question is "Wrong or "Partial", the error type annotation will be triggered.

we have computed the Cohen's Kappa scores for our annotations. The results are in Table 7. For the QS-Dataset, the overall Cohen's Kappa is 0.9424, with all individual scores (Wrong, Partial, Correct) exceeding 0.86. This indicates an "almost perfect agreement" among our annotators regarding the grading of answers. For the EA-Dataset, the global average Kappa score is 0.8352. According to the widely accepted interpretation criteria by Landis and Koch (1977), a score between 0.61–0.80 represents "substantial agreement," and a score between 0.81–1.00 represents "almost perfect agreement." These metrics objectively demonstrate that, our expert annotators successfully achieved a highly reliable consensus, which validates the high quality of our ground-truth labels.

The datasets and codes of this project will be open-sourced under the MIT license agreement. All research data are collected from public channels, and we have ensured strict compliance with relevant data usage policies and privacy protection regulations. This data is intended for academic research, and the use has been explained in detail in

the instructions.

F Test Settings

F.1 Closed-Source Model

Closed-source models, including GPT series and Gemini series, do not require local computing resources. Instead, they integrate by calling remote API interfaces through the standard 'openai' client library in the Python environment. In terms of parameter configuration, the model maintains the zero-sample prompt strategy, uses the Instructions provided in the Appendix B, and outputs the results under the default inference settings.

F.2 Open-Source Model

Qwen3 series Qwen3-4B-Instruct-2507 and Qwen3-VL-4B-Instruct are deployed locally on NVIDIA GPUs based on the Hugging Face Transformers framework. The model weights are loaded in bfloat16 half-precision format; during inference, the maximum generation length (max_new_tokens) is set to 128 (Quick Scoring Task), 512 (Error Analysis Task), and other generation parameters remain at default values.

Kimi-VL-A3B-Instruct is deployed locally on an NVIDIA GPU based on the Hugging Face Transformers framework, enabling trust_remote_code=True; during the inference stage, the greedy decoding strategy (setting do_sample=False) is adopted to eliminate the randomness in the generation process (although we set temperature=0.7 and top_p=0.9, it should not have worked), and the maximum generation length (max_new_tokens) is limited to 128 (Quick Scoring Task), 512 (Error Analysis Task).

MiMo-VL-7B-RL is deployed on a single NVIDIA GPU based on the Hugging Face Transformers framework, with trust_remote_code=True enabled; the model weights are loaded in the bfloat16 half-precision format. During the inference stage, do_sample=False is explicitly set to adopt the greedy decoding strategy to ensure the determinacy of the output (ignoring the temperature parameter), and the maximum generation length (max_new_tokens) is set to 2048 for all tests.

F.3 HiEval

Our HiEval-QS is based on the Qwen3-VL-4B-Instruct base model. Using the peft library, we merge and load the pre-trained weights with the fine-tuned LoRA adapter weights, and uniformly adopt bfloat16 precision to adapt to the NVIDIA GPU environment. During the inference stage, we limit the maximum generation length (max_new_tokens) to 8, and keep the other parameters at their default values. The HiEval-EA model needs to load the corresponding weights first, set the maximum generation length to 128, and keep the other parameters at their default values.

G Case Study

G.1 Quick Scoring Task

We demonstrate the performance of the Quick Scoring model and some baseline methods on specific examples. In Figure 7, both of these examples are correct because the meanings of the response and the ground truth answer are consistent. HiEval-QS makes an accurate judgment, but the other baseline methods are not able to make correct judgments in all cases. In Figure 8, this is the response provided by the "thinking" version model and the response's content is only a subset of the ground truth answer, so it is partially correct. Our HiEval-QS correctly classifies it as "Partial", but the baseline methods misjudge it. In Figure 9, the response shows overthinking, which has led to it being truncated due to the token limit, so it is wrong, but most baseline methods(except for Relax-Accuracy) mistakenly consider it correct. These examples fully demonstrate that our HiEval-QS has excellent semantic understanding capabilities and can provide accurate scores.

G.2 Error Analysis Task

To more intuitively demonstrate the performance differences between our model and baseline models on the error-analysis task, we select five representative cases.

