Not Just Reason, Not Just Scan: Reinforcement Learning for Proactive Scientific Error Verification over Academic Papers

Rongjin Li[1], Yuanxin Liu[2], Hao Zhou[3], Fandong Meng[3], Jie Zhou[3], Xu Sun[2]*

[1]Beijing University of Posts and Telecommunications
[2]State Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[3]WeChat AI, Tencent Inc., China

lirongjin@bupt.edu.cn, xusun@pku.edu.cn

Abstract

Multimodal large language models (MLLMs) are increasingly capable scientific assistants, yet they remain far from fully autonomous research. This transition requires models to actively inspect academic papers, build global evidence views, and make traceable judgments without prespecified issues or evidence. However, existing work provides limited task paradigms or training studies for such issue- and evidence-absent verification. We study this challenge through scientific error detection, where models must determine whether errors exist and justify them with evidence-based reasoning. To fill this gap, we present VERA-RL, a reinforcement-learning formulation for scientific error detection over academic papers. Following a *Reason–Verify–Scan* progression, we construct VERA-13K, a 12,900-sample dataset organized into 4,300 matched chains, covering 6 scientific-error categories across the research workflow and broad natural-science domains. We further introduce fine-grained rewards for reasoning completeness, evidence alignment, and error precision. Training Qwen3-VL-8B with VERA-RL substantially improves verifiable reasoning, approaching flagship MLLMs such as Gemini 3 Pro and Qwen3-VL-235B-A22B on *Scan*.

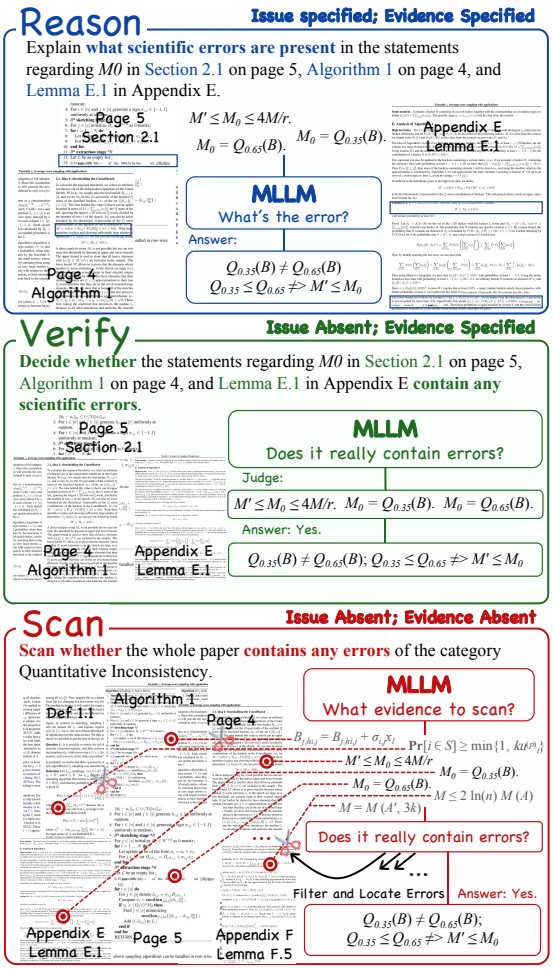


Figure 1: Overview of *Reason-Verify-Scan* paradigm.

1 Introduction

End-to-end independent scientific research by multimodal large language models (MLLMs) (Bai et al., 2025; ByteDance Seed Team, 2025; OpenAI, 2026; Anthropic, 2025) is a long-standing goal of Deep Research and a milestone toward AGI (Ge et al., 2023; Morris et al., 2024). In recent years, MLLMs have become increasingly embedded in the research workflow, automating various aspects of scientific work (He et al., 2025). For example, Google Deep Research (Comanici et al., 2025) aggregates multi-source materials into structured reports, while Paper Decision (Zhang et al., 2026) synthesizes reviews and rebuttals with multi-agent systems to predict final decisions. Together, these systems point to a broader shift from passive information synthesis toward active scientific inquiry.

As research tasks move beyond answering given questions or summarizing existing work, paper-level judgment becomes central to autonomous research: deciding what should be checked, gathering dispersed evidence, and assessing whether claims are supported by the paper. Such judgment un-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 9587–9609
October 24-29, 2026 ©2026 Association for Computational Linguistics

derlies research-quality assessment, idea discovery, and refinement of scientific work.

However, current MLLMs remain far from this goal. A key limitation is that most existing methods still operate in specified settings, where both the target problem and the supporting evidence are provided in advance (Xu et al., 2025; Wang et al., 2025a). Recently, ScholScan (Li et al., 2026) explicitly defined the *Scan* task paradigm, which requires models to perform global verification and consistency checking across the full text, rather than relying on evidence anchored in specific sections. However, existing work still provides limited discussion on methodologies and training paths for systematically enhancing the *Scan* capability.

Building on ScholScan, we study how to systematically train MLLMs for *Scan* rather than treating it only as an evaluation task. We instantiate this setting as scientific error detection and propose VERA-RL, a training formulation that couples progressive task decomposition with reinforcement learning (RL) rewards for verifiable reasoning over academic papers (Rafailov et al., 2024; Schulman et al., 2017; Shao et al., 2024; Yu et al., 2025). To support training, we construct VERA-13K from controlled edits on accepted papers and objective errors extracted from peer reviews, converting each error into *Reason–Verify–Scan* chains (Figure 1) across six categories.

In our experiments, we observe that evidence-specified reasoning does not naturally transfer to *Scan*, leading to a cue-removal gap and making direct end-to-end training on the full *Scan* setting less effective. VERA-RL substantially improves Qwen3-VL-8B on VERA-13K and shows measurable transfer to ScholScan. Further analyses show that *Reason* provides an internal-knowledge reference, while ablations confirm that both staged tasks and multi-dimensional rewards are necessary for stable gains.

In summary, our contributions are as follows:

- We formulate a three-stage *Reason–Verify–Scan* paradigm that turns *Scan* into a trainable progression from evidence-specified reasoning to issue- and evidence-absent verification.
- We build VERA-13K with a reusable construction pipeline and 12,900 filtered samples covering 6 scientific-error categories across typical risk points in the research process and a wide range of natural-science domains.
- We introduce fine-grained rewards for reasoning completeness, evidence alignment, and error precision, enabling RL to target the core dimensions required by scientific error verification.
- Experiments show that VERA-RL substantially improves Qwen3-VL-8B-Instruct and reaches performance comparable to the Qwen3-VL-235B-A22B series. Ablations further show that both the staged task paradigm and the reward system are essential for stable gains.

2 Related Work

2.1 Academic Paper Understanding

Academic papers differ from general documents in their dense domain knowledge and rigorous logic. While previous work focused on isolated elements such as paragraphs or figures, it often overlooked the added challenge of reasoning over complete documents. (Chen et al., 2026; Wang et al., 2024; Auer et al., 2023; Li et al., 2024) Recent studies have adopted full-document inputs, but often treat the paper as a sparse mix of key passages and irrelevant text. (Ma et al., 2024; Yan et al., 2025; Lou et al., 2025; Zhao et al., 2024) This framing narrows the evaluation scope, reducing it to long-context retrieval paired with localized reasoning.

More critically, most benchmarks still adopt a QA paradigm, diverging from real-world scientific tasks. Efforts like PRISMM-Bench (Selch et al., 2025; Xi et al., 2025; Tu et al., 2026) simulate reviewer-style understanding, yet still embed explicit clues and presuppose answer existence. ScholScan (Li et al., 2026) introduces the *Scan*-oriented task, targeting assumption-absent and evidence-absent conditions. However, it treats *Scan* as a static capability without examining how it can be effectively trained. A summary table comparing representative benchmarks is provided in Appendix A.

2.2 RL for LLM Reasoning

Since DeepSeek-R1 and OpenAI-o1 (Guo et al., 2025; OpenAI et al., 2024), large-scale reinforcement learning has become a key approach for improving the reasoning capabilities of LLMs. Recent studies have extended RL to more complex settings, including multimodal and long-context inputs. However, their task and data designs are not yet well aligned with document-level scientific verification. LoongRL and QwenLong-L1 (Wang et al., 2025c; Wan et al., 2025) extend RL to long-context settings through paragraph concatenation, while

