ANABENCH For Enhancing Scientific Table & Figure Analysis

Xuehang Guo[1] **Zhiyong Lu**[2] **Tom Hope**[3] **Qingyun Wang**[1]
[1]William & Mary [2]National Library of Medicine [3]The Allen Institute for AI

Abstract

In scientific research, analysis requires accurately interpreting complex multimodal knowledge, integrating evidence from different sources, and drawing inferences grounded in domain-specific knowledge. However, current artificial intelligence (AI) systems struggle to consistently demonstrate such capabilities. The complexity and variability of scientific tables and figures, combined with heterogeneous structures and long-context requirements, pose fundamental obstacles to scientific table & figure analysis. To quantify these challenges, we introduce ANABENCH, a large-scale benchmark featuring 63,178 instances from nine scientific domains, systematically categorized along seven complexity dimensions. To tackle these challenges, we propose ANAGENT, a multi-agent framework for enhanced scientific table & figure analysis through four specialized agents: PLANNER decomposes tasks into actionable subtasks, EXPERT retrieves task-specific information through targeted tool execution, SOLVER synthesizes information to generate coherent analysis, and CRITIC performs iterative refinement through five-dimensional quality assessment. We further develop modular training strategies that leverage supervised finetuning and specialized reinforcement learning to optimize individual capabilities while maintaining effective collaboration. Comprehensive evaluation across 170 subdomains demonstrates that ANAGENT achieves substantial improvements, up to ↑ 13.43% in training-free settings and ↑ 42.12% with finetuning, while revealing that task-oriented reasoning and context-aware problem-solving are essential for high-quality scientific table & figure analysis.

Correspondence to: Xuehang Guo <xguo15@wm.edu>, Zhiyong Lu <zhiyong.lu@nih.gov>, Tom Hope <tomh@allenai.org>, Qingyun Wang <qwang16@wm.edu>.

1 Introduction

AI has made notable progress in assisting scientifists across diverse domains (Boiko et al., 2023; Gao et al., 2024; Wang et al., 2026) and stages of the research lifecycle, such as hypothesis discovery (Wang et al., 2024; Garikaparthi et al., 2025), literature review (Zhang et al., 2024b), citation recommendation (Choi et al., 2025; Press et al., 2024), etc. With the growing trend of *human-AI co-discovery* (Gottweis et al., 2025), these advances reveal AI's potential in serving as AI co-scientists to accelerate scientific discovery and improve research communication (Gridach et al., 2025; Zhang et al., 2024a). However, to function effectively as AI co-scientists, AI systems draw on capabilities in *multimodal reasoning* (Bai et al., 2025c; Zhao et al., 2025), *long-context comprehension* (Reddy and Shojaee, 2024; Sundar et al., 2024), and *domain-specific understanding*, which remain challenging for current AI systems (Zhou et al., 2025).

A fundamental yet important task that reflects these capabilities is **scientific table & figure analysis**, as tables and figures provide critical information that is often difficult to express through text alone in scientific papers. Analyzing these artifacts requires AI systems to accurately: (1) interpret complex multimodal data across diverse layouts and formats (*e.g.*, `LaTeX` tables, bar charts, architectural diagrams), (2) integrate evidence from multiple sources and lengthy contexts (*e.g.*, captions, sections, citations), and (3) generate task-oriented insights grounded in specialized terminology, related contexts, and domain-specific knowledge. Despite recent advances in multimodal large language models (MLLMs), scientific table & figure analysis remains challenging, particularly when handling the ***heterogeneity*** of scientific literature across different *authoring formats* (*e.g.*, `LaTeX`, `XML`), *rendered formats* (*e.g.*, PDF, HTML), and *dissemination platforms* (*e.g.*, arXiv (arXiv, 1991), PubMed (PubMed,

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 33199–33269
October 24-29, 2026 ©2026 Association for Computational Linguistics

1996)). This is further complicated by *error propagation* (Gridach et al., 2025), as mistakes in structural parsing, numerical extraction, or contextual interpretation can cascade into factual incorrectness.

Where existing benchmarks fall short? Several benchmarks have been proposed for scientific table & figure understanding (Li et al., 2024; Singh et al., 2024; Lu et al., 2023; Zhang et al., 2025; Pramanick et al., 2024; Lou et al., 2023; Jin et al., 2019; Liu et al., 2026). However, those benchmarks primarily focus on narrowly defined tasks, such as *question answering*, *claim verification*, or *caption generation*. As such, they fail to capture the full spectrum of challenges inherent in scientific analysis writing (Tab. 4), including varying levels of analytical depth, diverse reasoning requirements across scientific domains, and synthesis of information across multiple modalities and long contexts (Fig. 1). Moreover, our preliminary exploration reveals that current MLLMs struggle significantly with scientific analysis (§2.2). These limitations are particularly pronounced for scientific analysis tasks requiring complex reasoning across different scopes, depths, and objectives (Fig. 6).

Our Approach. To tackle these challenges, we introduce ANABENCH, a scientific table & figure analysis benchmark encompassing tables and figures from 9 scientific domains across 170 fine-grained disciplines, systematically categorized along seven complexity dimensions that capture multifaceted challenges of scientific analysis (Fig. 1). Building on insights from how human researchers approach scientific writing (Fig. 9), we propose ANAGENT (Fig. 3), a multi-agent framework that decomposes scientific analysis into specialized subtasks handled by four collaborative agents: PLANNER for *task decomposition and planning*, EXPERT for *knowledge searching and retrieval*, SOLVER for *reasoning and generation*, and CRITIC for *reflection and refinement*. To enhance agent-wise performance on their specialized tasks, we implement test-time optimization (§3.3) and modular finetuning (§3.4) to enhance individual agent capabilities while maintaining effective collaboration. Our main contributions are:

- We introduce ANABENCH (§2), a benchmark consisting of 63, 178 instances for evaluating and advancing AI systems in scientific tale & figure analysis, spanning seven data and analysis complexities (Fig. 1).

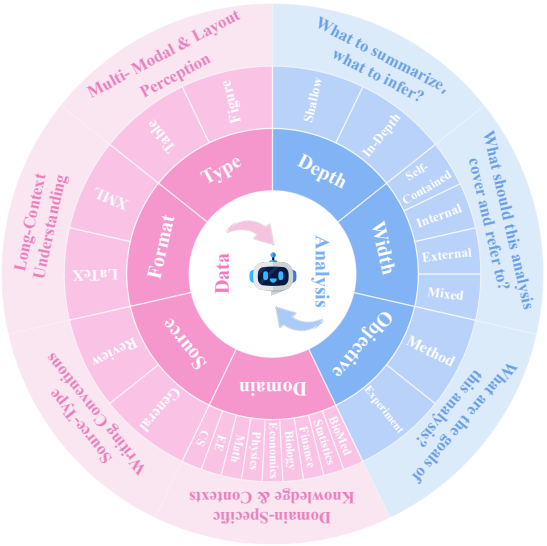


Figure 1: **Challenges In Scientific Table & Figure Analysis.** The heterogeneity of scientific literature presents great challenges for high-quality analysis of scientific tables and figures (Fig. 6).

- We propose ANAGENT (§3), a multi-agent framework for scientific table & figure analysis writing, comprising four specialized agents equipped with specialized tools, enabling complex reasoning, systematic knowledge integration, and collaborative scientific analysis writing.
- By presenting specialized evaluation metrics for assessing scientific analysis quality (§2.4), results in §5 demonstrate that ANAGENT significantly improves scientific table & figure analysis through test-time optimization ($\Delta_{rel} \geq \uparrow 13.43\%$) and modular training ($\Delta_{rel} \geq \uparrow 42.12\%$).

2 ANABENCH: Evaluating Scientific Analysis

2.1 Problem Formulation: Scientific Analysis

We formulate the task of *scientific table & figure analysis* as a context-aware generation problem: Given an input x comprising one or more tables $\{x_t\}$ and/or one or more figures $\{x_f\}$, together with their source information s and input query q that specifies the analysis requirements and objectives, the goal is to generate a well-written analysis y that accurately interprets the provided tabular and visual data, integrates evidence across all available contexts, situates findings within the broader research, and delivers domain-specific insights. For-

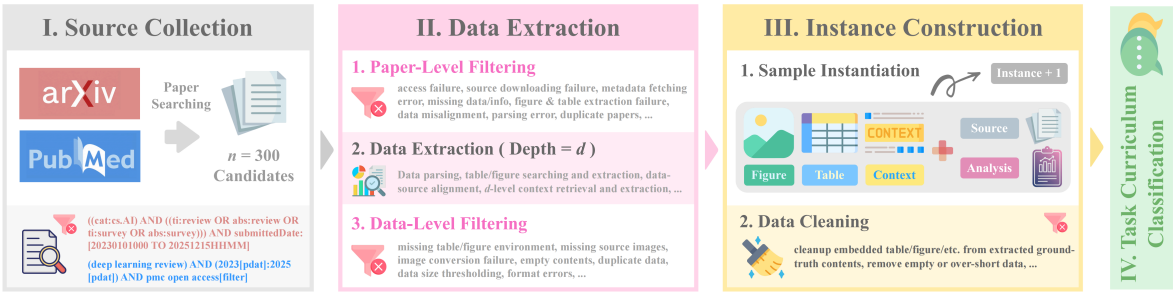


Figure 2: **AnaBench For Evaluating Autonomous Scientific Analysis.** We implement four-stage benchmark construction method to build AnaBench, with multi-level filtering to enhance data quality.

mally, the task can be expressed as:

$$y = f(x, s, q) \quad \text{where } x \in \{\{x_t\}, \{x_f\}, \{x_t, x_f\}\} \tag{1}$$

As such, this scientific analysis writing problem encompasses ***multimodal long-context reasoning*** for input tables and figures with different formats and layouts.

2.2 Preliminary: MLLM Agents In Scientific Analysis

To empirically assess the challenges faced by MLLM agents in scientific analysis, we conduct a preliminary study (§A) evaluating their performance across seven complexity dimensions (Fig. 1). We randomly select 120 samples from AnaBench with all seven challenges evenly distributed, and employ `Qwen3-VL-8B` as MLLM agent backbone to generate scientific analysis. Performance is evaluated using SciBERT (Eq. 23). With performance struggling to exceed 60% across all metrics, Fig. 6 reveals pronounced difficulties in multimodal, multi-layout understanding and in-depth analysis that demand inferential generation. These findings highlight that *MLLM agents face substantial challenges in interpreting complex heterogeneous scientific artifacts*.

2.3 Benchmarking Scientific Analysis

Benchmark Construction. Our benchmark construction method comprises four stages (Fig. 2): (1) *Source collection*, which identifies and collects candidate source papers that satisfy predefined relevance and retrieval criteria. (2) *Data extraction*, which extracts tables, figures, and their associated contexts. A context retrieval depth d controls the level of context referenced by each table or figure. Extracted data are augmented via two-level filtering: *paper-level filtering* removes papers that fail to meet validity requirements, and *data-level filtering* excludes tables and figures with formatting errors, missing information, or other quality issues. (3) *Instance construction*, which transforms each filtered data into a scientific analysis instance. Each instance consists of table and/or figure data, corresponding contexts, metadata, and the gold analysis. Resulting instances are further refined through a specialized *data cleaning* step using configurable thresholds, including the maximum number of samples and the minimum length of ground truths. (4) *MLLM-assisted task classification*, which combines rule-based heuristics with MLLM classification (§C.2.3) to categorize AnaBench along seven dimensions (§C.2). Through four-stage construction, AnaBench achieves large-scale coverage across seven complexity dimensions while faithfully reflecting real-world distributions of data characteristics and analytical challenges.