Specifically, in Figure 10, this response clearly shows a misunderstanding of the image. In the picture, there is only one peak, but it mistakenly believes there are three peaks. Our HiEval-EA accurately provides the score and determined the cause of the error as "Image Misunderstanding". However, in the baseline models, only MImo-RL makes the correct judgment, while the other models all fail.

In Figure 11, this response contains the error in reasoning. It is not necessarily true that any three randomly selected points out of the 18 points can form a triangle, and it also fails to calculate the

final numerical result. The HiEval-EA successfully provides the correct score and accurately classifies the error type. However, in the baseline model, only MIMO-RL is able to make the correct judgment, while the other models all make incorrect judgments.

In Figure 12, the response exhibits clear text truncation (abruptly ending), typically indicating the model becomes stuck in a reasoning loop leading to token exhaustion—a classic 'Overthinking' characteristic. The HiEval-EA successfully classifies this, whereas baseline models fail to detect it.

In Figure 13, the response explicitly states it can not answer due to insufficient information, constituting a standard 'Unanswerable' error. Our HiEval-EA accurately captures this refusal intent; however, in the baseline model, Mimo-RL does not follow the task instruction of providing scores first and then analyzing the reasons for the errors. Gemini-2.0-Flash misjudges the scores and the reasons for the errors. Other models fail to determine the reasons for the errors correctly.

In Figure 14, the response is incorrect and extremely brief (only outputting options), lacking specific reasoning processes or image descriptions. It also does not exhibit overthinking or unanswerable phenomena, thus should be classified as 'Other'. Our HiEval-EA can accurately determine the scores and the causes of errors, while the baseline models have all failed to correctly find the causes.

The above qualitative analysis fully demonstrates that, compared to the baseline model, our model can more precisely analyze the type of error.

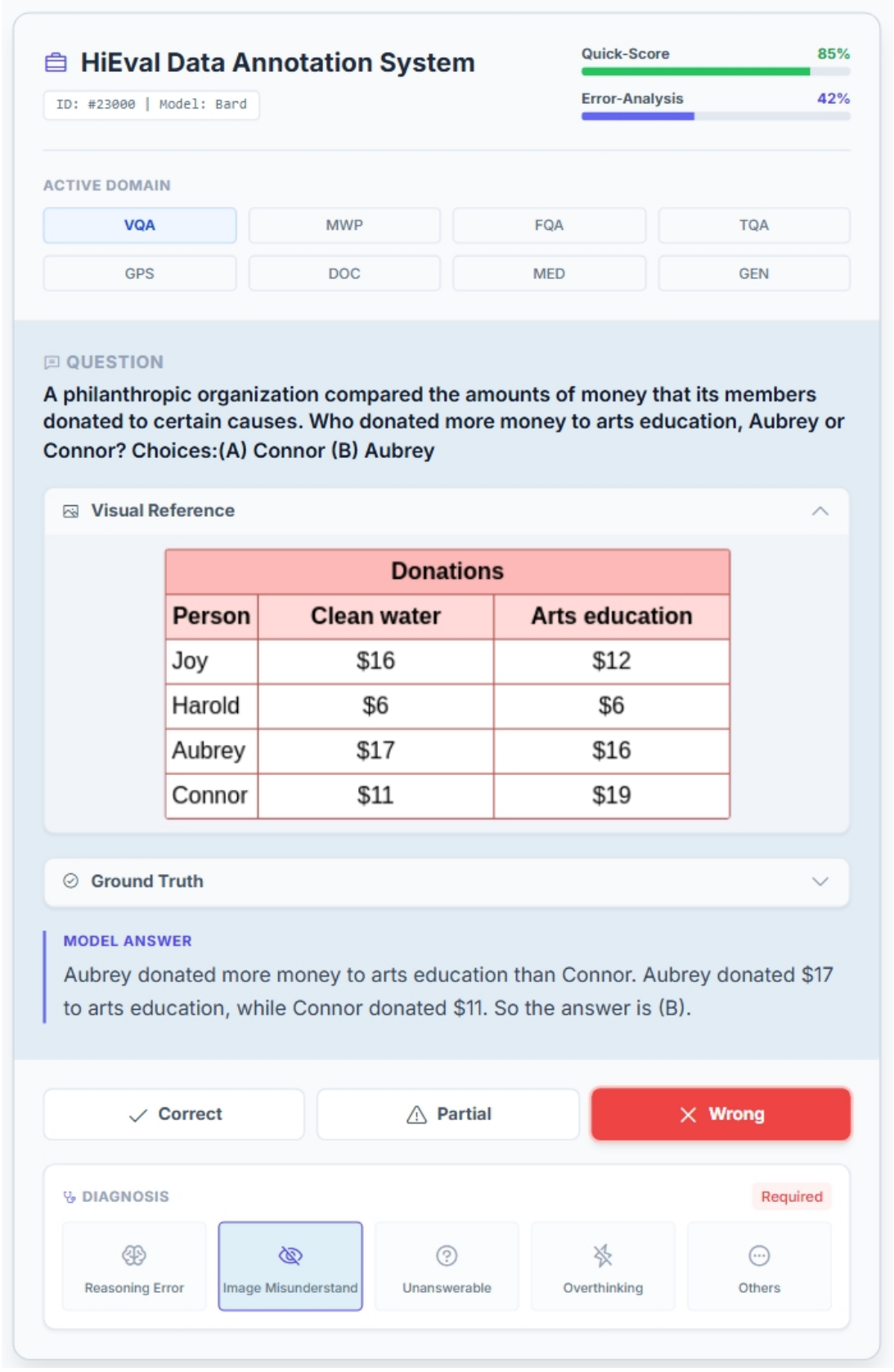
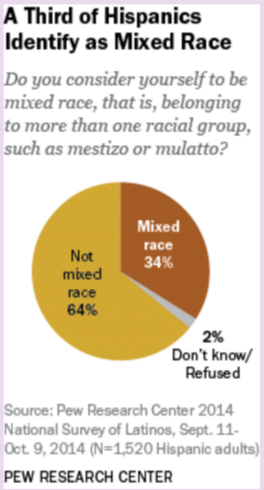