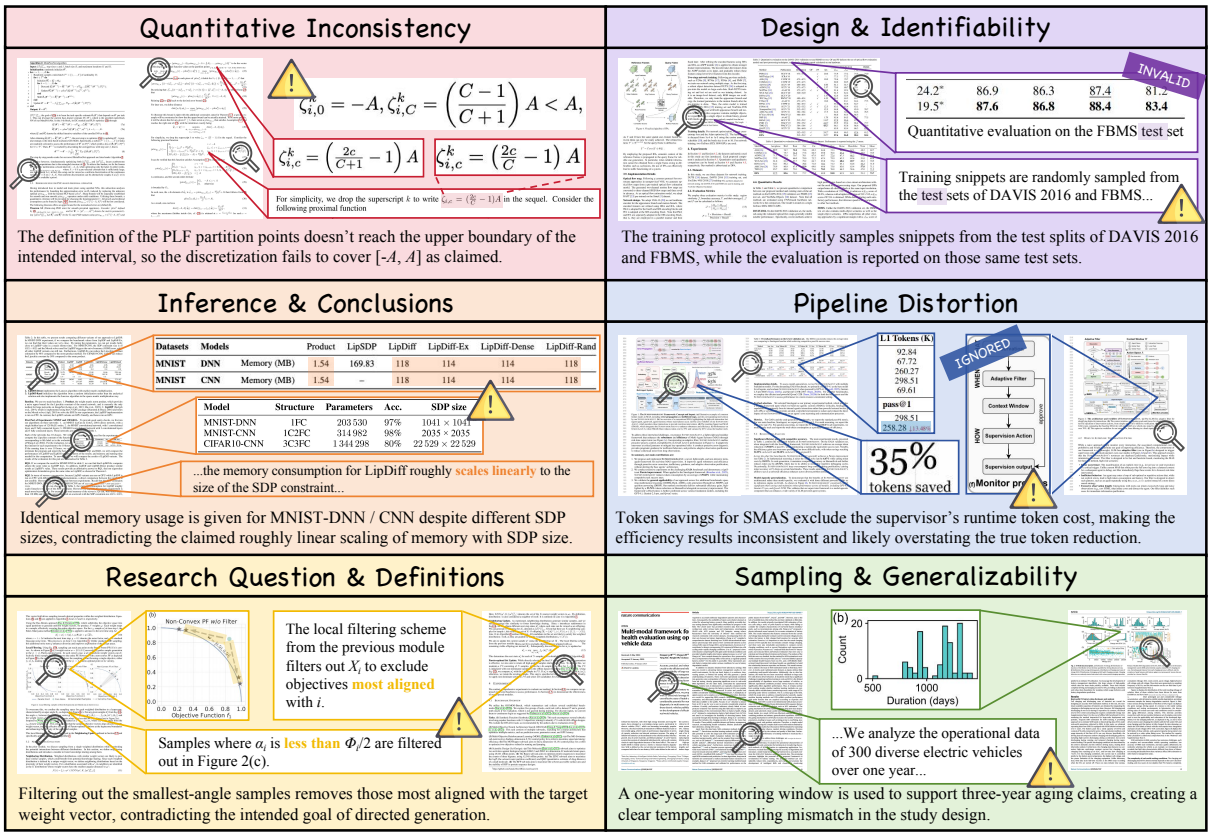


Figure 2: Sampled VERA-13K examples with 6 error types. We consolidate and extend the taxonomy proposed in ScholScan, resulting in a more coherent and well-structured categorization scheme.

VRAG-RL (Wang et al., 2025b) incorporates retrieval into the RL pipeline. These designs improve long-context grounding or retrieval-conditioned reasoning, but provide limited guidance for training *Scan*-style verification over full academic papers.

3 Methodology

3.1 Task Definition

We formulate scientific error detection as producing a structured answer from a paper input, consisting of evidence points, reasoning steps, and a final judgment. Based on the availability of issue and evidence cues, we define 3 stages: *Reason* assumes an error with specified evidence; *Verify* provides candidate evidence without assuming an error; and *Scan* is both issue- and evidence-absent, specifying only a broad scanning target such as an error type. Question examples are shown in Figure 1, and the prompt and answer format is provided in Table 2.

Each underlying error instance can be converted into a matched *Reason–Verify–Scan* chain by rewriting the query while preserving the corresponding evidence and reasoning structure.

3.2 Dataset Construction

Overview Following ScholScan, we consolidate and extend its error taxonomy into 6 categories aligned with major failure points in the research workflow (Figure 2). We construct VERA-13K, a dataset of 12,900 samples for training and evaluation, derived by rewriting 4,300 errors into *Reason–Verify–Scan* chains. Table 1 summarizes the splits and composition of VERA-13K, where each subset consists of complete three-stage task chains.

Data Collection We collect data from two complementary sources. First, we curate accepted papers from top-tier journals (*e.g.*, Nature Communications) and premier conferences (*e.g.*, ICML), which provide high-quality and clean materials. Second, we incorporate reviews from ICLR submissions, which provide human-written feedback on potential weaknesses in submitted papers. Additional statistics are provided in Appendix A.

Data Generation and Quality Control For accepted papers, we instruct Gemini 3 Flash (Google DeepMind, 2025) to inject scientific errors through paragraph-level edits spanning multiple sections,

Table 1: Detailed statistics and distributions of our train and test datasets. Length is calculated by the Qwen3 tokenizer. *Scan*, *Verify*, and *Reason* are strictly balanced in every split and are constructed as a chained curriculum from the same underlying instances. **QI**: Quantitative Inconsistency; **DI**: Design & Identifiability; **IC**: Inference & Conclusions; **PD**: Pipeline Distortion; **RQD**: Research Question & Definitions; **SG**: Sampling & Generalizability.

Statistics	Train Dataset		Test Dataset						
	SFT	RL	Avg.	QI	DI	IC	PD	RQD	SG
# Examples	10,500	1,500	-	150	150	150	150	150	150
Avg. Length	31,134	30,649	31,837	29,956	37,066	31,985	31,100	32,792	28,124
Avg. Evidence	3.54	3.52	3.58	3.28	3.28	3.80	3.60	3.92	3.62
Avg. Reasoning Step	3.81	3.79	3.79	3.77	3.85	3.72	3.84	3.81	3.74

Table 2: QA format used in VERA-13K.

Question Prompt:
Read the provided paper and answer the question. The question will ask you to check and reason about scientific errors in specific parts of the paper. In your answer, show your reasoning process and explain in detail the exact nature of the errors.
`<Question>` `<Interleaved context of the paper>`

Gold Answer:
Evidence: (Only for *Scan* questions)
`- <Evidence 1>` `- ...` `- <Evidence n>`
Reasoning:
`- <Step 1>` `- ...` `- <Step n>`
Answer: ...

under strict category definitions and constraints. For review-derived papers, it extracts objective scientific errors while filtering out subjective feedback. Each instance is rewritten into a *Reason–Verify–Scan* chain, with formulations standardized as in Table 2. We then use Seed-1.6-Thinking for Pass@4 filtering, retaining a sample only judged as correct or partially correct. This step filters out unverifiable or weakly supported annotations. Prompts and other details are provided in our repository; see Appendix G.

3.3 RL for Verifiable Reasoning

Algorithm We optimize the policy with DAPO, a mature RL algorithm suitable for the structured outputs required by academic-paper verification, and use it to study whether the *Scan* capability can be learned through *Reason–Verify–Scan* staged tasks. For each training sample (q, L, a), where q is the question, L is the input paper, and a is the gold answer, DAPO samples G trajectories $\{y_i\}_{i=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$ and updates the current policy π_θ by maximizing:

$$J_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,L)\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{\sum_{i=1}^{G}|y_i|}\right] \cdot \Bigg(\sum_{i=1}^{G}\sum_{t=1}^{|y_i|} \min\Bigg(\frac{\pi_\theta(y_{i,t} \mid q, L, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid q, L, y_{i,<t})} A_{i,t}, \text{clip}\left(\frac{\pi_\theta(y_{i,t} \mid q, L, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid q, L, y_{i,<t})}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right) A_{i,t} \Bigg)\Bigg) - \beta \cdot \mathrm{D}_{\text{KL}}\left[\pi_\theta(\cdot \mid q, L) \,\|\, \pi_{\text{ref}}(\cdot \mid q, L)\right]$$

Here $A_{i,t}$ is the token-level advantage for trajectory y_i at token t, ϵ_{low} and ϵ_{high} are asymmetric clipping bounds, β controls the KL penalty, and π_{ref} is the reference policy. The advantage is normalized within the sampled trajectory group:

$$A_{i,t} = \frac{r_i - \text{mean}(\{r_i\}_{i=1}^{G})}{\text{std}(\{r_i\}_{i=1}^{G})}$$

where r_i is the reward assigned to trajectory y_i.

Rewards For a given question q, reference answer a, and model trajectory y, we design rewards to capture both process-level correctness and task-specific verification requirements. $R_{\text{completeness}}$ extends conventional answer correctness to the reasoning process. Since scientific error detection often admits partially correct answers, it measures the proportion of reference answer points covered by the trajectory rather than only judging the final answer. Beyond this correctness-oriented signal, *Scan*-style verification imposes two additional requirements. First, the model must ground its judgment in the paper rather than produce unsupported explanations, so $R_{\text{alignment}}$ measures the overlap between generated and reference evidence points. Second, open-ended error detection may incentivize excessive candidate errors, so $R_{\text{precision}}$ penalizes unsupported error claims. Together, these terms define rewards for reasoning completeness, evidence grounding, and error precision:

$$
\begin{aligned}
R_{\text{final}} &= \omega_1 R_{\text{completeness}} + \omega_2 R_{\text{alignment}} + \omega_3 R_{\text{precision}} \\
&= \omega_1 \frac{|\hat{\mathcal{R}} \cap \mathcal{R}^*|}{|\mathcal{R}^*|} + \omega_2 \frac{2|\hat{\mathcal{E}} \cap \mathcal{E}^*|}{|\hat{\mathcal{E}}| + |\mathcal{E}^*|} \\
&\quad + \omega_3 \mathbb{I}_{\text{error}} e^{-0.4m}.
\end{aligned}
$$

Here $\hat{\mathcal{R}}$ and $\hat{\mathcal{E}}$ denote the reasoning points and evidence points extracted from the trajectory, while $\mathcal{R}^*$ and $\mathcal{E}^*$ denote the corresponding reference points. m denotes the number of unsupported error claims, and $\mathbb{I}_{\text{error}}$ is 1 for an error prediction and 0 otherwise. These quantities are extracted and matched by Seed-1.6-Thinking as a fixed structured evaluator, while the final reward is computed by the above rule rather than directly assigned by the evaluator. This structured evaluation is less dependent on unconstrained LLM-as-a-Judge preferences.

For *Reason* and *Verify*, evidence is specified, so we disable $R_{\text{alignment}}$ and set $(\omega_1, \omega_2, \omega_3) = (0.6, 0, 0.4)$, keeping reasoning as the main target while using precision as a constraint. For *Scan*, where evidence must be identified rather than given, we set $(0.4, 0.4, 0.2)$ so that $R_{\text{alignment}}$ receives a role comparable to $R_{\text{completeness}}$. Appendix D reports statistics supporting this choice.

4 Experiments

4.1 Setup

Inputs Prior work typically relies on either plain-text OCR, which discards all visual elements, or full-page image rendering, which demands high resolution to retain dense textual content. (Ma et al., 2024; Selch et al., 2025) These approaches fall short for our task, where figure-text alignment and layout preservation are essential. We instead adopt DeepSeek-OCR (Wei et al., 2025) to convert papers into interleaved text and images, preserving most of the original information.