Data Complexity. We consider four data complexity dimensions (§C.2.1): (1) ***Type***: the type of analysis data (*table*, *figure*, or *both*); (2) ***Domain***: domain disciplines that the source paper belongs to, with AnaBench spanning 9 broad domains across 170 disciplines; (3) ***Format***: the format of analysis writing (`LaTeX` or `XML`); (4) ***Source***: the type of source paper (*general* research papers or *reviews & surveys*).

Analysis Complexity. We characterize analysis complexity along three complementary dimensions (§C.2.2): (1) ***Width***: the reference scope of the analysis (*self-contained*, *internal*, *external*, or *mixed*); (2) ***Depth***: the level of analytical rigor (*shallow* or *in-depth*); (3) ***Objective***: the primary goal and focus of the analysis (*methodology* or *experiment*).

2.4 Evaluating Scientific Analysis

Rule-Based Evaluation. Rule-based evaluation metrics cover both lexical and semantic assess-

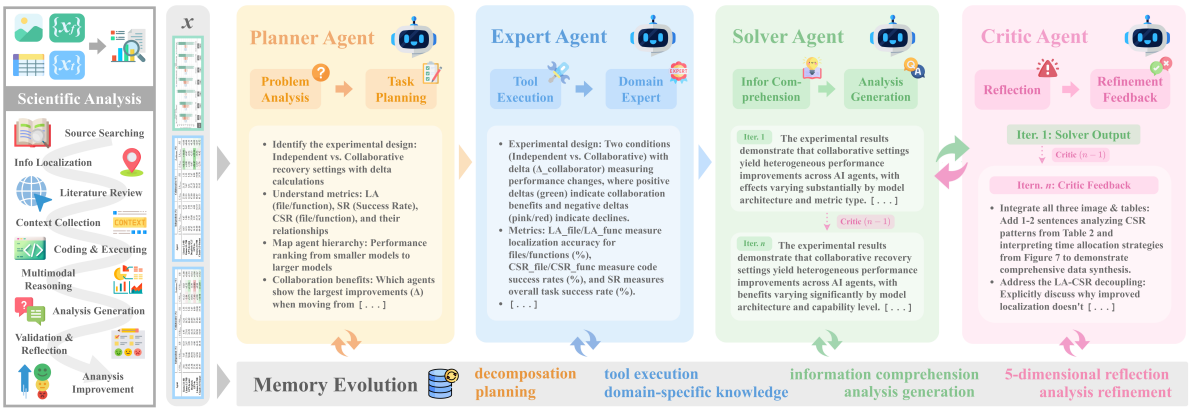


Figure 3: **Multi-Agent Coordinative Scientific Analysis.** Our multi-Agent scientific analysis framework, **ANAGENT**, is developed to cover various stages to analyze scientific tables and figures through four collaborative agents: **PLANNER**, **EXPERT**, **SOLVER**, and **CRITIC**. Some example details are omitted as **[...]** for clarity.

ment, where lexical evaluation include ROUGE-L (Eq. 19) (Lin, 2004), BLEU (Eq. 20) (Papineni et al., 2002), and word overlap (Eq. 21); while semantic assessment calculates similarity between model generated analysis y and ground-truth analysis y^* through cosine similarity (Eq. 22), SciBERT (Eq. 23) (Beltagy et al., 2019), and METEOR (Eq. 24) (Banerjee and Lavie, 2005) scores.

MLLM-As-Judge. For more reliable evaluation, we implement MLLM-as-judge by leveraging Gemini-2.5-Flash and GPT-4.1-mini to grade each generated analysis across five dimensions (Fig. 44, §A.2 & E), including *analysis consistency*, *query-analysis alignment*, *knowledge utilization*, *format correctness*, and *grounding accuracy*.

3 ANAGENT: Multi-Agent Scientific Analysis

3.1 ANAGENT For Scientific Table & Figure Analysis

Facing challenges in both data and analysis levels (Fig. 1), traditional approaches that directly map inputs to outputs struggle with varying task complexities due to their lack of systematic reasoning and knowledge retrieval capabilities.

How do human scientists analyze tables and figures? Instead of simply describing what we observe, we engage in a deliberate process of understanding the research question, planning the problem-solving, gathering relevant domain knowledge, interpreting the data in context, and rigorously evaluating our findings and conclusions (Fig. 9).

Our Approach. Inspired by human analysis workflow (Fig. 9), we propose **ANAGENT** (Fig. 3), a multi-agent system for enhanced table & figure analysis. Given input x, source s, and query q, **ANAGENT** operates through four interactive stages (§D):

Stage 1: Task Decomposition. **PLANNER** analyzes the input and decomposes the complex task into actionable subtasks τ_i ($i = 1, \ldots, M_p$):

$$\textbf{PLANNER}(x, s, q) = \{\tau_1, \tau_2, \ldots, \tau_{M_p}\} \quad (2)$$

Stage 2: Task-Oriented Knowledge Retrieval. **EXPERT** performs iterative knowledge acquisition through multi-turn tool executions. At each turn e, the knowledge base $\mathcal{K}_e$ is expanded by incorporating new knowledge retrieved based on subtask τ_e and previously accumulated knowledge $\mathcal{K}_{e-1}$:

$$\mathcal{K}_e = \mathcal{K}_{e-1} \cup \textbf{EXPERT}(\tau_e, \mathcal{K}_{e-1}), \quad e = 1, \ldots, M_e \quad (3)$$

Stage 3: Solution Generation. **SOLVER** synthesizes the accumulated knowledge $\mathcal{K}_n$ with the input to generate candidate analysis. At iteration i, it incorporates feedback f_{i-1}:

$$y_i = \textbf{SOLVER}(x, s, q, \mathcal{K}_n, f_{i-1}), \quad i = 1, \ldots, M_s \quad (4)$$

Stage 4: Reflective Refinement. **CRITIC** assesses generated analysis through five-dimensional evaluation protocol (§E.2) and provides feedback for iterative improvement:

$$f_i = \textbf{CRITIC}(y_i, x, s, q, \mathcal{K}_n), \quad i = 1, \ldots, M_c \quad (5)$$

The interactive refinement between **SOLVER** (Eq. 4) and **CRITIC** (Eq. 5) produces the final analysis $y = y_M$.

Model	Size	Semantic Accuracy (%)			Lexical Accuracy (%)			Overall Accuracy (%)		
		COSINE	BERT	METEOR	ROUGE-L	BLEU	WORD	S_{SEM}	S_{LEX}	S_{AVG}
Baselines										
GPT-4.1-mini	-	**56.34**	**59.74**	**19.47**	16.74	**3.39**	**11.49**	**45.18**	**10.54**	**27.86**
Gemini-2.5-Flash	-	52.41	55.99	19.01	14.90	2.76	9.95	42.47	9.20	25.84
InternVL-3.5	**4B**	54.38	58.19	18.76	15.67	2.66	9.80	43.78	9.37	26.58
	8B	55.73	59.10	19.30	16.80	2.86	10.28	44.71	9.98	27.34
Qwen2.5-VL	**3B**	54.74	58.49	17.82	15.89	2.56	10.02	43.68	9.49	26.59
	7B	55.65	59.66	18.90	16.40	2.98	10.38	44.74	9.98	27.31
Qwen3-VL	**4B**	55.41	58.15	18.41	15.77	2.77	10.06	43.99	9.53	26.76
	8B	55.94	59.11	19.16	**17.06**	3.02	10.39	44.73	10.16	27.44
ANAGENT (Zero-Shot)										
GPT-4.1-mini	-	**59.94**	**61.63**	22.75	**18.19**	4.81	**12.26**	**↑6.49%**	**↑11.48%**	**↑7.43%**
Gemini-2.5-Flash	-	55.60	59.37	19.40	16.04	3.15	11.10	**↑5.46%**	**↑9.67%**	**↑6.19%**
InternVL-3.5	**4B**	58.26	59.86	21.21	16.10	3.29	11.11	**↑6.08%**	**↑8.54%**	**↑6.51%**
	8B	59.46	61.25	22.59	17.00	3.88	11.68	**↑6.84%**	**↑8.72%**	**↑7.21%**
Qwen2.5-VL	**3B**	57.50	60.01	21.03	17.34	3.87	11.53	**↑5.72%**	**↑14.96%**	**↑7.37%**
	7B	58.91	60.41	21.59	17.47	4.11	11.85	**↑4.98%**	**↑11.62%**	**↑6.41%**
Qwen3-VL	**4B**	59.41	60.21	21.23	16.27	3.90	11.33	**↑6.73%**	**↑10.18%**	**↑7.36%**
	8B	59.76	61.53	**23.07**	17.75	**4.98**	12.20	**↑7.58%**	**↑14.57%**	**↑8.89%**
ANAGENT (One-Shot)										
GPT-4.1-mini	-	60.87	**63.28**	24.26	**20.65**	5.73	12.55	**↑9.50%**	**↑23.15%**	**↑12.06%**
Gemini-2.5-Flash	-	**61.06**	61.34	20.52	17.40	4.06	11.47	**↑12.17%**	**↑19.35%**	**↑13.43%**
InternVL-3.5	**4B**	59.11	60.52	22.60	18.04	3.82	11.50	**↑8.29%**	**↑18.68%**	**↑10.12%**
	8B	60.26	62.12	23.18	19.14	4.56	**12.97**	**↑8.52%**	**↑22.44%**	**↑11.08%**
Qwen2.5-VL	**3B**	58.89	60.70	22.19	18.41	3.99	11.54	**↑8.20%**	**↑19.18%**	**↑10.15%**
	7B	60.24	61.00	23.41	19.41	4.98	12.47	**↑7.78%**	**↑23.15%**	**↑10.77%**
Qwen3-VL	**4B**	59.64	60.61	22.42	18.05	4.03	11.51	**↑8.09%**	**↑17.52%**	**↑9.79%**
	8B	60.55	62.27	**24.65**	20.06	**5.92**	12.95	**↑9.88%**	**↑27.76%**	**↑13.23%**

Table 1: **Evaluation of Training-Free Agents.** Comparing between baselines and training-free **ANAGENT** ($M_e = 5$) on **ANABENCH** (§C.4), ***relative differences*** (Eq. 29) are shown as *positive* ↑ $\Delta_{rel}\%$ or *negative* ↓ $\Delta_{rel}\%$.

3.2 Scientific ToolKits

To facilitate complex scientific analysis spanning multiple stages from source searching to analysis writing (Fig. 3), we develop 5 scientific toolkits with 16 specialized tools (Tab. 13) to enable efficient scientific analysis with improved accuracy and comprehensiveness (§D.3).

3.3 Multi-Agent Optimization

Few-Shot Optimization. To enhance the adaptability of individual agents, we include few-shot learning by providing each agent with k-shot exemplars. These examples guide agents to perform specialized tasks effectively, enabling test-time adaptation without extensive task-specific training.

Critic-Guided Reflective Optimization. To further improve collaborative performance, we incorporate a dedicated **CRITIC** that assesses and optimizes **SOLVER**'s analysis solutions. Through five-dimensional protocol (§E.2), **CRITIC** provides targeted feedback to guide **SOLVER** optimizing analysis solution, reducing errors, improving logical consistency, and mitigating hallucinations.