Figure 6: Data Annotation System.

Correct

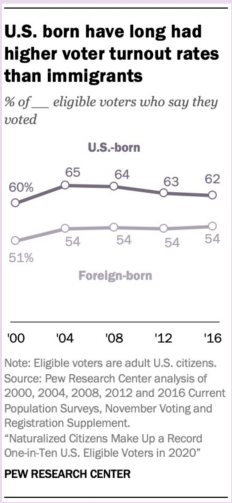
Question: What's the name of biggest pie in the section?
Ground truth answer: Not mixed race
Response: <think>\n</think>According to <img 1>, the answer is Not mixed race



Relaxed-Accuracy: Correct ✔
Qwen3-4B-Ins: Wrong ✖
Gemini-2.0-Flash: Wrong ✖
GPT-5: Wrong ✖
HiEval-QS: Correct ✔

Correct

Question: What is the percentage of U.S.-born eligible voters in 2016?
Ground truth answer: 62
Response: <think>\n</think>In 2016, 62% of U.S.-born eligible voters voted.



Relaxed-Accuracy: Wrong ✖
Qwen3-4B-Ins: Correct ✔
Gemini-2.0-Flash: Wrong ✖
GPT-5: Wrong ✖
HiEval-QS: Correct ✔

Figure 7: Correct Cases.

Partial

Question: What is the central wavelength of the peak?
Ground truth answer: 490;495;510;520
Response: <think>\nSo, let's look at the graph. The X-axis is Wavelength, ranging from 400 to 700. The peak is the main one in the graph. Let's find the center of the peak. The peak seems to be around 500 nm. Let's check the scale. The peak's highest point is between 490-510, but the central wavelength would be around 500. So the central wavelength is approximately 500.\n</think>500

Relaxed-Accuracy: Wrong ✖
Qwen3-4B-Instruct-2507: Correct ✖
Gemini-2.0-Flash: Correct ✖
GPT-5: Wrong ✖
HiEval-QS: Partial ✔