Training Setup We train Qwen3-VL-8B-Instruct with 1 epoch of SFT using a batch size of 8, followed by 30 RL steps with a batch size of 32 and random sampling over all task types. Further details are provided in Appendix B.

Baselines We evaluate 8 models spanning flagship proprietary models and Qwen3-VL variants. For selected Qwen3-VL models, we compare their Instruct and Thinking versions to assess how training schemes affect verifiable reasoning.

4.2 Main Results

Table 3 reports the main *Scan* results, with the full *Reason* and *Verify* results provided in Appendix B. Based on these results, we make 5 observations.

Current MLLMs remain generally weak. Although Gemini 3 Pro is among the strongest baselines, it only reaches 60.0 on *Reason*, near the passing threshold, and remains far from reliable on *Scan* (24.3), revealing a persistent limitation of current MLLMs in constructing and reasoning over global evidence views.

Evidence removal creates a sharp capability gap. Even Gemini 3 Pro, the strongest baseline on Reason, only barely achieves a passing threshold on the *Reason* task, while performing considerably worse on *Scan*. This reveals fundamental limitations in current models and training paradigms when it comes to constructing and reasoning over global evidence views. The two instruction-tuned versions of Qwen3-VL perform especially poorly, scoring below 10 and 20 points on *Verify* and *Scan* respectively, highlighting the difficulty of these higher-level tasks.

Post-training leads to significant performance improvements. The SFT- and RL-finetuned variants of Qwen3-VL-8B demonstrate consistent overall improvements across all rewards. A similar pattern of improvement is observed for the Qwen3-VL-235B-A22B series. Our RL-trained model achieves substantial improvements on both *Verify* and *Scan*. It approaches the Qwen3-VL-235B-A22B-Thinking on $R_{\text{completeness}}$ and $R_{\text{alignment}}$, surpasses it in terms of the overall composite score, and further narrows the gap to Gemini 3 Pro. In particular, SFT helps regularize the output through basic format and distributional constraints, leading to a noticeable increase in $R_{\text{precision}}$. This improvement is still evident in the RL phase, where the reward metrics continue to grow, indicating that all 3 metrics play a crucial role in the final performance.

VERA-RL shows measurable transfer to ScholScan. We further evaluate on ScholScan (Table 4), and the results mirror those observed above. Notably, although absolute scores remain low in this harder cross-benchmark setting, VERA-RL moves Qwen3-VL-8B from near-zero performance to non-trivial scores and brings it close to Qwen3-VL-235B-A22B-Instruct on several metrics. This suggests that the learned *Reason–Verify–Scan* capability is not limited to VERA-13K. More details

Table 3: *Scan*-task evaluation results (scaled by 100) for baselines and models trained in the main experiments.

Models	Avg.	QI	DI	IC	PD	RQD	SG
Scan $R_{\text{completeness}}$							
Gemini 3 Pro	22.8	24.4	13.2	30.0	17.7	25.8	26.0
Qwen3-VL-235B-A22B (Thinking)	16.7	22.2	12.5	21.7	10.2	13.3	20.3
Qwen3-VL-235B-A22B (Instruct)	3.4	7.0	3.3	7.8	1.5	0.0	0.5
Qwen3-VL-8B (Instruct)	1.5	0.0	0.0	5.3	1.5	0.0	2.0
Qwen3-VL-8B (SFT, ours)	5.3	5.2	1.3	5.8	6.3	3.0	10.3
Qwen3-VL-8B (SFT+RL, ours)	8.2	7.8	0.5	10.0	6.8	11.3	12.8
Scan $R_{\text{alignment}}$							
Gemini 3 Pro	17.8	22.7	9.4	25.7	15.0	17.4	16.9
Qwen3-VL-235B-A22B (Thinking)	13.0	21.4	8.9	17.5	7.1	11.0	12.2
Qwen3-VL-235B-A22B (Instruct)	2.4	4.7	2.0	6.3	0.8	0.0	0.8
Qwen3-VL-8B (Instruct)	1.0	0.0	0.0	3.9	2.0	0.0	0.0
Qwen3-VL-8B (SFT, ours)	2.5	1.1	0.8	5.3	1.1	0.8	4.8
Qwen3-VL-8B (SFT+RL, ours)	6.2	9.2	1.0	7.7	4.5	6.4	8.4
Scan $R_{\text{precision}}$							
Gemini 3 Pro	40.4	45.7	43.5	41.3	42.1	35.3	34.4
Qwen3-VL-235B-A22B (Thinking)	27.5	38.2	32.1	26.5	26.2	21.3	20.9
Qwen3-VL-235B-A22B (Instruct)	5.4	6.8	3.8	19.0	0.4	0.0	2.3
Qwen3-VL-8B (Instruct)	5.0	3.0	3.3	14.8	2.2	2.3	4.5
Qwen3-VL-8B (SFT, ours)	56.3	52.3	57.1	55.7	54.0	58.3	60.2
Qwen3-VL-8B (SFT+RL, ours)	68.6	67.0	65.2	68.9	70.4	69.6	70.3
Scan R_{final}							
GPT-5.4	29.9	49.9	26.2	22.8	31.8	26.2	23.8
Gemini 3 Pro	24.3	28.0	17.7	30.3	21.5	24.4	24.0
Seed-1.6-Thinking	21.2	25.9	17.1	27.5	16.1	22.6	17.6
Qwen3-VL-Plus	26.2	35.7	17.7	30.0	22.1	24.5	27.3
Qwen3-VL-235B-A22B (Thinking)	17.4	25.1	15.0	21.0	12.2	14.0	17.2
Qwen3-VL-235B-A22B (Instruct)	3.4	6.0	2.9	9.4	1.0	0.0	1.0
Qwen3-VL-8B (Instruct)	2.0	0.6	0.7	6.7	1.8	0.5	1.7
Qwen3-VL-8B (SFT, ours)	14.4	13.0	12.3	15.6	14.2	13.2	18.1
Qwen3-VL-8B (SFT+RL, ours)	19.5	20.2	13.6	20.9	18.6	21.0	22.5

and analysis are provided in Appendix C.

***Reason* provides a reference for the internal-knowledge upper bound.** We observe that *Reason* shows relatively limited improvement during post-training, while *Verify* approaches *Reason*. We consider *Reason* and $R_{\text{completeness}}$ to be more strongly shaped by the model's parametric prior knowledge, making them useful reference points for the upper bound of capability under internal knowledge alone. In contrast, *Verify* and *Scan* demand stronger external evidence grounding and higher-level verifiable reasoning, thus reflecting a shift in the nature of capability from internal recall to evidence-based inference.

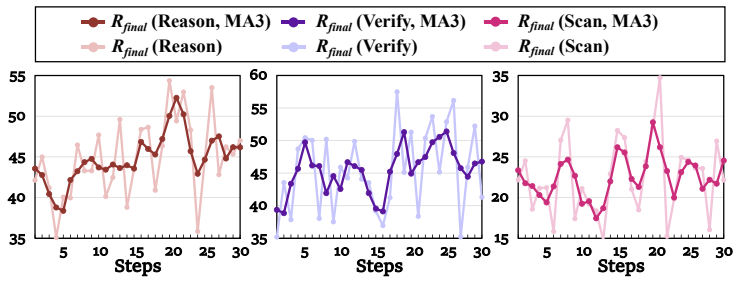
4.3 Additional Analysis

Training Dynamics. Figure 3 shows the rollout dynamics of the main experiment. Rewards increase in phases with moderate oscillations, indicating that the staged tasks provide usable optimization signals despite the long and structured outputs. Additionally, *Reason* and *Verify* exhibit converging patterns in the fine-grained perspective of a single training run. This is consistent with our observation that *Reason* serves as an internal-knowledge reference point.

The improvements also vary across error categories. For RQD, PD, and SG, post-training brings larger gains than parameter scaling. These error types often involve domain-specific concepts introduced within the paper, reducing reliance on the model's parametric knowledge and shifting the demand toward evidence-grounded verification. DI and IC yield limited improvements, suggesting that errors involving experimental design and final conclusions require stronger holistic understanding and a higher degree of internal scientific knowledge. For QI, the model shows broad improvements under both paths, which we attribute to its additional demand on numerical computation and analysis.

Table 4: Evaluation results (scaled by 100) for baselines and models trained in the main experiments. S_{reason}, S_{location}, $P_{\text{unrelated_err}}$, and $S(m)$ are evaluation metrics introduced in ScholScan, with detailed definitions provided in Appendix C. $R_{\text{precision}}$ is not applicable in this evaluation setting and is therefore omitted.

Models	$R_{\text{completeness}}$	$R_{\text{alignment}}$	S_{reason}	S_{location}	$P_{\text{unrelated_err}}$	$S(m)$
Gemini 3 Pro	11.3	7.2	11.0	6.0	10.2	5.5
Qwen3-VL-235B-A22B-Thinking	4.7	2.8	4.6	2.4	5.3	2.6
Qwen3-VL-235B-A22B-Instruct	0.8	0.4	0.8	0.3	0.4	0.1
Qwen3-VL-8B-Instruct	0.0	0.0	0.0	0.0	0.0	0.0
Qwen3-VL-8B (SFT)	0.4	0.1	0.4	0.0	3.8	0.0
Qwen3-VL-8B (SFT+RL)	0.5	0.3	0.5	0.2	0.5	0.2



	R_{final}	R_{comp}	R_{align}	R_{pre}
↑	285	40	36	287
↓	6	5	4	3
=	9	255	260	10
Total	300	300	300	300

Figure 3: **Left**: Training dynamics of main experiments. **Right**: Metrics change after VERA-RL on *Scan* (*test* set).