Agent-Level Capability Augmentation. In multi-agent systems, overall performance is significantly influenced by individual agents' capabilities. To this end, we introduce agent-level capability augmentation, a strategy in which individual agents can be independently enhanced by more capable models to improve system-level outcomes, enabling selective upgrades at test time.

3.4 Modular Training

How to train **ANAGENT** *to enhance individual agent capabilities while maintaining effective global collaboration?* We develop a modular training paradigm that aligns with the functional decomposition of **ANAGENT**. Each agent is first initialized via supervised finetuning (SFT) to establish analysis and reasoning foundations, followed by agent-specific reinforcement learning (RL) to optimize specialized behaviors and capabilities.

Supervised Finetuning. All agents in **ANAGENT** are initialized through the SFT phase on the sci-

entific analysis writing training set (Tab. 7) randomly sampled from ANABENCH (§2). Each training instance consists of the multimodal input $x \in \{x_t, x_f\}$, source information s, query q, and the corresponding ground-truth analysis y^*. Let θ denote the shared model parameters. The SFT objective (Eq. 6) is to minimize the token-level negative log-likelihood of the reference analysis conditioned on the input (§3.4):

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \mathbb{E}_{(x,s,q,y^*)} \left[-\sum_{t=1}^{|y^*|} \log p_\theta(y_t^* \mid y_{<t}^*, x, s, q) \right] \tag{6}$$

RL Optimization. After SFT initialization, each agent is further optimized via RL with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each agent $a \in \{\text{PLANNER}, \text{EXPERT}, \text{SOLVER}, \text{CRITIC}\}$, we define an agent-specific policy π_{θ_a} derived from SFT initialization and optimized on specialized RL datasets (§D.4). Given an input state ξ_a and a sampled action $z_a \sim \pi_{\theta_a}(\cdot \mid \xi_a)$, GRPO maximizes the expected relative advantage within a sampled group $\mathcal{G}_a = \{z_a^{(1)}, \ldots, z_a^{(K)}\}$:

$$\mathcal{L}_{\mathrm{RL}}^{(a)}(\theta_a) = -\mathbb{E}_{\xi_a} \left[\frac{1}{K} \sum_{k=1}^{K} \hat{A}_a^{(k)} \log \pi_{\theta_a}\left(z_a^{(k)} \mid \xi_a\right) \right] \tag{7}$$

where K represents the number of sampled candidate actions in each GRPO group, $\hat{A}_a^{(k)}$ denotes the normalized advantage computed from relative rewards within the group $\mathcal{G}_a$, and $\mathbb{E}_{\xi_a}[\cdot]$ shows the expectation over the agent-specific input distribution $\mathcal{D}_a$.

Specialized Rewards. Each agent a is optimized with a specialized reward tailored to its functional role (§D.4). Let R_a denote the total reward for agent a, which decomposes into a weighted sum of multiple components:

$$R_a = \sum_m \lambda_{a,m}\, r_{a,m} \tag{8}$$

where m is the index over reward components for agent a, $r_{a,m}$ is individual reward term, and $\lambda_{a,m}$ is its corresponding weight with $\sum_m \lambda_{a,m} = 1$.

4 Experiments

Model. Across baselines and ANAGENT (§D.1), we evaluate two *close-source* MLLMs, `GPT-4.1-mini` (OpenAI, 2025) and `Gemini-2.5-Flash` (Google Developers Blog, 2025), and six *open-source* MLLMs, `InternVL-3.5` (Wang et al., 2025a), `Qwen2.5-VL` (Bai et al., 2025b), and `Qwen3-VL` (Bai et al., 2025a) of different sizes.

Data. To maintain computationally efficiency while reflect real-world complexity, we randomly sample from ANABENCH (§2) across 170 different scientific domains (Tab. 6), reducing the overall training size while maintaining the real-world data distribution.

Implementation Details. During SFT, agents are trained for one *epoch* on the SFT training set (Tab. 7) with initial *learning rate* 1×10^{-4} and *cosine* scheduler. During RL training, each agent is optimized on tailored datasets (Tab. 8) for one *epoch* with initial *learning rate* 1×10^{-6} to develop specialized skills. See detailed configurations and computation overhead in Tabs. 10-11.

5 Results

5.1 Scientific Table & Figure Analysis

Effectiveness of Training-Free ANAGENT. Tab. 1 summarizes the performance of baselines and training-free ANAGENT. Across eight backbone MLLMs (§4), ANAGENT consistently achieves notable gains, demonstrating the effectiveness and robustness of multi-agent collaboration. Under zero-shot setting, ANAGENT yields marked improvements over baselines, with $\Delta_{rel} \geq 4.98\%$ on S_{SEM}, $\Delta_{rel} \geq 8.54\%$ on S_{LEX}, and $\Delta_{rel} \geq 6.19\%$ on overall S_{AVG}. When extending to one-shot setting, the advantages of ANAGENT are further amplified, with relative improvements $\Delta_{rel} \geq 7.78\%$ on S_{SEM}, $\Delta_{rel} \geq 17.52\%$ on S_{LEX}, and $\Delta_{rel} \geq 9.79\%$ on S_{AVG}. Results of training-free ANAGENT demonstrate its effectiveness in consistently enhancing scientific analysis across diverse backbone models, with few-shot prompting in synergizing with ANAGENT to further unlock its potential.

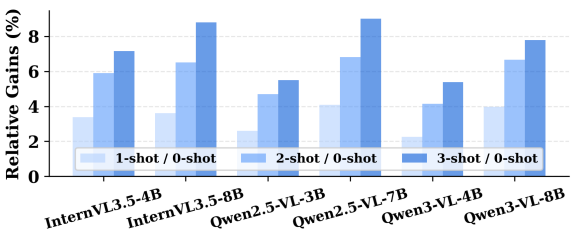


Figure 4: **Few-Shot Learning Optimization (§3.3)**

Effectiveness of Finetuned ANAGENT. ANAGENT consistently benefits from finetuning,

Model	Size	Semantic Accuracy (%)			Lexical Accuracy (%)			Overall Accuracy (%)		
		COSINE	BERT	METEOR	ROUGE-L	BLEU	WORD	S_{SEM}	S_{LEX}	S_{AVG}
ANAGENT | SFT (Zero-Shot)										
InternVL-3.5	**4B**	62.38	64.78	28.38	24.77	13.10	18.90	↑18.43%	↑101.92%	↑33.15%
	8B	64.16	65.82	29.62	25.26	14.27	20.56	↑18.99%	↑100.70%	↑33.91%
Qwen2.5-VL	**3B**	60.84	65.77	25.09	24.27	11.19	17.58	↑15.77%	↑86.30%	↑28.32%
	7B	63.21	65.97	27.63	24.91	13.96	20.21	↑16.83%	↑97.29%	↑31.75%
Qwen3-VL	**4B**	62.98	64.72	28.86	25.75	14.11	19.68	↑18.64%	↑108.29%	↑34.60%
	8B	**64.70**	**66.98**	**31.33**	**27.93**	**16.47**	**22.09**	↑21.48%	↑118.11%	↑39.40%
ANAGENT | SFT (One-Shot)										
InternVL-3.5	**4B**	62.79	65.60	27.99	24.51	13.72	18.75	↑19.07%	↑102.67%	↑33.78%
	8B	64.97	66.63	29.39	25.01	14.63	20.79	↑20.02%	↑101.80%	↑34.97%
Qwen2.5-VL	**3B**	61.25	66.54	25.27	24.27	11.26	17.53	↑16.80%	↑86.41%	↑29.18%
	7B	63.93	66.78	29.31	26.47	14.22	20.60	↑19.22%	↑104.71%	↑35.08%
Qwen3-VL	**4B**	63.51	65.22	28.30	26.34	14.05	19.77	↑18.98%	↑110.39%	↑35.28%
	8B	**65.07**	**67.13**	**31.45**	**28.08**	**16.01**	**22.59**	↑21.95%	↑118.70%	↑39.91%
ANAGENT | RL (Zero-Shot)										
Qwen2.5-VL	**3B**	56.54	60.49	21.37	18.87	5.67	12.72	↑5.61%	↑30.87%	↑10.12%
Qwen3-VL	**4B**	**58.97**	**60.99**	**22.33**	**19.09**	**5.91**	**12.76**	↑7.82%	↑32.00%	↑12.22%
ANAGENT | RL (One-Shot)										
Qwen2.5-VL	**3B**	58.19	61.10	22.01	19.40	6.20	**13.63**	↑7.83%	↑37.72%	↑13.16%
Qwen3-VL	**4B**	**60.42**	**62.24**	**22.70**	**19.51**	**6.21**	12.95	↑10.14%	↑35.26%	↑14.61%
ANAGENT | SFT+RL (Zero-Shot)										
Qwen2.5-VL	**3B**	62.61	65.90	27.22	25.22	12.59	18.78	↑18.84%	↑98.74%	↑33.10%
Qwen3-VL	**4B**	**63.13**	**66.66**	**29.46**	**26.87**	**14.84**	**20.73**	↑20.66%	↑118.36%	↑38.08%
ANAGENT | SFT+RL (One-Shot)										
Qwen2.5-VL	**3B**	62.92	66.63	27.64	26.82	14.41	19.63	↑19.96%	↑113.80%	↑36.67%
Qwen3-VL	**4B**	**63.75**	**67.89**	**30.79**	**27.91**	**15.80**	**22.03**	↑23.07%	↑130.01%	↑42.12%

Table 2: **Evaluation of Finetuned Agents.** Comparing baselines (Tab. 1) with finetuned **ANAGENT** ($M_e = 5$) on **ANABENCH** (§C.4), ***relative performance differences*** (Eq. 29) are shown as *positive* ↑ $\Delta_{rel}\%$ or *negative* ↓ $\Delta_{rel}\%$.

with the combination of SFT+RL exhibiting the strongest optimization effects (*e.g.*, `Qwen3-VL-4B`: S_{AVG}=38.03%) in comparison to their SFT (*e.g.*, `Qwen3-VL-4B`: S_{AVG} = 36.02%) and RL (*e.g.*, `Qwen3-VL-4B`: S_{AVG} = 30.03%) counterparts, highlighting the cumulative advantages of combining SFT with RL. On the other hand, jointly considering performance (Tab. 2) and computation overhead (Tab. 10) indicates SFT provides a more favorable trade-off than RL for long-context comprehension and long-output generation (§G.2). Also, similar to training-free variants (Tab. 1), finetuned **ANAGENT** presents consistent improvements from zero-shot to one-shot scientific analysis ($\Delta \geq 3.04\%$). This reveals the effectiveness of one-shot learning in multi-agent systems, where agents with specialized objectives can better coordinate and reason with minimal task-specific demonstrations.

5.2 Test-Time Optimization

Enhancing Scientific Analysis via Few-Shot Learning. To better accommodate heterogeneous inputs, we incorporate k-shot learning to improve both individual adaptability and overall coordination (§3.3). Compared with zero-shot generation, one-shot learning yields marked gains (Tabs. 1-2). Increasing the number of shots further improves performance ($k > 1$), with $k = 3$ achieving the highest, though relative gains diminish as k grows. Considering computational efficiency, $k = 1$ provides the most favorable trade-off between performance and cost. As shown in Fig. 4, few-shot learning enables **ANAGENT** to more effectively leverage prior knowledge and achieve improved coordination.