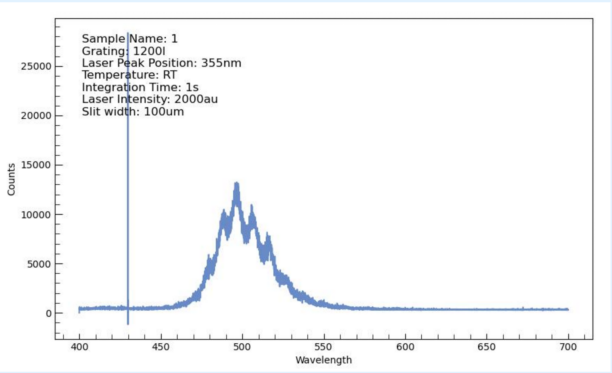


Figure 8: Partial Case.

Wrong

Question: Based on the image provided, can you determine the per capita GDP of Country 1, Country 4 for the year 2017?(A)$ 60739.60, $ 25285.71(B)$ 24451.71, $ 60739.60(C)$ 60739.60, $ 24451.71(D)$ 25285.71, $ 60739.60

Ground truth answer: C

Response: <think>\nGot it, let's figure out the per capita GDP for Country 1 and Country 4 in 2017..... Wait, that's 60,739.6? No, 60.7396 million? No, per capita GDP is in dollars. Wait, no, I think

Relaxed-Accuracy: Wrong ✔

Qwen3-4B-Instruct-2507: Correct ✖

Gemini-2.0-Flash: Correct ✖

GPT-5: Correct ✖

HiEval-QS: Wrong ✔

Year	Country	GDP($)	Population (hundred millions)
2017	Country 1	39.42	6.49
2017	Country 2	6.88	3.04
2017	Country 4	33.67	13.77
2019	Country 5	5.31	2.1
2020	Country 1	35.2	8.73
2020	Country 3	24.42	5.0
2021	Country 1	17.52	1.85
2021	Country 2	27.52	4.41
2021	Country 5	4.49	11.96
2023	Country 3	9.44	16.72

Figure 9: Wrong Case.

Error (Image Misunderstanding)

Question: Please point out where each peak is located in this graph?

Ground truth answer: 305

Response: <think>\nSo, let's look at the graph. The X-axis is Wavelength, ranging from 260 to 400. We need to find the peaks. First, the leftmost peak: around 275 (since the tick is at 280, so the peak is near 275). Then the middle peak: around 310 (since it's between 300 and 320, the peak is around 310). Then the rightmost tall peak: around 365 (since it's near 360-370, the tall one is around 365). Wait, let's check each peak. The first peak is at approximately 275, the second around 310, the third around 365. Let's confirm. The X-axis ticks: 260, 280, 300, 320, 340, 360, 380, 400. So the first peak is before 280, so ~275. The second peak is around 310 (between 300 and 320). The third peak is at ~365 (between 360 and 370). So those are the three peaks.\n</think>275;310;365

GPT-5: Wrong; Reasoning Error ✖

Gemini-2.0-Flash: Partial; Reasoning Error ✖

Mimo-RL: Wrong; Image Misunderstanding ✔

Qwen3-VL-4B: Wrong; Reasoning Error ✖

HiEval-EA: Wrong; Image Misunderstanding ✔

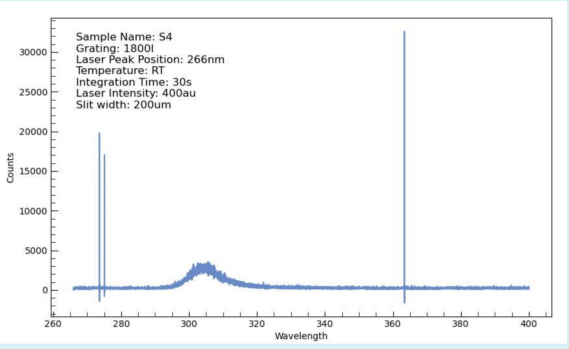


Figure 10: Image Misunderstanding Case.