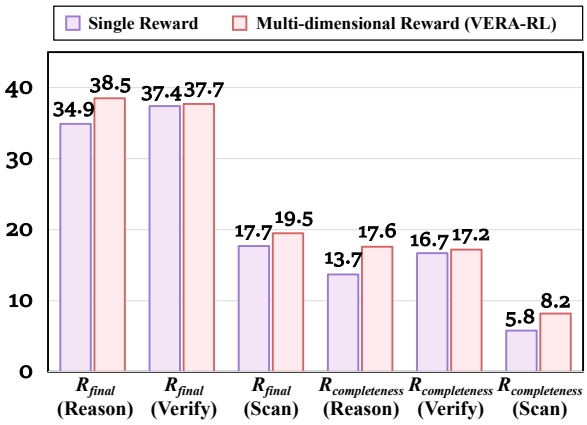


Figure 4: Performance comparison between multi-dimensional and single reward configurations.

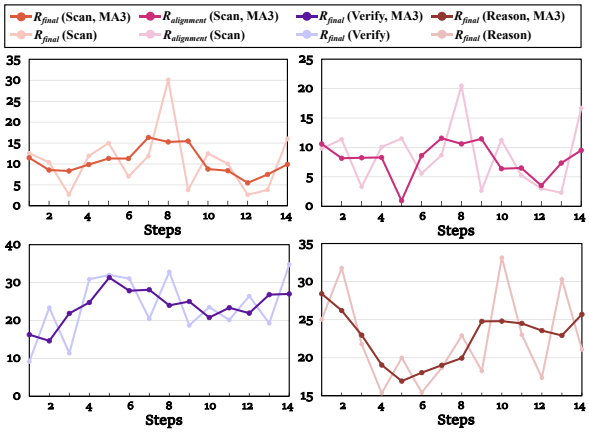


Figure 5: Training dynamics of the ablation study using only $R_{\text{completeness}}$.

4.4 Ablation Study of Reward Design

To explore the necessity of our multi-dimensional reward system, we compare it with a single-reward variant using only $R_{\text{completeness}}$. As shown in Figure 4, the multi-dimensional reward consistently outperforms the single-reward variant on both R_{final} and $R_{\text{completeness}}$. Furthermore, the training dynamics in Figure 5 reveal that reverting to such a sparse reward signal leads to degradation in overall performance. From a task-paradigm perspective, the model shows little meaningful improvement on *Reason* and *Verify*, while its performance on the *Scan* task collapses. More notably, from the perspective of reward optimization itself, the model struggles to stably improve even the explicitly rewarded metric, with training exhibiting clear signs of instability and regression. The reward components are not independent. Instead, they form a tightly coupled positive-feedback structure that collectively and precisely defines the capability dimensions required by the task.

4.5 Ablation Study of Task Framework

To evaluate the effectiveness of the three-stage task paradigm, we compare the main setting with two alternatives in Table 5. The first setting, *Pure Scan*, trains the model only on *Scan* samples. When controlling for the amount of *Scan* exposure at

Table 5: Performance comparison across different configurations for *Reason*, *Verify*, and *Scan* tasks. The metric R_{com} refers to $R_{\text{completeness}}$ and R_{align} refers to $R_{\text{alignment}}$. The *Pure Scan* configuration isolates the *Scan* task for training.

Config.	**Step**	***Reason***			***Verify***			***Scan***			
		R_{final}	R_{com}	$R_{\text{precision}}$	R_{final}	R_{com}	$R_{\text{precision}}$	R_{final}	R_{com}	R_{align}	$R_{\text{precision}}$
Main	30	38.5	17.6	69.9	37.7	17.2	68.5	19.5	8.2	6.2	68.6
Main	15	36.1	14.6	68.4	35.0	14.9	65.2	15.0	2.8	2.3	65.0
Pure Scan	10	37.0	16.7	67.4	34.0	12.8	65.8	16.8	5.3	3.7	66.0
Δ *to Main*(30)	-	-1.5	-0.9	-2.5	-3.7	-4.4	-2.7	-2.7	-2.9	-2.5	-2.6
Pure Scan	20	37.2	15.9	69.1	34.1	12.6	66.3	16.9	5.3	4.2	65.4
Δ *to Main*(30)	-	-1.3	-1.7	-0.8	-3.6	-4.6	-2.2	-2.6	-2.9	-2.0	-3.2
Curricular	20	36.3	14.3	69.2	35.8	15.1	66.8	18.2	6.8	5.2	67.3
Δ *to Main*(30)	-	-2.2	-3.3	-0.7	-1.9	-2.1	-1.7	-1.3	-1.4	-1.0	-1.3
Curricular	30	36.5	15.5	67.9	36.1	15.1	67.6	16.5	4.5	4.2	64.8
Δ *to Main*(30)	-	-2.0	-2.1	-2.0	-1.6	-2.1	-0.9	-3.0	-3.7	-2.0	-3.8

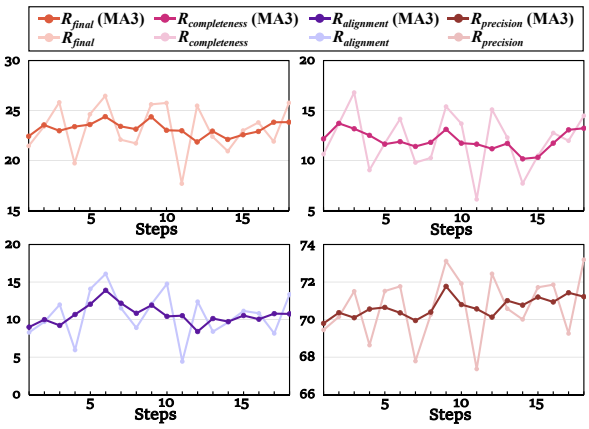


Figure 6: Training dynamics of the ablation experiment using only *Scan* tasks.

step 10, *Pure Scan* performs worse than the main setting not only on *Reason* and *Verify*, but also on *Scan* itself. Extending *Pure Scan* training to step 20 still fails to close the gap. Just as the reward components are complementary, the task stages are also coupled in training. *Reason* and *Verify* remain non-trivial, and removing their supervision weakens the evidence-conditioned reasoning and judgment needed for *Scan*. Neglecting the training of these tasks may hinder the development of *Scan* performance.

Further training does not yield stable improvement on *Reason*, and *Scan* also improves slowly. Figure 6 illustrates the reward dynamics when training solely on the *Scan* task, which reveal the struggle across nearly all metrics from a fine-grained perspective.

Furthermore, we introduced an attempt with curricular learning (Yuan et al., 2025). Specifically, we used the main experiment setup for the first 15 steps, then switched to training using only *Scan* samples, and evaluated models at total steps 20 and 30. The results in Table 5 indicate that this setup faces similar challenges as the setup using only *Scan* task samples. This further underscores the tight coupling of our task paradigm framework. From a more practical perspective, it enables the expansion of sample size by breaking down capability dimensions, efficiently utilizing valuable high-quality papers, and reserving greater potential for scaling up.

4.6 Case Study

We further examine *Scan* samples improved by VERA-RL. Statistics in Figure 3 show broad gains across many instances, and the cases are provided in Appendix E. Qualitatively, the base Qwen3-VL-8B-Instruct often stops at surface-level consistency checks and gives broad no-error judgments. After VERA-RL, the model searches more actively across the paper, identifies localized evidence from dispersed content, and connects it into traceable error reasoning. These cases suggest that the gains are not merely from catering to evaluator heuristics, but reflect stronger full-paper exploration and evidence-grounded verification required by *Scan*.

5 Conclusion

We presented VERA-RL for training verifiable scientific error detection over academic papers. By decomposing *Scan* into *Reason*, *Verify*, and *Scan*, and by using rewards for reasoning completeness, evidence alignment, and error precision, VERA-RL provides a principled training path from evidence-specified reasoning to issue- and evidence-absent verification. We also constructed VERA-13K, a 12,900-sample dataset of matched three-stage

chains across 6 scientific-error categories. Experiments on VERA-13K and ScholScan show consistent improvements, and ablations confirm that both task staging and reward design are necessary for stable gains. These results demonstrate that *Scan*-style verification is a trainable capability that can be systematically improved through aligned task design, data construction, and reward modeling, offering a concrete step from scientific assistance toward more autonomous scientific research.

Limitations

This work focuses on scientific error detection as a concrete setting for Scan-style verification over academic papers. It does not cover the full range of peer-review judgments, such as novelty, significance, writing quality, or broader research impact, which often depend on community context and subjective assessment. VERA-13K emphasizes errors that can be verified from the paper itself, making supervision and reward computation more structured. As a result, more implicit weaknesses or errors requiring extensive external domain knowledge may be underrepresented. Our training experiments are mainly conducted on Qwen3-VL-8B. Although we compare with multiple strong MLLMs and external benchmarks, larger-scale RL training, broader model families, and alternative paper representations remain for future study.

Acknowledgments

This work is supported by WeChat AI, Tencent Inc., China and "The Fundamental Research Funds for the Central Universities, Peking University". We would also like to thank the anonymous reviewers and area chairs for constructive discussions and feedback.

References

Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. https://www.anthropic.com/claude-4-system-card. Updated Sep 2, 2025.

S. Auer, Dante Augusto Couto Barone, Cassiano Bartz, E. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry I. Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

ByteDance Seed Team. 2025. Introduction to techniques used in seed1.6. https://seed.bytedance.com/en/blog/introduction-to-techniques-used-in-seed1-6. Official blog post describing Seed1.6 techniques.

Yelin Chen, Fanjin Zhang, Suping Sun, Yunhe Pang, Yuanchun Wang, Jian Song, XiaoYan Li, Lei Hou, Shu Zhao, Jie Tang, and Juanzi Li. 2026. RPC-bench: A fine-grained benchmark for research paper comprehension.

Gheorghe Comanici and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. Openagi: When llm meets domain experts. In *Advances in Neural Information Processing Systems*, volume 36, pages 5539–5568. Curran Associates, Inc.

Google DeepMind. 2025. Gemini 3 Pro Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Daming Guo, Dongdong Yang, Hongyi Zhang, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638.