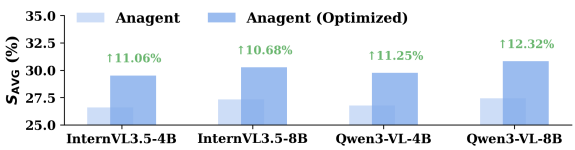


Figure 5: **Agent-Level Ability Augmentation (§3.3)**

Enhancing Scientific Analysis via Agent-Level Capability Augmentation. We conduct controlled experiments in which `GPT-4.1-mini` pow-

ers PLANNER, EXPERT, and CRITIC, with four different MLLMs instantiating SOLVER, respectively. Results in Fig. 5 show that agent-level capability augmentation (§3.3) consistently improves the overall performance of ANAGENT across all four SOLVER backbones ($\Delta_{rel} \geq 10.68\%$). Notably, augmenting only selected agents with a more capable MLLM leads to marked gains over homogeneous ANAGENT, despite leaving SOLVER unchanged. These findings highlight the significance of agent-level capability differentiation in multi-agent systems and demonstrate that selectively augmenting critical roles, especially those tasked with global guidance and complex reasoning, can effectively enhance coordination performance.

5.3 Ablations On ANAGENT Variants

Effectiveness of Multi-Agent Scientific Analysis. The performance of ANAGENT variants (Tab. 9) varies across training-free and finetuned settings (Fig. 17). Comparing training-free variants, *Omnion* consistently underperforms baselines ($\downarrow 3.89\% \leq \Delta_{abs} \leq \downarrow 6.90\%$), revealing that providing a standalone SOLVER with diverse tools can overwhelm reasoning and fails to enable effective scientific analysis. *Symnion* improves upon *Omnion* ($\Delta_{abs} \geq \uparrow 1.39\%$) by including EXPERT to assist tool invocation and context comprehension, yielding performance that is generally above baselines but remains unstable and occasionally inferior. This unveils the key insight that the absence of global planning can lead to suboptimal coordination and misleading intermediate decisions. In contrast, by integrating high-level planning, interactive executing, context-aware problem-solving, and reflective refinement (§3), ANAGENT consistently achieves the highest performance. Among all variants, finetuning leads to marked gains over training-free counterparts ($\Delta_{abs} \geq \uparrow 3.12\%$), even finetuned *Omnion* surpassing baselines, highlighting the importance of targeted finetuning in optimizing multi-agent coordination.

Effectiveness of Critic-Guided Optimization. Comparing ANAGENT with and without CRITIC reveals contrasting effects (Fig. 17). For training-free ANAGENT, incorporating CRITIC can degrade collaborative performance for smalls-size MLLM agents as a result of their ineffective reflection. This effect is different for more capable agents, unveiling the limited reasoning and reflection abilities of smaller MLLM agents. In contrast, finetuned CRITIC is able to more accurately assess intermediate solutions and identify key deficiencies, guiding effective refinements to improve overall performance. These findings underscore both the challenges and the significance of equipping agentic systems with robust reflection and refinement abilities in tackling complex problems.

Model	Size	S_{SEM}	S_{LEX}	S_{AVG}
ANAGENT (Training-Free)				
Qwen2.5-VL	**3B**	46.18	10.91	↑7.37%
Qwen3-VL	**4B**	46.95	10.50	↑7.36%
ANAGENT (End-to-End)				
Qwen2.5-VL	**3B**	48.32	14.56	↑18.24%
Qwen3-VL	**4B**	49.17	14.87	↑19.66%
ANAGENT (Modular)				
Qwen2.5-VL	**3B**	51.91	18.86	↑33.10%
Qwen3-VL	**4B**	**53.08**	**20.81**	↑38.08%

Table 3: **End-to-End Training vs. Modular Training.**

5.4 In-Depth Analysis

Validation via MLLM-As-Judge & Case Studies. Tab. 12 showcases consistent performance gains across six backbone MLLMs (§G.1), with overall S_{MLLM} achieving up to $\Delta_{rel} = 29.24\%$. This validates our design of multi-metric evaluation (§2.4). Through dedicated case studies (§I) on seven error patterns (Fig. 6), Fig. 35 reveals substantial reductions across all error types, indicating the effectiveness of ANAGENT in advancing scientific reasoning & understanding across seven complexity dimensions (§I).

Modular Training Is Better Than End-to-End Training For Multi-Agent Optimization. Comparing modular training (§3.4) with end-to-end training, we evaluate their impact on multi-agent collaboration. As shown in Tab. 3, modular training consistently outperforms end-to-end training across all metrics with notable gains ($\Delta_{rel} \geq \uparrow 33.10\%$). These results reveal that modular training more effectively supports coordinated behaviors and leads to stronger overall performance. In contrast, end-to-end training markedly constrains agents from developing and preserving specialized capabilities for designated roles. For example, in some cases, PLANNER directly generates final solutions during the planning stage (§I.8), significantly undermining role specialization and leading to degraded performance ($\Delta_{abs} \geq \downarrow 2.68\%$). This loss of specialization ultimately hampers collaborative effectiveness, highlighting the significance of modular optimization in multi-agent systems.

Unpacking the Training Data Recipe For Multi-Agent Finetuning. To understand how training data affects multi-agent finetuning, we conduct ablation studies along four dimensions of the training set (Fig. 10). As shown in Fig. 10(a), training 30K subset consistently underperforms training on the full set ($\Delta_{rel} \geq \downarrow 10.27\%$), unveiling the benefits of larger-scale training data. Fig. 10(b) compares domain-specific training with training on nine-domain full set. Domain-specific learning results in pronounced performance degradation ($\Delta_{rel} \geq \downarrow 26.55\%$), revealing that restricting training domains significantly limits agents' generalizability to out-of-domain tasks. Fig. 10(c) illustrates that single-format training impairs cross-format generalization, leading to consistent performance drops ($\Delta_{rel} \geq\downarrow 8.85\%$). Fig. 10(d) further demonstrates that limiting training to a single data type markedly degrades performance ($\Delta_{rel} \geq\downarrow 9.53\%$). We extend our discussion in §G.3.

Tools Are The Key To Open The Door of Good Scientific Analysis. Tools play a pivotal role in enabling high-quality scientific analysis by exposing ANAGENT to extended knowledge and context. Figs. 32-33 demonstrate that performance gains arise not merely from the availability of tools, but from their strategic and objective-aligned utilization. When tool functionalities are accurately matched to task demands, ANAGENT is able to effectively retrieve relevant context and domain knowledge, ground reasoning in additional evidence, and adapt analysis to task-specific scientific scenarios. We extend our discussion in § G.9.

6 Conclusion

In advancing real-world scientific table & figure analysis, we introduce ANABENCH, a large-scale benchmark of 63,178 instances built upon real-world scientific publications. By covering nine domains and seven complexity dimensions, ANABENCH casts scientific table & figure analysis as open-ended analysis writing. Building on ANABENCH, we propose ANAGENT, a multi-agent framework that improves analysis quality by up to $\uparrow$ 13.43% with test-time optimization and $\uparrow$ 42.12% with modular training. We hope ANABENCH and ANAGENT meaningfully support future research on agentic reasoning and scientific analysis.

Use of AI

We used AI assistants only to polish writing and correct grammar. All research questions, methodologies, experimental design, analyses, and conclusions are the authors' own. This is separate from our methodological and experimental use of MLLMs within ANABENCH and ANAGENT, which is documented in the corresponding sections.

Limitations

In this work, we address scientific table & figure analysis by proposing (1) ANABENCH (§2), a benchmark with 63,178 instances along seven complexity dimensions (Fig. 1), and (2) ANAGENT (§3), a multi-agent system for enhanced scientific table & figure analysis. Through test-time optimization (§3.3) and modular training (§3.4), ANAGENT achieves substantial improvements on ANABENCH (§5), revealing the effectiveness of task-oriented decomposition, strategic knowledge retrieval, and context-aware problem-solving in tackling complex scientific problems. Nevertheless, we acknowledge our limitations in four aspects, including computation overhead, domain coverage, tool dependency, and evaluation (§J). We hope ANABENCH and ANAGENT provide meaningful foundations to facilitate future research.

Acknowledgments

The authors acknowledge William & Mary Research Computing for providing computational resources and/or technical support that have contributed to the results reported within this paper. This work is supported by William & Mary Applied Research & Innovation Initiative. This work used DeltaAI at NCSA through allocation CIS260012 and CIS230280 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904. This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH author(s) are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views

of the NIH or the U.S. Department of Health and Human Services.

References

arXiv. 1991. arxiv e-print archive. https://arxiv.org/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, and 22 others. 2025a. Qwen3-vl technical report. *arXiv*, abs/2511.21631. Preprint.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *arXiv*, abs/2502.13923. Preprint.

Yu Bai, Baoqiang Liu, Shuang Xue, Fang Cai, Na Ye, and Guiping Zhang. 2025c. Reasoning knowledge filter for logical table-to-text generation. In *Proceedings of Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025*, pages 18–30, Abu Dhabi, UAE. ELRA and ICCL.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Daniil A. Boiko, R. MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624:570 – 578.

Yee Man Choi, Xuehang Guo, Yi R. Fung, and Qingyun Wang. 2025. Citeguard: Faithful citation attribution for llms via retrieval-augmented validation. *ArXiv*, abs/2510.17853.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. 2025. Tabarena: A living benchmark for machine learning on tabular data. *arXiv preprint arXiv:2506.16791*.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and M. Zitnik. 2024. Empowering biomedical discovery with ai agents. *Cell*, 187:6125–6151.

Aniketh Garikaparthi, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. 2025. IRIS: Interactive research ideation system for accelerating scientific discovery. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 592–603, Vienna, Austria. Association for Computational Linguistics.

Google Developers Blog. 2025. Start building with gemini 2.5 flash.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2025. Towards an ai co-scientist. *CoRR*, abs/2502.18864.

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and C. Mack. 2025. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions. *ArXiv*, abs/2503.08979.

Xuehang Guo, Xingyao Wang, Yangyi Chen, Sha Li, Chi Han, Manling Li, and Heng Ji. 2025. Syncmind: Measuring agent out-of-sync recovery in collaborative software engineering. *arXiv preprint arXiv:2502.06994*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Chuhan Li, Ziyao Shangguan, Yilun Zhao, Deyuan Li, Yixin Liu, and Arman Cohan. 2024. M3sciqa: A multi-modal multi-document scientific qa benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15419–15446, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain. Association for Computational Linguistics.

Tengxiao Liu, Deepak Nathani, Zekun Li, Kevin Yang, and William Yang Wang. 2026. Wildsci: Advancing scientific reasoning from in-the-wild literature. *arXiv preprint arXiv:2601.05567*.

Yuze Lou, Bailey Kuehl, Erin Bransom, Sergey Feldman, Aakanksha Naik, and Doug Downey. 2023. S2abel: A dataset for entity linking from scientific tables. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 3089–3101, Singapore. Association for Computational Linguistics.

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7813, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing scientific literature into tables using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9631, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI. 2025. Introducing gpt-4.1 in the api.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. In *Advances in Neural Information Processing Systems*.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. 2024. Citeme: Can language models accurately cite scientific claims? *ArXiv*, abs/2407.12861.

PubMed. 1996. Pubmed: Database of biomedical literature. https://pubmed.ncbi.nlm.nih.gov/.

Chandan K. Reddy and Parshin Shojaee. 2024. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *AAAI Conference on Artificial Intelligence*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300.