Error (Reasoning Error)

Question: How many triangles can be drawn with vertices in the 18 points shown in the figure?<image1>

Ground truth answer: 711

Response: We first choose 3 vertices from the 18 points. This can be done in \\({}18 \\choose 3}\\) ways. For each choice of 3 vertices, we can draw a unique triangle. Hence, there are \\({}18 \\choose 3}\\) triangles.\n\n\\(\\boxed{ {}18 \\choose 3} \\)

GPT-5: Wrong; Image Misunderstanding ✖

GPT-4o: Wrong; Image Misunderstanding ✖

Mimo-RL: Wrong; Reasoning Error ✔

Qwen3-VL-4B: Wrong; Image Misunderstanding ✖

HiEval-EA: Wrong; Reasoning Error ✔

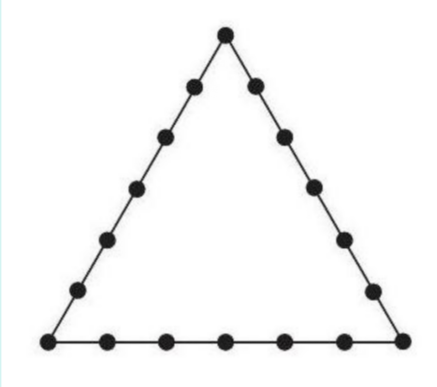


Figure 11: Reasoning Error Case.

Error (Overthinking)

Question: What is the value of each peak?

Ground truth answer: 760

Response: ◁think▷Okay, let's see. The user is asking about identifying the vertical coordinates (Y-axis values) of the peaks in the provided graph. The graph shows a spectrum with counts on the Y-axis and wavelength on the X-axis.unts..... Let me check the X-axis at 360 nm. The main peak is centered at 365, but there's a slight increase before that. Maybe around 363-364 nm, a small peak. But that's part of the rising edge. Alternatively, maybe it's a separate peak. But I think that's part of the same peak's tail.\n\nSo the main peak is at around 900 counts, then the next peak is around 750, then 650. But I need to see the exact positions. Alternatively, maybe the graph has two peaks: the main one and one to the right. Wait, looking at the graph, after the main peak, there's a dip and then a small peak around 370

GPT-5: Wrong; Reasoning Error ✖

Gemini-2.0-Flash: Wrong; Reasoning Error ✖

Mimo-RL: Wrong; Reasoning Error ✖

Qwen3-VL-4B: Wrong; Image Misunderstanding ✖

HiEval-EA: Wrong; Overthinking ✔

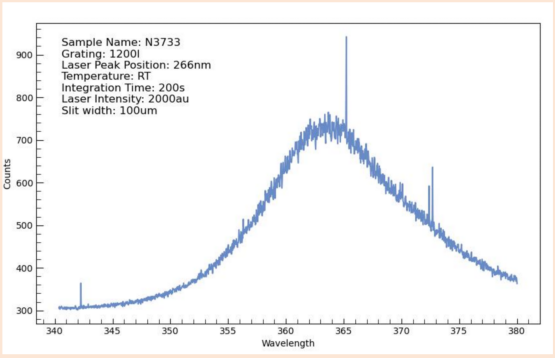


Figure 12: Overthinking Case.

Error (Unanswerable)

Question: How many miles per gallon do the average motorcycle get on the highway?

Ground truth answer: 40

Response: I do not have enough information to provide an exact integer answer for how many miles per gallon the average motorcycle gets on the highway.

GPT-5: Wrong; Reasoning Error ✖

Gemini-2.0-Flash: Partial; Reasoning Error ✖

Mimo-RL: Unanswerable ✖

Qwen3-VL-4B: Wrong; Image Misunderstanding ✖

HiEval-EA: Wrong; Unanswerable ✔

Figure 13: Unanswerable Case.

Error (Other)

Question: Is the number of tiny gray bicycles that are on the left side of the brown metal sedan greater than the number of things that are to the left of the tiny green bicycle? Choices: (A) Yes (B) No

Ground truth answer: Yes

Response: B

GPT-5: Wrong; Image Misunderstanding ✖

Gemini-2.0-Flash: Wrong; Reasoning Error ✖

Mimo-RL: Wrong; Reasoning Error ✖

Qwen3-VL-4B: Wrong; Image Misunderstanding ✖

HiEval-EA: Wrong; Other ✔

Figure 14: Other Case