Yichen He, Guanhua Huang, Peiyuan Feng, Yuan Lin, Yuchen Zhang, Hang Li, and Weinan E. 2025. Pasa: An llm agent for comprehensive academic paper search. *Preprint*, arXiv:2501.10120.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Rongjin Li, Zichen Tang, Xianghe Wang, Xinyi Hu, Zhengyu Wang, Zhengyu Lu, Yiling Huang, Jiayuan Chen, Weisheng Tan, Jiacheng Liu, Zhongjun Yang, and Haihong E. 2026. Not search, but scan: Benchmarking MLLMs on scan-oriented academic paper reasoning. In *The Fourteenth International Conference on Learning Representations*.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang,

Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. 2025. AAAR-1.0: Assessing AI's potential to assist research. In *Forty-second International Conference on Machine Learning*.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLONGBENCH-DOC: Benchmarking long-context document understanding with visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Forty-first International Conference on Machine Learning*.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI. 2026. GPT-5.4 Thinking System Card. https://openai.com/index/gpt-5-4-thinking-system-card/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Lukas Selch, Yufang Hou, M. Jehanzeb Mirza, Sivan Doveh, James Glass, Rogerio Feris, and Wei Lin. 2025. Prismm-bench: A benchmark of peer-review grounded multimodal inconsistencies. *Preprint*, arXiv:2510.16505.

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

Songjun Tu, Yiwen Ma, Jiahao Lin, Qichao Zhang, Xiangyuan Lan, Junfeng. Li, Nan Xu, Linjing Li, and Dongbin Zhao. 2026. Paperaudit-bench: Benchmarking error detection in research papers for critical automated peer review. *Preprint*, arXiv:2601.19916.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *Preprint*, arXiv:2505.17667.

Chengye Wang, Yifei Shen, Zexi Kuang, Arman Cohan, and Yilun Zhao. 2025a. SciVer: Evaluating foundation models for multimodal scientific claim verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8562–8579, Vienna, Austria. Association for Computational Linguistics.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025b. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *Preprint*, arXiv:2505.22019.

Siyuan Wang, Gaokai Zhang, Li Lyna Zhang, Ning Shang, Fan Yang, Dongyao Chen, and Mao Yang. 2025c. Loongrl: Reinforcement learning for advanced reasoning over long contexts. *Preprint*, arXiv:2510.19363.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Preprint*, arXiv:2406.18521.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *Preprint*, arXiv:2510.18234.

Sarina Xi, Vishisht Rao, Justin Payan, and Nihar B. Shah. 2025. Flaws: A benchmark for error identification and localization in scientific papers. *Preprint*, arXiv:2511.21843.

Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025. Can LLMs identify critical limitations within scientific research? a systematic evaluation on AI research papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20652–20706, Vienna, Austria. Association for Computational Linguistics.

Dawei Yan, Yang Li, Qing-Guo Chen, Weihua Luo, Peng Wang, Haokui Zhang, and Chunhua Shen. 2025. Mmcr: Advancing visual language model in multimodal multi-turn contextual reasoning. *Preprint*, arXiv:2503.18533.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, Hao Zhang, and Yu Rong. 2025. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *Preprint*, arXiv:2507.22607.

Yi-Fan Zhang, Yuhao Dong, Saining Zhang, Kai Wu, Liang Wang, Caifeng Shan, Ziwei Liu, Ran He, Hao Zhao, and Chaoyou Fu. 2026. Iclr 2026 acceptance prediction: Benchmarking decision process with a multi-agent system. https://github.com/PaperDecision/PaperDecision. Accessed: 2026-01-18.

Yilun Zhao, Yitao Long, Hongjun Liu, Ryo Kamoi, Linyong Nan, Lyuhao Chen, Yixin Liu, Xiangru Tang, Rui Zhang, and Arman Cohan. 2024. DocMath-eval: Evaluating math reasoning capabilities of LLMs in understanding long and specialized documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16103–16120, Bangkok, Thailand. Association for Computational Linguistics.

Table 6: Comparison with representative academic document understanding benchmarks. **T**: text; **I**: image; **MD**: multimodal document; **CS**: Computer Science.

Benchmark	Modality	Paradigm	Eval.	Domains	Count (K)
CharXiv	I	Reason	Close	8	11.6
ArXivQA	I	Reason	Close	10	100
MMCR	T+MD	Reason	Close	CS	310
AAAR-1.0	T+MD	Reason	Mixed	CS	13.5
PRISMM-Bench	I	Reason	Close	CS	0.4
ScholScan	T+MD	Scan	Open	13	1.8
VERA-13K (ours)	**T+MD**	**Reason+Verify+Scan**	**Open**	**9**	**12.9**

A Details of VERA-13K

Comparison with other benchmarks Table 6 summarizes the differences in modality, task paradigm, etc. with existing academic document understanding benchmarks.

Distribution Figure 7 presents the source distribution of papers in VERA-13K. Table 9 reports the distribution of the 6 categories in the *train* subset. The categories cover typical risk points across the research process, from problem formulation and experimental design to quantitative analysis and final inference.

AI/ML papers are used as a major source because they represent a rapidly growing and highly active research area with abundant public materials and reviews. Also, they contain diverse subfields. We further categorize them according to the ICML call-for-papers taxonomy and report the major represented subfields in Table 8.

Beyond computer science, VERA-13K includes 655 papers and 1,605 samples from broader scientific domains. We group these papers into high-level scientific domains. As shown in Table 7, VERA-13K covers diverse areas beyond computer science.

Source Composition. VERA-13K combines two complementary data sources: controlled-edit samples constructed from accepted papers and review-derived samples based on objective errors extracted from peer reviews. Table 10 reports their proportions in each split. Review-derived samples account for 27.2%, 32.8%, and 30.7% of the SFT, RL, and test sets, respectively, yielding a similar source composition across training and evaluation splits.

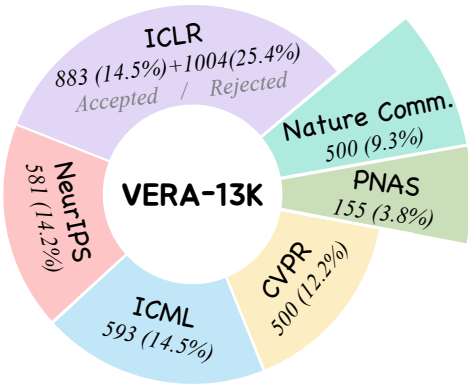


Figure 7: Distribution of paper sources.

B Supplementary Materials of Experiments

B.1 Supplementary Results

Table 11 provides supplementary *Scan* sub-metrics for additional baseline models. Tables 12 and 13 report the full *Reason* and *Verify* results across error categories. For *Reason* and *Verify*, we report $R_{\text{completeness}}$, $R_{\text{precision}}$, and R_{final}, since evidence alignment is disabled in these evidence-specified settings. All scores are scaled by 100.

Beyond ScholScan, we further evaluate the model on three related external benchmarks covering long-document understanding and scientific reasoning. As shown in Table 14, VERA-RL maintains or modestly improves performance across all three benchmarks, suggesting that its gains do not come at the cost of related general capabilities.

We further remove test samples whose source papers overlap with the training set and reevaluate the models. As shown in Table 15, the results remain nearly unchanged, and the gains from SFT and RL persist under the paper-disjoint setting.

Table 7: Broad-domain distribution beyond Computer Science.

Domain	Count
Life Sciences	227
Medicine	107
Materials	67
Ecology	34
Chemistry	32
Environment	30
Physics	23
Interdisciplinary	135
Total	655

Table 8: Major AI/ML subdomain distribution. Counts are paper-level statistics, and only the most represented subfields are listed.

Subdomain	Count
Foundation Models & LLM	424
Computer Vision	412
Multimodal Learning	287
Generative Modeling	243
Reinforcement Learning	194
Evaluation	190
Graph ML	162
Safety	156
Learning Theory	119
AI Agents	115
Efficiency	94

B.2 Training Configurations

SFT Training We perform SFT training on Qwen3-VL-8B-Instruct using a standard teacher-forcing objective. Training is conducted for 1 epoch with a global batch size of 8 and a micro-batch size of 1 per GPU. Optimization is performed using the AdamW optimizer with a learning rate of 5×10^{-6}, no weight decay, and a cosine learning rate schedule. A warm-up ratio of 0.03 is applied, and gradient norms are clipped at 1.0. All experiments use a fixed random seed of 42.

RL Training We fine-tune the SFT-initialized model using the DAPO setup. The base model is Qwen3-VL-8B initialized from the SFT checkpoint. We sample 8 responses per prompt. Policy optimization is performed with a global training batch size of 32. We use asymmetric clipping with clip ratios $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.1, 0.5)$ and adopt a token-level loss aggregation strategy (`token-mean`) to mitigate length bias. The learning rate is set to 2×10^{-6}, with no weight decay and a cosine scheduler with a warm-up ratio of 0.03. Training is run for 30 optimization steps with a fixed random seed of 42. We enable gradient checkpointing, activation offloading, and FSDP with parameter and optimizer offloading.

Table 9: Train-subset distribution across 6 scientific-error categories.

Category	Count
QI	3450
DI	540
IC	2190
PD	1470
RQD	3450
SG	900

B.3 Training Dynamics

We further evaluate the step-200 SFT checkpoint to test whether the early plateau in training loss is sufficient. As shown in Table 16, step 200 already improves over the Instruct model, but the 1-epoch checkpoint achieves stronger *Scan* completeness and a higher overall *Scan* score, so we use the latter to initialize RL training.

C Metrics Definition in ScholScan

Given a model answer a, ScholScan first parses it into a structured tuple:

$$\Psi(a) = \left(I_{\text{exist}}, I_{\text{contain}}, \hat{\mathcal{E}}, \hat{\mathcal{R}}, n\right),$$

where I_{exist} indicates whether the answer asserts any error, I_{contain} indicates whether it contains the annotated target error, $\hat{\mathcal{E}}$ and $\hat{\mathcal{R}}$ are the predicted evidence set and reasoning chain, and n is the number of unrelated error claims. Let $\mathcal{E}^*$ and $\mathcal{R}^*$ denote the gold evidence set and reasoning chain.