Shruti Singh, Nandan Sarkar, and Arman Cohan. 2024. SciDQA: A deep reading comprehension dataset over scientific papers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20908–20923, Miami, Florida, USA. Association for Computational Linguistics.

Anirudh S. Sundar, Jin Xu, William Gay, Christopher Richardson, and Larry Heck. 2024. cpapers: A dataset of situated and multimodal interactive conversations in scientific papers. *ArXiv*, abs/2406.08398.

Brian Wang, Bin Feng, Xiaoman Pan, Chenyang An, Felix Liu, Tangqi Fang, Gongbo Sun, Lingfeng Shen, Ning Wang, Handuo Zhang, and 1 others. 2026. Apodex discovery: Reality benchmarks and environments for evaluating and building discoverative artificial intelligence. *arXiv preprint arXiv:2608.11341*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 41 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv*, abs/2508.18265. Preprint.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and 1 others. 2025b. Perception-aware policy optimization for multimodal reasoning. *arXiv preprint arXiv:2507.06448*.

Xuanliang Zhang, Dingzirui Wang, Baoxin Wang, Longxu Dou, Xinyuan Lu, Keyan Xu, Dayong Wu, and Qingfu Zhu. 2025. Scitat: A question answering benchmark for scientific tables and text covering diverse reasoning types. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3859–3881, Vienna, Austria. Association for Computational Linguistics.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024a. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.

Zhi Zhang, Yan Liu, Sheng hua Zhong, Gong Chen, Yu Yang, and Jiannong Cao. 2024b. From references to insights: Collaborative knowledge mini-graph agents for automating scholarly literature review. In *AAAI Conference on Artificial Intelligence*.

Xuanle Zhao, Xianzhen Luo, Qi Shi, Chi Chen, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2025. Chart-Coder: Advancing multimodal large language model for chart-to-code generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7333–7348, Vienna, Austria. Association for Computational Linguistics.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China. Association for Computational Linguistics.

Zekun Zhou, Xiaocheng Feng, Lei Huang, Xiachong Feng, Ziyun Song, Ruihan Chen, Liang Zhao, Weitao Ma, Yuxuan Gu, Baoxin Wang, Dayong Wu, Guoping Hu, Ting Liu, and Bing Qin. 2025. From hypothesis to publication: A comprehensive survey of ai-driven research support systems. *ArXiv*, abs/2503.01424.

A Preliminary Exploration

A.1 What Challenges Exist In Scientific Table & Figure Analysis?

The heterogeneity of *table & figure scientific analysis* poses critical challenges for MLLM agents in accurately understanding various modalities, structures, formats, contexts, domains, and writing demands. To investigate how MLLM agents can perform in tackling these challenges, we employ Qwen3-VL-8B as the base model, generating analysis on 60 scientific tables and figures, respectively. These 120 data are randomly sampled from ANABENCH with evenly distributed features across both **data complexity** (including *data type*, *data format*, *data source*, and *data domain*) and **analysis complexity** (including *analysis objective*, *analysis depth*, and *analysis width*) (§2.3).

As shown in Fig. 1, the agent presents varying performance on these dimensions, showcasing strengths on table structures over visual figures, XML formats over LaTeX, arXiv (arXiv, 1991) papers over PubMed (PubMed, 1996), computer science related domains over biomedicine, experimental analysis over methodology, superficial summarization over in-depth analysis, and fully-grounded analysis over inferential writing. Analyzing case-by-case, we observe seven distinct patterns (Fig. 6) in agent's analysis failures, which are in close accordance with the main challenges we conclude in Fig. 1.

Employing Qwen3-VL-8B as the analysis agent baseline, the agent performs scientific analysis on different data types, formats, sources, and domains (Figs. 6-8). As shown in Fig. 7, the agent analyzes two distinct data types, *table* and *figure*, respectively. While the table is a text-only single modality input, it presents different structures than writing or drawing. On the other hand, the figure gives vision-language input, complicating the multimodal understanding that serves as the basis of accurate analysis. Compared with ground-truth analysis, the agent's outputs reveal their significant visual perception errors, together with their lack of scientific writing abilities with proper analysis depth and width. Fig. 8 shows two examples with different data types, sources, formats, and domains. As domain-specific knowledge further complicates the tasks, the agent analyzes both table and figure with notable hallucinated contents. Meanwhile, the agent's analysis contains significant misinterpretation on domain-specific terminologies and expressions, and lacks wider analysis coverage and deeper discussions.

A.2 How To Enhance Scientific Table & Figure Analysis?

Benchmarking Scientific Analysis Under Realistic Challenges. Our preliminary exploration (§A.1) reveals that failures in scientific table & figure analysis are not stemming from isolated weaknesses, but rather from the compound difficulty introduced by heterogeneous data representations and diverse analytical demands (Fig. 1). To this end, we propose ANABENCH (§2), a benchmark that systematically instantiates these challenges based on real-world scientific publications. By covering diverse data modalities and structures, types and formats, sources and domains, ANABENCH develops scientific analysis tasks with varying analysis widths, depths, and objectives, enabling controlled evaluation of MLLM agents' scientific analysis capabilities across various data and analysis complexities. Aligning benchmark construction (§2.3) directly with the observed challenges (Fig. 1), ANABENCH establishes a principled foundation for evaluating and advancing scientific table & figure analysis.

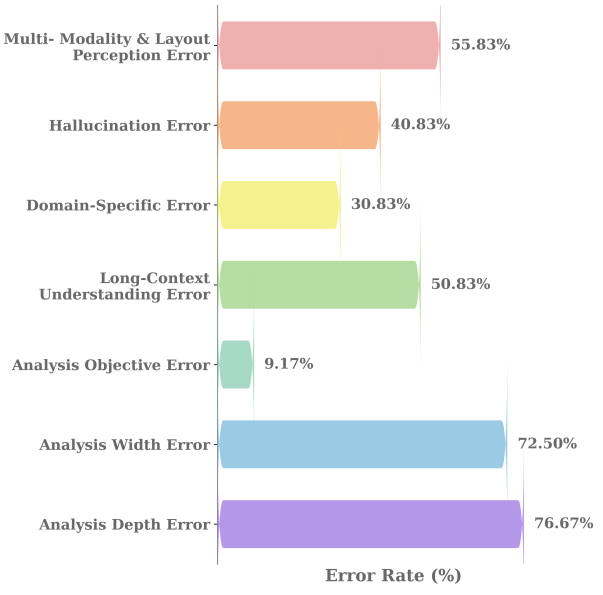


Figure 6: **Key Error Patterns.** Through our case study on 120 samples, we identify seven key error patterns that correspond to the analysis challenges shown in Fig. 1.

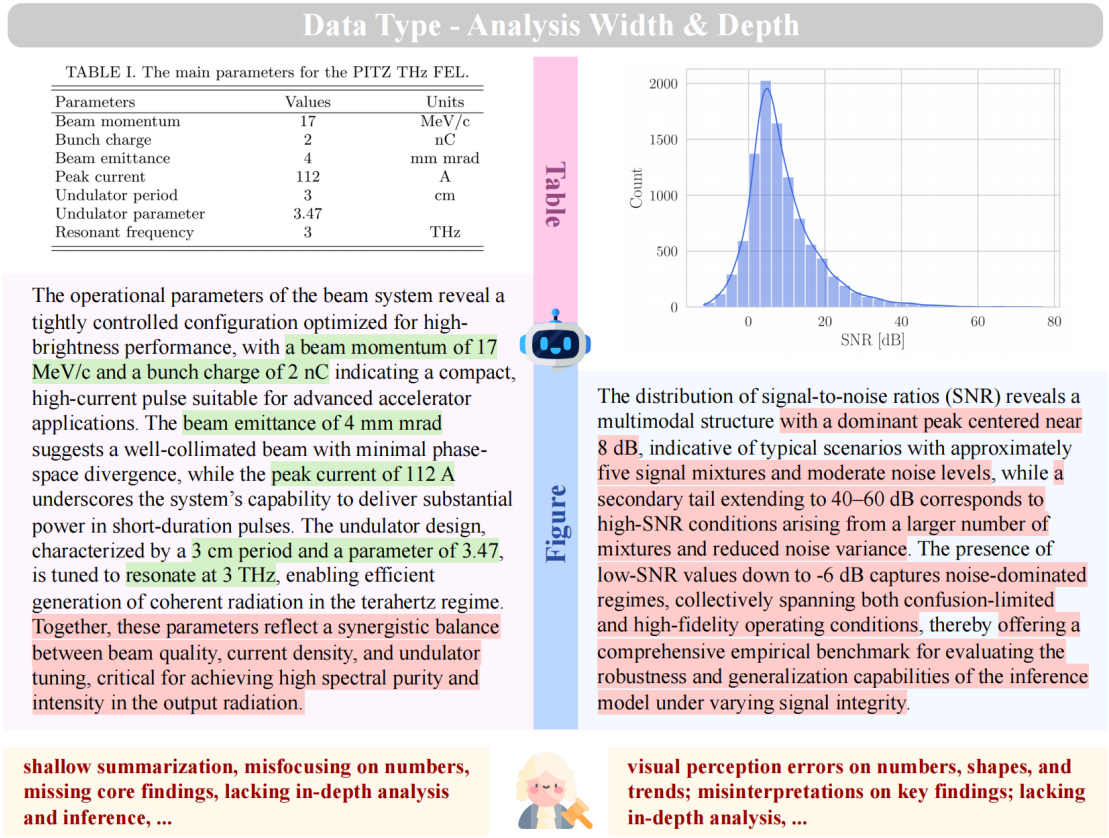


Figure 7: **Preliminary Analysis On Failure Patterns.** Analyzing on two different data types, *table* & *figure*, the agent shows its lack of accurate visual perception and its incapability of writing with proper analysis depth and width.

Enhancing Scientific Analysis Through Multi-Agent Collaboration. The identified error patterns present across different stages of scientific writing, from input interpretation and contextual understanding to analysis writing and reflective correction. As such, a high-quality scientific analysis demands deliberate planning, accurate and domain-aware information acquisition, coherent synthesis, and careful verification, which require capabilities that are difficult to reliably achieve within a monolithic agent (§5.3). *How to enhance these specialized capabilities in one agentic framework?* By decomposing the scientific analysis writing process into multiple stages, we introduce ANAGENT, a multi-agent framework capable of collaborative scientific analysis with specialized completion and optimization. Specifically, ANAGENT consists of four specialized agents: PLANNER for high-level task planning, alleviating potential errors in analysis width, depth, and objectives; EXPERT for task-oriented exploration and retrieval, avoiding domain-specific and contextual understanding errors across varied modalities, structures, and formats; SOLVER for scientific analysis writing, reasoning through all the available knowledge and information as guided and supported by PLANNER and EXPERT; and CRITIC for self- reflection and correction to rectify inaccurate analyses

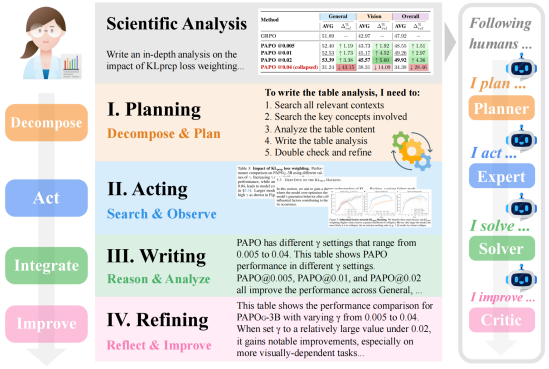


Figure 9: **Scientific Analysis Workflow.** Motivated by how human researchers perform scientific analysis, we decompose the scientific analysis workflow into dedicated stages, which leads to ANAGENT (Fig. 3).

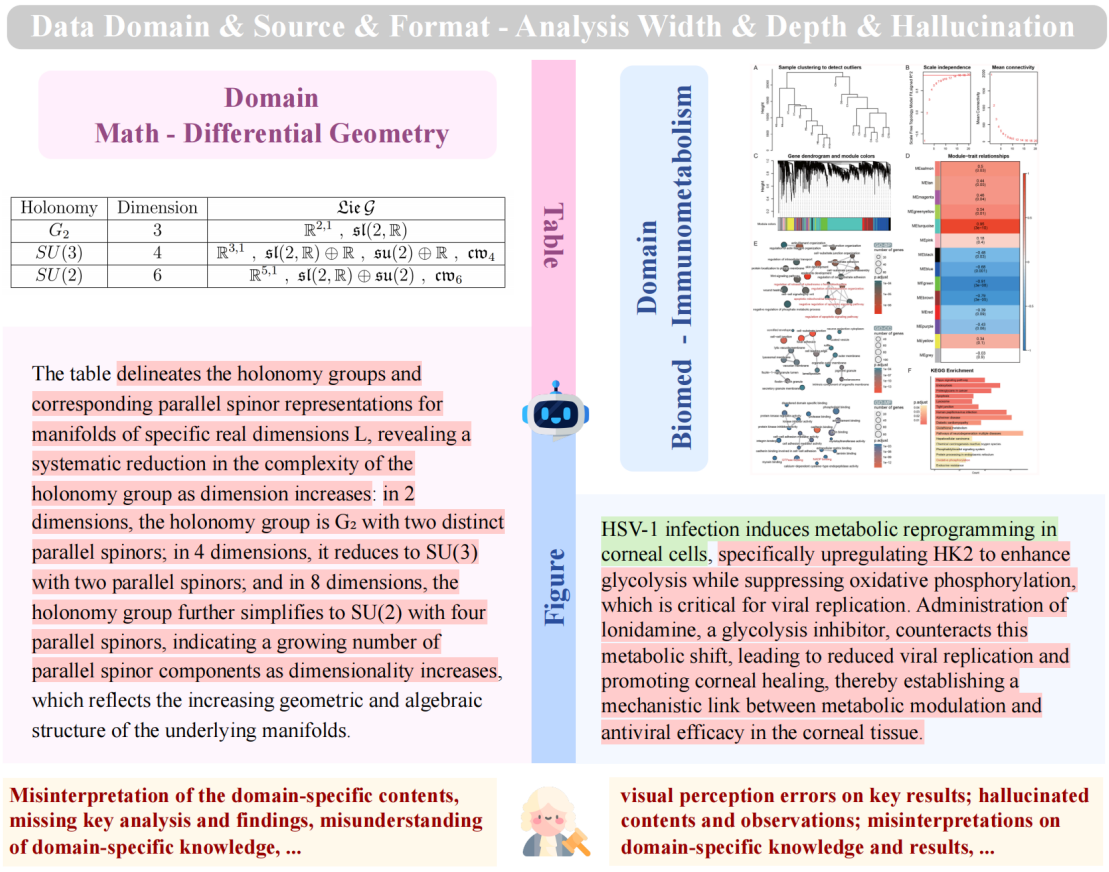


Figure 8: **Preliminary Analysis On Failure Patterns.** Analyzing on two data types with different data sources, formats, and domains, the agent generates analysis with significant hallucinated contents, meanwhile incapable of writing with proper analysis depth and width.

and hallucinated contents. Collectively, ANAGENT targets the key failure patterns observed in Fig. 6 to enhance agent's scientific analysis capabilities for improved overall performance.

Evaluating Scientific Analysis With Multi-Dimensional Assessment. Given the diversity of challenges (Fig. 1) and error patterns (Fig. 6) involved in scientific analysis, single-scalar judgment is insufficient to reflect analysis quality. Accurate scientific analysis requires faithful interpretation of presented data, comprehensive coverage of key findings, adherence to task-specific requirements, clear and coherent scientific writing, and strict grounding in available evidence. Accordingly, our assessment considers these aspects jointly, capturing both analytical correctness and writing quality through the five-dimensional evaluation protocol (Fig. 44): *content accuracy*, *analytical completeness*, *format correctness*, *clarity & coherence*, and *reliability & faithfulness*. To complement our rule-based assessment, we apply *five-dimensional protocol* to both MLLM-As-Judge and human expert assessment (§2.4), enabling fine-grained comparisons across models and settings.

B Related Work

AI For Scientific Table Understanding. Recent advances in AI greatly inspire research on table understanding (Erickson et al., 2025), particularly scientific tables that exhibit diverse formats, layouts, domains, and analytical objectives. As tables constitute a compact yet information-dense medium for conveying methodological details and empirical findings, benchmarks are proposed to evaluate distinct aspects of scientific table understanding: SCITAB (Lu et al., 2023) assesses table-based claim verification, S2abEL (Lou et al., 2023) targets entity linking, and other benchmarks address question answering (QA)

(Pramanick et al., 2024; Zhang et al., 2025), table-to-text generation (Bai et al., 2025c), literature-to-table (Newman et al., 2024), etc. However, existing benchmarks emphasize isolated tasks while lack principled curriculum to capture diverse data heterogeneity and reasoning complexity in long-horizon contexts, motivating our work to benchmark scientific table understanding across multiple complexity dimensions.

AI For Scientific Multimodal Understanding. Scientific papers are inherently multimodal, combining text with figures, tables, algorithms, etc., to communicate complex scientific evidence (Zheng et al., 2025; Zhang et al., 2024a). Accordingly, multimodal reasoning and long-context comprehension are essential for scientific research. However, existing benchmarks have significant limitations: SPIQA (Pramanick et al., 2024) for table & figure QA shows limited coverage of cross-domain generalization and reasoning complexity. WildSci (Liu et al., 2026) targets QA across domains, yet fails to incorporate multimodal long-context reasoning that is fundamental to scientific inquiry. These limitations motivate ANABENCH with structured reasoning curriculum to provide more comprehensive testbed for enhancing multimodal scientific understanding.

C ANABENCH: Benchmark Analysis

We construct ANABENCH that covers seven key challenges (Fig. 1), with our construction method scalable to different sizes for custom use (Fig. 2). By developing an automated multi-stage benchmark construction method (§2), ANABENCH captures a wide range of data complexity (§C.2.1) and analysis complexity (§C.2.2), enabling more comprehensive evaluation of scientific analysis. Our multi-level filtering and quality-control procedures further ensure high data reliability. The comparison between ANABENCH and recent scientific benchmarks is summarized in Tab. 4.

C.1 Benchmark Construction

As illustrated in Fig. 2, our dataset construction method comprises four progressive automated stages: (1) *source collection*, (2) *data extraction*, (3) *instance construction*, and (4) *task classification*. To ensure data quality, we implement multi-level filtering across stages, from *source collection* to *instance construction*. Here, we elaborate on our benchmark construction in further detail to complement ANABENCH:

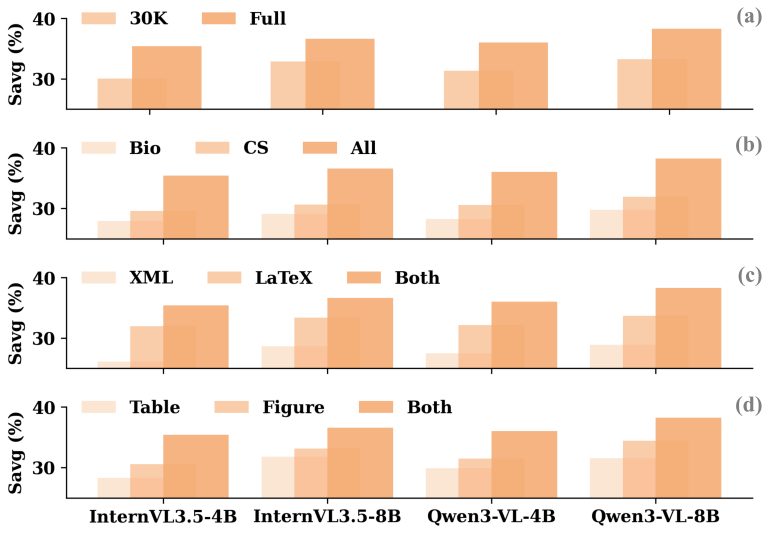


Figure 10: **Ablations On Training Data.** Performance visualization of ablation studies (§5.4), respectively on: (a) *data size*, (b) *data domain*, (c) *data format*, (d) *data type* (§C.2.1).

Source Collection. During the initial stage of *source collection*, we gather source papers from multiple dissemination platforms and apply a combination of paper-level filters, including domain-category filtering, publication-year filtering, keyword-based filtering, full-text access filtering, and maximum-source thresholding. In particular, to mitigate the risk of data contamination during model pretraining, we restrict sources to papers published after 2023. Moreover, to ensure data quality and better coverage of recent work, we set the maximum source threshold for papers published in or after 2025 to be twice that of papers published before 2025.

Data Extraction. In the second stage of *data extraction*, we perform both *paper-level* and *data-level* filtering based on automated data parsing. Specifically, we filter out papers and data instances that exhibit access failures or parsing errors. For each retained figure or table data, we extract the parsed data content along with the associated source files when available (*e.g.*, PNG images). In addition, we extract contextual information for each targeted data through d-depth hierarchical intra-document and inter-document reference retrieval. Our d-depth hierarchical context retrieval method (Alg. 1) is implemented in a recursive manner: the first-level context consists of elements that the target instance refers to or is referred to by; the second-level context includes elements that the first-level contexts refer to or are referred to by; and this process continues iteratively up to depth d. This hierarchical context retrieval enables the extraction of both internal and external relational information surrounding each data sample.

Gold Analysis. Every gold analysis in ANABENCH is author-written text from the peer-reviewed source paper (Fig. 2): for each target table or figure, we extract verbatim the analytical prose in which the original authors discuss it. No reference analysis is model-generated or model-edited. Also, contexts are retrieved separately through d-depth hierarchical reference retrieval (Alg. 1), so d controls only the context scope supplied to a model. MLLM assistance in constructing ANABENCH is confined merely to labeling two of the seven complexity dimensions, *analysis objective* and *analysis depth* (§C.2.3); the other five and all filtering are rule-based (§C.1, §C.4). Domain experts rate the gold analyses at 93.78% under the same five-dimensional protocol applied to model outputs (§G.8).

Instance Construction. Supported by the prior two stages, the *instance construction* stage integrates the targeted data, d-depth contexts, ground-truth analysis, and source metadata to create each scientific analysis instance. This stage performs multi-level *data cleaning*, including data filtering that excludes embedded elements, threshold-based filtering that removes instances with over-short or over-long inputs and outputs according to the predefined thresholds, and data validation that discards data with missing targeted samples or ground-truth analyses. The resulting cleaned instances are then stored in ANABENCH for subsequent task classification.