The detection score requires the target error to be identified:

$$S_{\text{det}} = I_{\text{exist}} I_{\text{contain}}.$$

The evidence location score measures overlap with the gold evidence while penalizing over-reporting:

$$D_{\mathcal{E}} = \frac{2|\hat{\mathcal{E}} \cap \mathcal{E}^*| + \mathbf{1}\{|\hat{\mathcal{E}}| + |\mathcal{E}^*| = 0\}}{\max(|\hat{\mathcal{E}}| + |\mathcal{E}^*|, 1)},$$

Table 10: Source composition of VERA-13K. Review-derived samples are constructed from objective errors extracted from peer reviews, while controlled-edit samples are generated by injecting errors into accepted papers.

Split	Total	Review-derived	Controlled-edit	Review (%)
SFT Train	10,500	2,856	7,644	27.2
RL Train	1,500	492	1,008	32.8
Test	900	276	624	30.7
Total	12,900	3,624	9,276	28.1

Table 11: Supplementary sub-metric scores on the *Scan* task (scaled by 100).

Models	**Avg.**	**QI**	**DI**	**IC**	**PD**	**RQD**	**SG**
Scan $R_{\text{completeness}}$							
GPT-5.4	39.5	54.3	42.6	33.0	36.8	37.7	32.9
Qwen3-VL-Plus	20.6	34.2	7.5	28.5	16.0	19.2	18.3
Qwen3-VL-32B	12.7	9.0	6.7	19.5	10.0	13.8	17.3
Seed-1.6-Thinking	13.4	19.1	11.2	15.7	8.8	14.0	11.5
Scan $R_{\text{alignment}}$							
GPT-5.4	28.5	60.0	17.0	19.5	37.5	23.1	16.7
Qwen3-VL-Plus	15.3	26.2	6.1	21.4	11.8	12.7	13.9
Qwen3-VL-32B	10.1	12.1	5.0	16.6	7.3	9.1	10.3
Seed-1.6-Thinking	11.1	20.1	8.1	14.7	6.6	9.2	8.2
Scan $R_{\text{precision}}$							
GPT-5.4	13.7	20.6	11.5	9.1	10.4	9.7	19.9
Qwen3-VL-Plus	59.2	57.9	61.4	50.1	54.8	58.9	71.9
Qwen3-VL-32B	60.2	63.1	56.2	58.5	67.7	58.1	57.7
Seed-1.6-Thinking	56.7	51.4	47.0	77.0	49.6	66.8	48.6

$$S_{\text{loc}} = \max\left\{0,\ D_{\mathcal{E}} - 0.8\left(\frac{|\hat{\mathcal{E}} \setminus \mathcal{E}^*|}{\max(|\hat{\mathcal{E}}|, 1)}\right)^2\right\}.$$

For reasoning, ScholScan counts the matched prefix length between the predicted and gold reasoning chains:

$$\hat{g} = \text{prefix_match}(\hat{\mathcal{R}}, \mathcal{R}^*), \qquad g_r = |\mathcal{R}^*|.$$

The reasoning score is then defined as

$$S_{\text{reason}} = \mathbf{1}\{g_r = 0\} + \mathbf{1}\{g_r > 0\}\left(\frac{\hat{g}}{g_r}\right)^2.$$

To penalize unrelated error claims, it defines

$$P_{\text{unrel}}(n) = 0.9^{\min(n,2)} e^{-0.6[\max(n-2,0)]^{1.5}}.$$

The overall score combines target-error detection, evidence location, reasoning faithfulness, and the unrelated-error penalty:

$$S(a) = S_{\text{det}} \cdot \sqrt{S_{\text{location}} \cdot S_{\text{reasoning}}} \cdot P_{\text{unrel}}(n).$$

D Reward Weight Analysis

We further analyze the choice of reward weights through rollout statistics and post-hoc reweighting. For *Reason* and *Verify*, the evidence is specified, so we disable $R_{\text{alignment}}$ and use $(\omega_1, \omega_2, \omega_3) = (0.6, 0, 0.4)$, keeping reasoning completeness as the primary signal while retaining precision as a constraint against unsupported error claims. For *Scan*, both issue and evidence cues are absent, so we use $(\omega_1, \omega_2, \omega_3) = (0.4, 0.4, 0.2)$, assigning symmetric primary weights to reasoning completeness and evidence alignment while using precision as an auxiliary constraint.

Table 17 reports the standard deviation of reward components over rollout steps. For *Reason* and *Verify*, $R_{\text{completeness}}$ consistently shows larger variance than $R_{\text{precision}}$, suggesting that it better differentiates reasoning quality across sampled trajectories. For *Scan*, $R_{\text{completeness}}$ and $R_{\text{alignment}}$ have comparable variability, while $R_{\text{precision}}$ is more stable. This supports treating completeness and alignment as the main objectives in *Scan*, with precision used to suppress excessive error reporting.

Table 12: Sub-metric scores on the *Reason* task (scaled by 100).

Models	Avg.	QI	DI	IC	PD	RQD	SG
Reason $R_{\text{completeness}}$							
GPT-5.4	71.1	82.1	62.8	58.4	78.6	72.0	71.9
Gemini 3 Pro	59.2	64.7	55.3	63.0	57.0	65.2	49.9
Seed-1.6-Thinking	52.4	58.4	49.8	52.0	58.4	59.6	36.3
Qwen3-VL-Plus	60.7	68.8	52.3	68.0	60.8	64.8	49.6
Qwen3-VL-235B-A22B (Thinking)	54.7	67.1	42.7	64.0	58.1	55.2	40.9
Qwen3-VL-235B-A22B (Instruct)	46.8	52.4	36.1	52.4	49.5	49.9	40.4
Qwen3-VL-32B	52.2	63.7	45.4	46.2	61.6	53.5	42.8
Qwen3-VL-8B (Instruct)	23.3	25.3	17.3	27.2	26.6	27.3	15.9
Qwen3-VL-8B (SFT)	13.9	16.1	13.6	4.7	20.6	19.5	8.9
Qwen3-VL-8B (SFT+RL)	17.6	12.7	9.8	17.3	24.2	24.0	17.6
Reason $R_{\text{precision}}$							
GPT-5.4	34.4	40.8	28.3	26.0	44.7	43.9	24.0
Gemini 3 Pro	61.2	63.1	57.6	58.0	65.0	67.1	56.6
Seed-1.6-Thinking	49.8	55.6	44.3	48.2	50.7	56.5	43.7
Qwen3-VL-Plus	53.3	58.1	50.8	50.9	62.8	53.0	44.1
Qwen3-VL-235B-A22B (Thinking)	47.4	58.2	39.7	41.4	53.5	51.5	39.9
Qwen3-VL-235B-A22B (Instruct)	44.5	55.2	42.2	39.1	49.6	44.7	36.3
Qwen3-VL-32B	44.8	54.3	40.2	38.2	49.4	47.3	39.4
Qwen3-VL-8B (Instruct)	38.8	40.9	38.2	42.1	46.8	38.1	26.8
Qwen3-VL-8B (SFT)	54.9	53.8	53.6	51.6	58.4	58.6	53.3
Qwen3-VL-8B (SFT+RL)	69.9	68.1	66.3	71.3	71.7	71.9	69.9
Reason R_{final}							
GPT-5.4	56.4	65.6	49.0	45.4	65.0	60.7	52.7
Gemini 3 Pro	60.0	64.0	56.2	61.0	60.2	66.0	52.6
Seed-1.6-Thinking	51.4	57.3	47.6	50.5	55.3	58.4	39.3
Qwen3-VL-Plus	57.7	64.5	51.7	61.2	61.6	60.1	47.4
Qwen3-VL-235B-A22B (Thinking)	51.8	63.5	41.5	55.0	56.3	53.7	40.5
Qwen3-VL-235B-A22B (Instruct)	45.9	53.5	38.5	47.1	49.5	47.8	38.8
Qwen3-VL-32B	49.2	59.9	43.3	43.0	56.7	51.0	41.5
Qwen3-VL-8B (Instruct)	29.5	31.5	25.7	33.2	34.7	31.6	20.3
Qwen3-VL-8B (SFT)	30.3	31.2	29.6	23.5	35.7	35.1	26.6
Qwen3-VL-8B (SFT+RL)	38.5	34.8	32.4	38.9	43.2	43.2	38.5

We also perform post-hoc reweighting of R_{final} without retraining, as shown in Table 18. The default setting corresponds to the weights used in the main experiments. The completeness-heavy setting increases the weight of $R_{\text{completeness}}$ to $(0.7, 0, 0.3)$ for *Reason*/*Verify* and $(0.5, 0.3, 0.2)$ for *Scan*. We additionally test an equal-weight setting for *Reason*/*Verify*, using $(0.5, 0, 0.5)$. Across these variants, the main conclusions remain stable: the trained Qwen3-VL-8B model consistently improves over its base and SFT variants, and the relative difficulty of *Scan* remains unchanged.

These analyses do not replace full retraining under alternative rewards, but they show that the reported trends are not an artifact of one exact coefficient choice. The selected weights follow the cue structure of the tasks: evidence alignment is disabled when evidence is already specified, becomes central when evidence must be discovered, and precision remains a constraint against unsupported error reporting.

E Case Study

Figures 8, 9, 10, 11 and 12 provide qualitative examples showing how VERA-RL improves Qwen3-VL-8B-Instruct on Scan-style scientific error detection. Across the three cases, the original instruction-tuned model tends to give broad no-error judgments based on surface-level checks, such as whether metrics, tables, or experimental descriptions appear generally consistent. After VERA-RL, the model more often identifies the specific evidence needed for verification and connects it to the corresponding error. In the quantitative inconsistency case, it compares numerical claims in the abstract with table values and detects inconsistent performance gains.

Table 13: Sub-metric scores on the *Verify* task (scaled by 100).

Models	**Avg.**	**QI**	**DI**	**IC**	**PD**	**RQD**	**SG**
		Verify $R_{\text{completeness}}$					
GPT-5.4	62.3	74.7	51.2	67.6	65.5	71.0	49.7
Gemini 3 Pro	52.3	68.8	40.5	46.3	62.4	55.0	40.5
Seed-1.6-Thinking	32.8	37.3	22.0	27.2	39.9	42.2	28.4
Qwen3-VL-Plus	46.4	64.2	36.7	43.0	52.1	46.9	35.7
Qwen3-VL-235B-A22B (Thinking)	41.0	57.6	15.8	42.0	50.3	47.5	32.5
Qwen3-VL-235B-A22B (Instruct)	13.0	14.2	9.0	16.0	16.2	13.0	9.3
Qwen3-VL-32B	40.7	53.6	29.5	49.4	41.0	42.2	28.7
Qwen3-VL-8B (Instruct)	3.9	4.3	5.5	5.0	3.3	3.7	1.5
Qwen3-VL-8B (SFT)	12.6	11.8	5.5	15.5	9.8	19.7	13.1
Qwen3-VL-8B (SFT+RL)	17.2	14.9	9.2	20.3	22.9	18.3	17.7
		Verify $R_{\text{precision}}$					
GPT-5.4	46.1	60.4	36.2	46.9	51.8	45.4	39.2
Gemini 3 Pro	57.8	64.2	46.1	54.7	69.2	57.8	55.0
Seed-1.6-Thinking	79.4	77.5	82.4	78.1	79.9	80.4	77.9
Qwen3-VL-Plus	70.4	70.6	65.7	73.6	74.1	73.0	65.2
Qwen3-VL-235B-A22B (Thinking)	55.5	70.2	33.2	49.9	62.3	65.3	52.2
Qwen3-VL-235B-A22B (Instruct)	16.1	19.9	7.6	21.1	22.2	13.6	12.2
Qwen3-VL-32B	66.8	69.1	61.3	65.6	68.0	72.2	64.7
Qwen3-VL-8B (Instruct)	8.5	14.4	9.1	10.9	7.0	4.2	5.3
Qwen3-VL-8B (SFT)	52.8	50.2	49.0	55.3	54.5	57.1	50.4
Qwen3-VL-8B (SFT+RL)	68.5	67.5	66.9	70.7	72.4	67.0	66.6
		Verify R_{final}					
GPT-5.4	55.8	69.0	45.2	59.3	60.0	60.7	45.5
Gemini 3 Pro	54.5	67.0	42.7	50.0	65.1	56.1	46.3
Seed-1.6-Thinking	51.4	53.4	46.2	47.5	55.9	57.5	48.2
Qwen3-VL-Plus	56.0	66.8	48.3	55.2	60.9	57.4	47.5
Qwen3-VL-235B-A22B (Thinking)	46.8	62.7	22.8	45.2	55.1	54.5	40.3
Qwen3-VL-235B-A22B (Instruct)	14.2	16.5	8.4	18.0	18.6	13.2	10.5
Qwen3-VL-32B	51.2	59.8	42.2	55.9	51.8	54.2	43.1
Qwen3-VL-8B (Instruct)	5.7	8.4	6.9	7.4	4.8	3.9	3.0
Qwen3-VL-8B (SFT)	28.6	27.2	22.9	31.4	27.7	34.7	28.0
Qwen3-VL-8B (SFT+RL)	37.7	35.9	32.3	40.5	42.7	37.8	37.3

Table 14: Additional external evaluation results.

Benchmark	Instruct	SFT+RL
MMLongBench-Doc	22.9	23.3
MMMU	44.3	46.1
PRISMM-Bench	51.6	52.0

Table 15: Performance on the full and paper-disjoint test sets. All scores are R_{final} and scaled by 100.

Model	Split	*Reason*	*Verify*	*Scan*
Instruct	Full	29.5	5.7	2.0
	Disjoint	30.2	5.2	1.6
SFT	Full	30.3	28.6	14.4
	Disjoint	31.3	28.6	14.3
SFT+RL	Full	38.5	37.7	19.5
	Disjoint	39.3	39.1	19.3

In the pipeline-distortion case, it traces the retrieval procedure across the main text, equation, algorithm, and appendix, identifying that the method starts from answer entities and therefore introduces information leakage. In the inference-and-conclusions case, it compares ablation results with the textual interpretation and detects a contradiction between the reported score and the claimed necessity of the current-state information. These examples illustrate that VERA-RL improves not only response format, but also evidence selection, cross-section consistency checking, and the ability to turn paper-level evidence into verifiable error judgments.

Table 16: Performance of the intermediate SFT checkpoint at step 200, compared with the original base model and the 1-epoch SFT model. Scores are scaled by 100.

Metric	**0 step**	**200 steps**	**1 epoch**
Scan $R_{\text{completeness}}$	1.5	1.6	5.3
Scan $R_{\text{alignment}}$	1.0	2.4	2.5
Scan $R_{\text{precision}}$	5.0	61.4	56.3
Scan R_{final}	2.0	13.8	14.4
Verify R_{final}	5.7	29.2	28.6
Reason R_{final}	29.5	31.1	30.3

F Reliability Analysis

Cross-Evaluator Agreement. We independently rescore the step-30 RL rollouts using Qwen3-27B and Gemini 2.5 Flash. As shown in Table 19, both evaluators show high correlation with the original scores across *Reason*, *Verify*, and *Scan*, suggesting that the reward signals are not specific to a single evaluator.

G Reproducibility and Ethics Statement

VERA-13K is constructed by the authors and does not reuse samples from previously released benchmarks. For papers from international conferences such as ICML, ICLR (including reviews), and NeurIPS, all content was crawled from the OpenReview platform. Papers from Nature Communications and PNAS were obtained entirely from open-access sources, ensuring that no privacy, ethical, or conflict-of-interest concerns arise.

The code and data are publicly available at https://github.com/Staudinger0325/VERA-RL.

Table 17: Standard deviation of reward components over rollout steps.

Step	*Reason*		*Verify*		*Scan*		
	$\sigma(R_{\text{com}})$	$\sigma(R_{\text{prec}})$	$\sigma(R_{\text{com}})$	$\sigma(R_{\text{prec}})$	$\sigma(R_{\text{com}})$	$\sigma(R_{\text{align}})$	$\sigma(R_{\text{prec}})$
1	0.36	0.16	0.42	0.21	0.34	0.28	0.19
5	0.24	0.14	0.34	0.20	0.28	0.23	0.17
10	0.29	0.21	0.26	0.19	0.20	0.15	0.13
15	0.32	0.17	0.32	0.18	0.28	0.24	0.14
20	0.35	0.17	0.37	0.23	0.29	0.29	0.14
25	0.28	0.13	0.32	0.18	0.30	0.25	0.16
30	0.34	0.20	0.41	0.19	0.30	0.26	0.16

Table 18: Post-hoc reward reweighting results. Scores are scaled by 100. Default weights are $(0.6, 0, 0.4)$ for *Reason*/*Verify* and $(0.4, 0.4, 0.2)$ for *Scan*. Completeness-heavy weights are $(0.7, 0, 0.3)$ for *Reason*/*Verify* and $(0.5, 0.3, 0.2)$ for *Scan*. Equal R/V uses $(0.5, 0, 0.5)$ for *Reason*/*Verify*.

Model	**Default**			**Completeness-heavy**			**Equal R/V**	
	Reason	**Verify**	**Scan**	**Reason**	**Verify**	**Scan**	**Reason**	**Verify**
Gemini 3 Pro	60.0	54.5	24.3	59.8	54.0	24.8	60.2	55.1
Qwen3-VL-235B-A22B Thinking	51.8	46.8	17.4	52.5	45.4	17.8	51.0	48.3
Qwen3-VL-235B-A22B Instruct	45.9	14.2	3.4	46.1	13.9	3.5	45.7	14.6
Qwen3-VL-8B Instruct	29.5	5.7	2.0	28.0	5.3	2.1	31.1	6.2
Qwen3-VL-8B SFT	30.3	28.6	14.4	26.2	24.7	14.7	34.4	32.7
Qwen3-VL-8B SFT+RL	38.5	37.7	19.5	33.3	32.6	19.7	43.8	42.9

Table 19: Correlation (%) between reward scores extracted by alternative evaluators and those from Seed-1.6-Thinking on the step-30 RL rollouts.

Setup	Reason (R_{final})	Verify (R_{final})	Scan (R_{final})
Qwen3-27B	78.3	88.4	88.7
Gemini 2.5 Flash	84.4	82.3	82.2

Case 1

ID: 1319
Error Category: Inference & Conclusions
Paper ID: ICLR_accept_01439_modified_2_10325
Question: Assess whether Section 5.2 contains any errors of the category Inference & Conclusions.

Gold Answer:
Evidence:
- Section 5.2 (page 10): Text claims that 'retaining this information is essential' but also claims 'removing current state information... improves performance'.
- Table 3 (page 10): 'w/o current state' variant has a score of '89.55' while 'Base Diffusion Planner' has a score of '89.19'.
Reasoning:
- The paper reports a higher performance score for a model variant (89.55) than the baseline configuration (89.19).
- Section 5.2 contains contradictory interpretations, suggesting information is both essential to keep and beneficial to remove.
- These conflicting statements lead to an invalid inference regarding the necessity of ego vehicle state variables.
Answer: ERROR. Section 5.2 contains mutually exclusive conclusions and an invalid inference because the text claims removing information both improves performance and is essential to retain, contradicting the data in Table 3.