Task Classification. We combine *rule-based* task classification with *MLLM-assisted* curriculum categorization to classify scientific analysis instances into fine-grained curriculum categories across seven complexity dimensions (§C.2.3). We summarize the complexity curriculum categories in Tab. 5, with 23 task complexity categories across four data complexity dimensions (§C.2.1) and three analysis complexity dimensions (§C.2.2).

Quality Control. To ensure data quality, we implement multi-level filtering and data cleaning across different benchmark construction stages (§C). Furthermore, to mitigate the risk of data contamination during model pretraining, we restrict paper sources to those published after 2023 at the initial source-collection stage of our benchmark construction (Fig. 2), with 2025 accounting for the majority of instances to mitigate data contamination (Fig. 11). Accordingly, our evaluation set (§C.4) is obtained by filtering ANABENCH to instances derived from papers published in 2025 and then downsampling this subset.

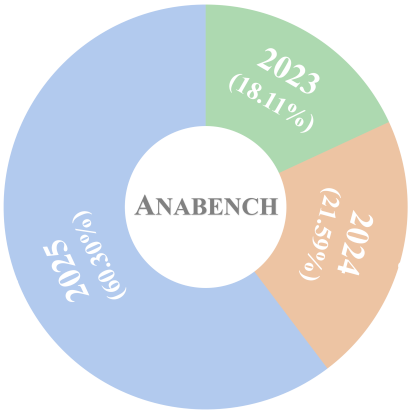


Figure 11: **Year Distribution of ANABENCH.** Visualization of year distribution, with 2025 comprising the largest proportion to mitigate data contamination (§C).

C.2 Benchmark Curriculum

According to the difficulty and diversity of the data, ANABENCH is organized into a curriculum along two overarching dimensions (§2.3), **data complexity** (§C.2.1) and **analysis complexity** (§C.2.2), to capture and reflect real-world variations in both scientific inputs and analytical demands, enabling systematic

Algorithm 1 k-Depth Hierarchical Context Retrieval

Require: Target data instance d, maximum context depth k, reference graph $\mathcal{G}$
Ensure: Hierarchical context set $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$

```
Initialize C ← ∅
Initialize visited set V ← {d}
Initialize frontier F_0 ← {d}
for i = 1 to k do
    Initialize F_i ← ∅
    for all e ∈ F_{i−1} do
        Retrieve R_in(e) and R_out(e) from G
        for all e′ ∈ R_in(e) ∪ R_out(e) do
            if e′ ∉ V then
                Add e′ to F_i
                Add e′ to V
            end if
        end for
    end for
    C_i ← F_i
end for
return C
```

evaluation across heterogeneous scenarios. To determine benchmark curriculum, we perform fine-grained task classification across data and analysis complexities (§C.2.3).

C.2.1 Data Complexity

Data Type. ANABENCH covers different data modalities and structures commonly encountered in scientific literature. Specifically, the input data include structured *tables* that present single-modality data with explicit tabular organization, and *figures* that are inherently multimodal and consist of both visual and textual elements. The *table* category includes diverse tabular organizations with varying layouts, levels of sparsity, and semantic density, requiring structured parsing and relational reasoning. On the other hand, the *figure* category spans a wide range of visual structures, such as charts, plots, frameworks, diagrams, *etc*., introducing additional challenges in visual interpretation and cross-modal alignment between textual and visual elements.

Data Format. To reflect the real-world diversity of scientific document representations, the input data are supplied in both `LaTeX` and `XML` formats. These formats differ substantially in syntactic structure and parsing complexity, requiring models to handle distinct markup conventions while preserving the underlying semantic content.

Data Source. The benchmark incorporates data collected from publications across not only dissimilar literature categories, including *general* papers and *review or survey* papers, but also different dissemination platforms, such as *arXiv* (arXiv, 1991) and *PubMed* (PubMed, 1996). These sources vary in writing structures, submission formats, and disciplinary emphasis, contributing to increased heterogeneity in data and domain distributions.

Data Domain. ANABENCH spans 9 broad scientific domains, covering 170 fine-grained disciplines (Tab. 6 & Figs. 12-14). This domain diversity supports ANABENCH to systematically evaluate the analytical capabilities of MLLM agents across varied domain-specific knowledge, terminologies, methodological conventions, and writing norms.

Table 4: **Benchmark Comparison.** We compare ANABENCH with existing scientific benchmarks across multiple dimensions of data complexity (M-mod.: multimodal, M-Layout: multi-layout, M-Doc.: multi-document, M-Src.: multi-source, M-Fmt.: multi-format, M-Dom.: multi-domain) and reasoning complexity (Long-Ctx.: long-context reasoning, M-Width: multi-width, M-Depth: multi-depth, M-ObJ.: multi-objective), highlighting the comprehensiveness of our approach to evaluating autonomous scientific analysis capabilities of MLLM-powered scientific agents. Note: 9/26 denotes 9 domains and 26 subdomains; 9/170 denotes 9 domains and 170 subdomains. Compared benchmarks are listed in alphabetical order.

Benchmark	Task	Source	Data Complexity						Reasoning Complexity			
			M-mod.	M-Layout	M-Doc.	M-Src.	M-Fmt.	M-Dom.	Long-Ctx.	M-Width	M-Depth	M-Obj.
M3SciQA (Li et al., 2024)	*QA*	*Partial*	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
SCIDQA (Singh et al., 2024)	*QA*	*Full*	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
SCITAB (Lu et al., 2023)	*Claim Verification*	*Partial*	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
SCITAT (Zhang et al., 2025)	*QA*	*Partial*	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
SPIQA (Pramanick et al., 2024)	*QA*	*Full*	✓	✓	✗	✓	✗	✗	✓	✗	✗	✗
S2abEL (Lou et al., 2023)	*Entity Link*	*Partial*	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
PubMedQA (Jin et al., 2019)	*QA*	*Partial*	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
WildSci (Liu et al., 2026)	*QA*	*Partial*	✗	✗	✗	✗	✗	✓ (9/26)	✗	✗	✗	✗
ANABENCH	*Scientific Analysis*	*Full*	✓	✓	✓	✓	✓	✓ (9/170)	✓	✓	✓	✓

Table 5: **AnaBench Complexity Curriculum.** This table summarizes the *task classification* stage of our benchmark construction method for the complexity curriculum of AnaBench, which defines 23 task complexity categories across seven data complexity and analysis complexity dimensions.

Curriculum	Broad	Fine-Grained	Category Definition
Data Complexity	Type	Table	Tabular-structured data that organize information into rows and columns to systematically present structured contents, such as method characteristics, parameter configurations, benchmark comparisons, categorical classifications, experimental results, quantitative or qualitative summaries, and analytical breakdowns, enabling precise lookup, comparison, and reference.
		Figure	Visual representations of scientific information that employ graphical, diagrammatic, or illustrative elements, such as plots, charts, schematics, diagrams, images, framework architectures, etc., to show relationships, patterns, processes, structures, conceptual approaches, and other critical information that may be complex or inefficient to express in textual form
	Format	LaTeX	Scientific data obtained from documents in LaTeX, which is a structured format widely used in scholarly publishing to encode mathematical expressions, tables, figures, algorithms, and document structures
		XML	Scientific data obtained from documents encoded in XML, which is a structured, machine-readable format that represents document content, metadata, and hierarchical relationships through tagged elements, commonly used for standardized archival and data exchange in scholarly publishing
	Source	General	Scientific literature that primarily focuses on the development, analysis, or evaluation of a specific method, model, framework, or benchmark, typically proposing novel techniques or reporting empirical results within a defined research problem.
		Review	Scientific literature that systematically examines, summarizes, and synthesizes existing research within a specific domain or research direction, aiming to provide an overview of prior work, identify trends, compare approaches, and highlight open challenges for future research.
	Domain	9/170	Scientific data can span multiple research domains, among which AnaBench covers 9 broad domains encompassing 170 fine-grained disciplinary categories
Analysis Complexity	Objective	Methodology	Scientific analysis that explains and rationalize the proposed method, such as model designs, training algorithms, framework architectures, etc.
		Experiment	Scientific analysis that focuses on empirically evaluating and analyzing diverse aspects like performance, robustness, trustworthiness, comparative effectiveness, etc., across different datasets, settings, or baselines using quantitative and qualitative results
	Width	Self-Contained	Scientific analysis that relies solely on the given input, without referring to additional sources
		Internal	Scientific analysis that integrates the given input with references drawn from within the same source paper
		External	Scientific analysis that integrates the given input with references drawn from sources outside the source paper
		Mixed	Scientific analysis that combines the given input with references drawn from both within and outside the source paper
	Depth	Shallow	Scientific analysis that focuses on surface-level observations, straightforward patterns, or direct summarization of the input, without extended reasoning, interpretation, or inference
		In-Depth	Scientific analysis that involves extended reasoning, deeper interpretation, and evidence-grounded inference beyond surface-level summarization to derive deeper insights, explanations, findings, or conclusions

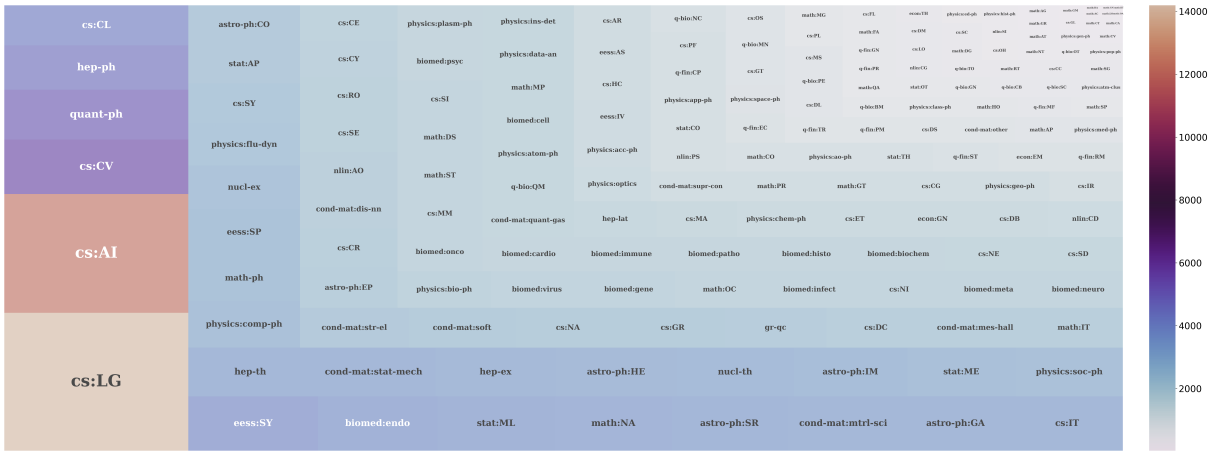
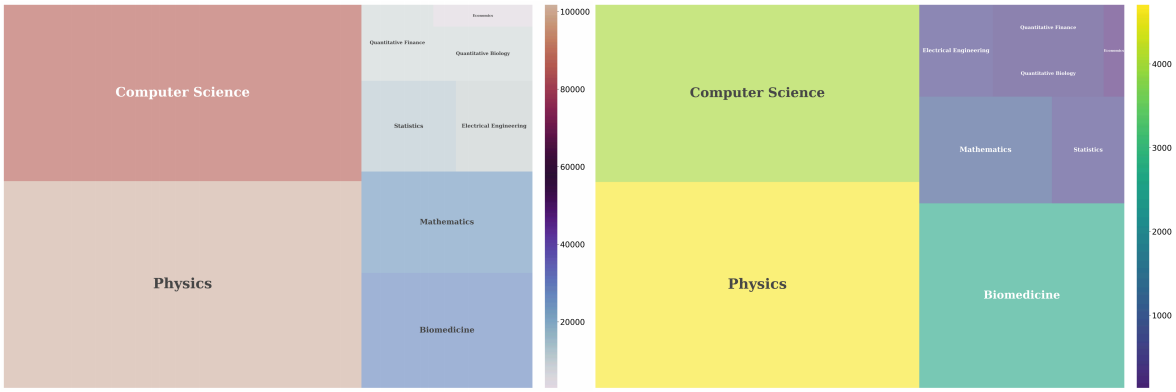


Figure 12: **ANABENCH Domain Distribution.** Domain distribution of ANABENCH across 9 broad domain categories (Fig. 13.a) and 170 fine-grained subdomain categories (Fig. 12).