Response from Qwen3-VL-8B-Instruct (R_{final}=0.0):
Based on a thorough review of Section 5.2 ("Ablation Studies") in the provided paper, there are no errors of the category "Inference & Conclusions".
The section clearly states its purpose: to demonstrate the effectiveness of key components through ablation.
Each claim is directly supported by Table 3 and Figures 6 and 7. For example, the drop from 89.19 to 85.02 supports the claim about z-score normalization; the drop to 83.78 supports the claim about trajectory interpolation.
The interpretations drawn from the ablation results are reasonable. The conclusion that retaining velocity/acceleration/yaw rate may lead to shortcuts is consistent with the lower scores for `w/ ego state` and `w/SDE`, while `w/o current state` improves performance. The section acknowledges nuance.
The authors do not overreach beyond their framework. There is no internal contradiction within Section 5.2 or with other parts of the paper.
Therefore, Section 5.2 does not contain any errors of the category "Inference & Conclusions".

Response after VERA-RL (R_{final}=1.0):
Reasoning:
- In Section 5.2, a contradiction is identified regarding the effect of removing the ego vehicle's current state.
- Table 3 provides statistical evidence showing the 'w/o current state' variant (89.55) outperforms the standard model (89.19).
- The text explicitly claims that retaining velocity and acceleration leads to increasing planning capabilities.
- This contradictory relationship between data and text affects the validity of the "Design Choices for Inference" claims in Section 6.
Answer: ERROR. Section 5.2 contains contradictory information about the importance of ego vehicle state and performance claims.

Analysis:
After RL, the model becomes better at detecting contradictions between experimental results and textual conclusions. In Section 5.2, the Instruct model overlooks that the “w/o current state” variant scores 89.55, higher than the 89.19 baseline, whereas the RL model correctly identifies this inconsistency and cites the result as evidence.

Figure 8: Case study on inference and conclusions.

Case 1

ID: 1208
Error Category: Pipeline Distortion
Paper ID: ICLR_accept_01038_modified_2_9607
Question: Assess whether the whole paper contains any errors of the category Pipeline Distortion.

Gold Answer:
Evidence:
- Section 4.4 (page 6): The text claims the retrieval process finds paths that "start from the answer entities ea".
- Section A.3 (page 16): States the queue is initialized with "entities in the answer Aq".
- Algorithm 1 (page 17): Explicitly seeds the search with "ea in Aq".
- Page 3: Defines the KGQA task goal as predicting answers a from question q.
- Figure 2: Illustrates an operational flow intended to deduce the answer.

Reasoning:
- The paper describes a retrieval module that requires the answer entities as input to start the search process.
- The stated goal of the task is to predict these very answers from the question.
- This creates a data handling contradiction where the retrieval process assumes the outcome it is supposed to produce.
- This flaw breaks the link between the question and the reasoning process, leading to information leakage.

Answer: ERROR. The paper contains a systematic data handling error where the retrieval and reasoning modules assume knowledge of the answer entities to find paths, contradicting the task goal of predicting those answers.

Response from Qwen3-VL-8B-Instruct (R_{final}=0.0):
After a thorough review of the paper "Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning," no errors in the category of Pipeline Distortion were found.
The paper presents a well-structured experimental evaluation with clear metrics (Hits@1, F1), appropriate baselines, and detailed ablation studies. All reported results are consistent with the described methodology and datasets. The authors provide comprehensive appendices detailing dataset statistics, training configurations, and additional experiments, ensuring reproducibility and transparency in data handling.
Key observations supporting this assessment:
1. Consistent Metrics: The use of Hits@1 and F1 is standard for KGQA tasks and is applied consistently across all experiments and comparisons.
2. Clear Dataset Description: Appendix A.4 provides detailed statistics for WebQSP, CWQ, and MetaQA-3hop, including train/test splits, max hops, and answer distributions, which are correctly referenced in the main text.
3. Accurate Reporting: Results in tables (e.g., Table 1, Table 2, Table 3) are internally consistent and align with textual claims (e.g., improvements over SOTA methods).
4. Reproducible Setup: Implementation details (Appendix A.6) specify model backbone, training epochs, batch size, learning rate, and hardware, allowing for replication.
5. Appropriate Analysis: Figures (e.g., Figure 3, Figure 4) accurately represent the relationship between K (top-K relation paths) and retrieval cost/faithfulness, supporting the choice of K=3.
6. No Inconsistent or Unexplained Data: There are no instances of unexplained outliers, contradictory results, or mislabeled axes/legends in figures or tables.
Therefore, the paper demonstrates rigorous measurement and data handling practices throughout.
Conclusion: The paper contains no errors in the category of Pipeline Distortion.

Figure 9: Case study on the first part of pipeline distortion.

Response after VERA-RL (R_{final}=0.678):
Evidence:
- Section 4.4 (page 6): 'Retrieval' by 'finding paths starting from the answer entities'.
- Equation 9 on page 6: W_z = { w_z(e_a, e^*)| ... }.
- Algorithm 1 (page 12): Line 5 initializes the queue with answer entities (e_a).
- Appendix A.3 (page 12): Defines reasoning paths using answer entities.

Reasoning:
- The paper constructs the retrieval process to pick an arbitrary set of entities from answer entities by design.
- This initialization is contradicting with the planning rationale where paths should lead from a question entity.
- The methodology in places 5, 9, and 12 mandates that any entity within the question-hopping horizon leading to the answer may be used as a start or end point.

Answer: ERROR. Section 4.4, Equation 9, Algorithm 1, and Appendix A.3 contradict the planning rationale by initiating the graph retrieval from answer entities rather than starting from question entities.

Analysis:
After VERA-RL, the model shifts from surface-level experimental checking to mechanism-level error identification. In this case, the gold error is that the KGQA task should predict answers from the question, but the paper's retrieval module initializes path search from answer entities, meaning that the model uses the target answer in advance and introduces information leakage. The Instruct model only checks surface-level issues such as Hits@1, F1, dataset statistics, result tables, and reproducibility details, and therefore concludes that there is no Pipeline Distortion error. In contrast, the RL model explicitly points out that Section 4.4, Equation 9, Algorithm 1, and Appendix A.3 all use answer entities as the starting point for retrieval or path construction, and further notes that this contradicts the KGQA logic of deriving answers from question entities. Through VERA-RL, the model reduces irrelevant responses and identifies hidden information leakage in the method pipeline.

Figure 10: Case study on the second part of pipeline distortion.

Case 1

ID: 2235
Error Category: Quantitative Inconsistency
Paper ID: ICML_2025_00622_modified_1_15940
Question: Assess whether the whole paper contains any errors of the category Quantitative Inconsistency.

Gold Answer:
Evidence:
- Abstract (page 1): Claims the model outperforms by 15% and 8% on CUB200 and Flowers102.
- Section 4 (page 4), fourth paragraph: Lists gains as 1%, 8%, and 15% on Flowers102, CUB200, and Food101.
- Table 1 (page 4): Shows 'S. KML Mixup' (75.20%) vs 'Places' (59.51%) on Flowers (15.69 points) and 'S. KML Mixup' (27.08%) vs 'Places' (19.09%) on CUB (7.99 points).

Reasoning:
- The paper presents conflicting numerical data across different sections regarding classification gains.
- Calculations based on Table 1 show a 15.69 point gain for Flowers and a 7.99 point gain for CUB.
- The Abstract incorrectly swaps these values, attributing the 15% gain to CUB and 8% to Flowers.
- Section 4 provides a third set of values (1%, 8%, and 15%) that contradicts both Table 1 and the Abstract.

Answer: ERROR. The paper contains numerical inconsistencies and calculation errors regarding classification gains across the Abstract, Section 4, and Table 1.

Response from Qwen3-VL-8B-Instruct (R_{final}=0.0):
After a thorough review of the entire paper, including all text, figures, tables, and appendices, no errors of the category "Quantitative Inconsistency" were found.
The paper presents its methodology, results, and analysis clearly and consistently. All reported metrics (accuracy, R^2, percentages) are internally consistent with the described experiments and visualizations. The mathematical notation used (e.g., for DINO objective, KNN, PCA, differential privacy definition) is standard and correctly applied within context. The formulas and computational procedures described (e.g., Shaders KML Mixup process, z-test in Appendix B.1) are logically sound and align with the presented data.
There are no instances of:
- Incorrect arithmetic or misreported numbers.
- Misuse of mathematical symbols or operators.
- Flawed derivations or unjustified formulaic assumptions.
- Inconsistencies between stated methods and reported results.

Therefore, the paper is free from Quantitative Inconsistency errors.

Figure 11: Case study on the first part of quantitative inconsistency.

Response after VERA-RL (R_{final}=0.754):
Evidence:
- Abstract (page 1): Claims a 'gap of only 10%' between procedural and real-world models and states the procedural model beats the Places model by 15% on CUB200 and 8% on Flowers102.
- Table 1 (column 2, page 5): Shows the researcher's own 'S. KML Mixup' fine-grained results are 75.20/27.08/48.70.
- Table 1 (column 2, page 5): Shows the 'Places' model results are 59.51/19.09/47.78 for CUB200, Flowers102, and Food101.
Reasoning:
- The claim of a 10% 'performance gap' and relative increases are from the Abstract.
- The reported 'difference' values in the Abstract's claims are incompatible with the reported ratios and absolute scores in Table 1.
- The 'Flowers102' value in the Abstract is 8%, whereas the methodology in the Table shows a difference of only 0.2%; this difference changes the accuracy range as defined by the authors.
Answer: ERROR. The Abstract and Table 1 contain contradictory arithmetic definitions for performance differences regarding the procedural models' accuracy gap and relative increases in performance.

Analysis:
After VERA-RL, the model better identifies numerical inconsistencies in the paper. The gold error is that the abstract claims 15% and 8% improvements on CUB200 and Flowers102, respectively, while the actual values in Table 1 show an improvement of about 15.69 points on Flowers102 and 7.99 points on CUB200, meaning the two gains are mixed up in the abstract. The Instruct model only gives a generic judgment that the formulas, metrics, and tables are consistent and finds no error; the RL model instead cites the specific numbers from the abstract and Table 1, and points out that the reported performance gains are inconsistent with the table data. VERA-RL improves the model’ s sensitivity to arithmetic contradictions between numerical claims and tabular results within the paper.

Figure 12: Case study on the second part of quantitative inconsistency.