(a) AnaBench (b) AnaBench - Test

Figure 13: **Broad Domain Distribution.** Broad domain distribution of **(a)** ANABENCH and **(b) the evaluation set of** ANABENCH across nine broad domain categories.

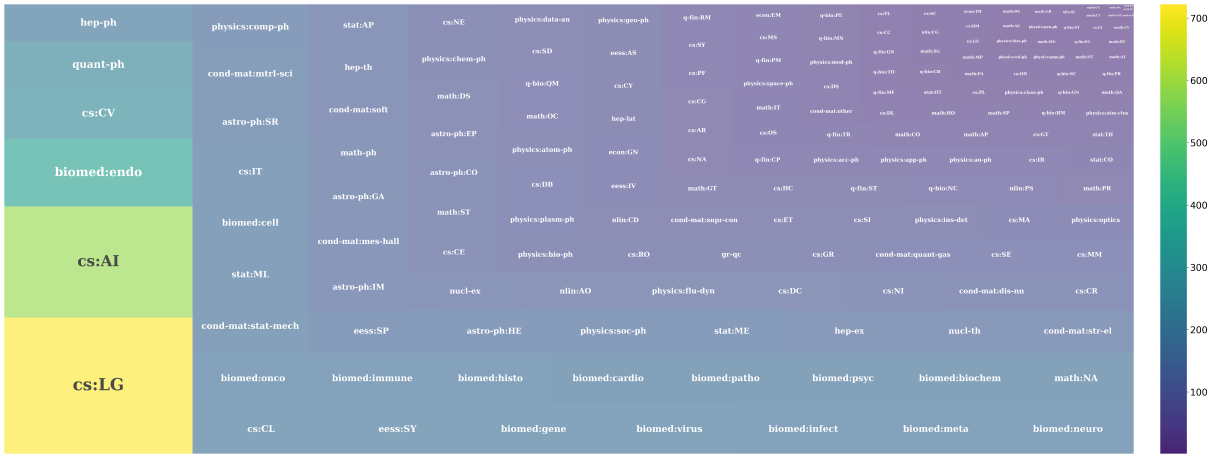


Figure 14: **ANABENCH Evaluation Domain Distribution.** Besides the domain distribution of ANABENCH (Fig. 12 & Fig. 13.a), we additionally visualize that of the downsampled evaluation set across 9 broad domains (Fig. 13.b) and 170 subdomains (Fig. 14).



Figure 15: **AnaBench: Benchmark Statistics.** We quantitatively analyze our benchmark in the dimensions of seven challenges (Fig. 1) across varying data and analysis complexities.

Table 6: **AnaBench Domains.** This table summarizes the 9 broad domains and 170 subdomains that **AnaBench** covers and collects data from. Directly summing subdomain categories is inaccurate due to subdomain overlapping.

Broad Domain	Fine-Grained Subdomain
Computer Science (40 Subdomains)	*Artificial Intelligence; Hardware Architecture; Computational Complexity; Computational Engineering, Finance, and Science; Computational Geometry; Computation and Language; Cryptography and Security; Computer Vision and Pattern Recognition; Computers and Society; Databases; Distributed, Parallel, and Cluster Computing; Digital Libraries; Discrete Mathematics; Data Structures and Algorithms; Emerging Technologies; Formal Languages and Automata Theory; General Literature; Graphics; Computer Science and Game Theory; Human-Computer Interaction; Information Retrieval; Information Theory; Machine Learning; Logic in Computer Science; Multiagent Systems; Multimedia; Mathematical Software; Numerical Analysis; Neural and Evolutionary Computing; Networking and Internet Architecture; Other Computer Science; Operating Systems; Performance; Programming Languages; Robotics; Symbolic Computation; Sound; Software Engineering; Social and Information Networks; Systems and Control*
Economics (3 Subdomains)	*Econometrics; General Economics; Theoretical Economics*
Electrical Engineering (4 Subdomains)	*Audio and Speech Processing; Image and Video Processing; Signal Processing; Systems and Control*
Physics (51 Subdomains)	*Astrophysics (Cosmology and Nongalactic Astrophysics; Earth and Planetary Astrophysics; Astrophysics of Galaxies; High Energy Astrophysical Phenomena; Instrumentation and Methods for Astrophysics; Solar and Stellar Astrophysics); Condensed Matter (Disordered Systems and Neural Networks; Mesoscale and Nanoscale Physics; Materials Science; Other Condensed Matter; Quantum Gases; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity); General Relativity and Quantum Cosmology (General Relativity and Quantum Cosmology); High Energy Physics - Experiment; High Energy Physics - Lattice; High Energy Physics - Phenomenology; High Energy Physics - Theory; Mathematical Physics; Nonlinear Sciences; Nuclear Experiment; Nuclear Theory; Physics (Accelerator Physics; Atmospheric and Oceanic Physics; Applied Physics; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Physics Education; Fluid Dynamics; General Physics; Geophysics; History and Philosophy of Physics; Instrumentation and Detectors; Medical Physics; Optics; Plasma Physics; Popular Physics; Physics and Society; Space Physics); Quantum Physics*

Broad Domain	Fine-Grained Subdomain
Mathematics (32 Subdomains)	*Commutative Algebra; Algebraic Geometry; Analysis of PDEs; Algebraic Topology; Classical Analysis and ODEs; Combinatorics; Category Theory; Complex Variables; Differential Geometry; Dynamical Systems; Functional Analysis; General Mathematics; General Topology; Group Theory; Geometric Topology; History and Overview; Information Theory; K-Theory and Homology; Logic; Metric Geometry; Mathematical Physics; Numerical Analysis; Number Theory; Operator Algebras; Optimization and Control; Probability; Quantum Algebra; Rings and Algebras; Representation Theory; Symplectic Geometry; Spectral Theory; Statistics Theory*
Quantitative Biology (10 Subdomains)	*Biomolecules; Cell Behavior; Genomics; Molecular Networks; Neurons and Cognition; Other Quantitative Biology; Populations and Evolution; Quantitative Methods; Subcellular Processes; Tissues and Organs*
Quantitative Finance (9 Subdomains)	*Computational Finance; Economics; General Finance; Mathematical Finance; Portfolio Management; Pricing of Securities; Risk Management; Statistical Finance; Trading and Market Microstructure*
Statistics (6 Subdomains)	*Applications; Computation; Methodology; Machine Learning; Other Statistics; Statistics Theory*
Biomedicine (15 Subdomains)	*General Pathology; Infectious Disease; Neurological Disease; Endocrine & Metabolic Disease; Psychiatry; Oncology; Cardiovascular System; Cell Biology; Genetics; Endocrinology; Immunology; Biochemistry; Metabolism; Histology; Virology*

Table 7: **SFT Data Distribution.** This table summarizes the data size, data type, data format, and domain categories of different downsampled datasets for SFT.

Data Size	Data Type		Data Format		Data Source		Domain
	Figure	Table	LaTeX	XML	General	Review & Survey	
Single-Format							
20,000	16,743	3,258	20,000	0	10,000	10,000	8
24,210	11,778	12,432	0	24,210	22,860	1,350	1
42,804	33,485	9,319	42,804	0	20,000	11,350	8
Multi-Format							
31,350	26,245	5,105	20,000	11,350	21,901	20,903	9
67,014	45,263	21,751	42,804	11,350	44,761	22,253	9

Table 8: **RL Data Distribution.** This table summarizes the data size, data type, data format, and domain categories of the downsampled datasets for agent-wise RL training.

Agent	Data Type		Data Format		Data Source		Domain
	Figure	Table	LaTeX	XML	General	Review & Survey	
Train							
Planner	7,870	2,584	7,632	2,822	5,829	4,625	9
Expert	10,628	2,894	8,058	5,464	7,883	5,639	9
Solver	18,737	6,444	16,331	8,850	16,868	8,313	9
Critic	8,761	4,065	7,655	5,171	7,650	5,176	9
Test							
Planner	1,168	416	1,000	584	1,000	584	9
Expert	1,594	499	1,342	751	1,312	781	9
Solver	1,992	749	1,773	968	1,838	903	9
Critic	1,242	431	1,089	584	1,034	639	9

C.2.2 Analysis Complexity

Analysis Objective. Data in ANABENCH are classified according to their analytical objectives. Specifically, each analysis is labeled as either (1) *methodology-oriented analysis* that describes methodological designs, theoretical formulations, algorithmic principles, and implementation of methods, models, or experiments; or (2) *experimental analysis* that interprets empirical results, identifies patterns or trends, and draws evidence-based conclusions. This distinction reflects the diverse objectives of scientific reasoning involved in research analysis.

Analysis Width. Analysis width characterizes the scope of information referenced in the ground-truth analysis. We define four fine-grained classes: (1) analyses with *no references*, which rely solely on the immediate inputs; (2) *internal references*, which draw upon other components within the same document; (3) *external references*, which incorporate information beyond the current document; and (4) *mixed references*, which combine both internal and external sources. This analysis width reflects the increasing breadth of contextual integration required for comprehensive analysis.

Analysis Depth. Analysis depth distinguishes between surface-level summarization and inference-driven analysis. *Shallow* analyses involve direct restatement or aggregation of explicitly stated information, whereas in-depth analyses require implicit reasoning, interpretation, or synthesis that is not directly observable from the input. This analysis depth captures the degree of cognitive and analytical complexity demanded by each task.

C.2.3 Task Curriculum

Of the seven complexity dimensions (§2.3), five are assigned by deterministic rule-based classification over the parsed document structure, and only two, *analysis objective* and *analysis depth*, require MLLM classification (§C.1). Employing Gemini-2.5-Flash (Google Developers Blog, 2025) for *MLLM-assisted classification* and conceptual criteria for *rule-based classification*, we perform fine-grained task curriculum classification (Tab. 5) on ANABENCH according to the data complexity and analysis complexity of each instance:

- **Data Type:** Rule-based classification, based on the data type of the input data (§C.2.1).
- **Data Format:** Rule-based classification, based on the data format of the input data(§C.2.1).

- **Data Source:** Rule-based classification, based on the data source of the input data (§C.2.1).
- **Data Domain:** Rule-based classification, based on the domain of the task (§C.2.1).
- **Analysis Objective:** MLLM classification, according to the analysis objective of the task (§C.2.2).
- **Analysis Width:** Rule-based classification, according to the references and citations included in the task (§C.2.2).
- **Analysis Depth:** MLLM classification, according to the analysis level of the task (§C.2.2).

C.3 Benchmark Statistics

Following the major challenges (Fig. 1) and failure errors (Fig. 6) identified through preliminary exploration (§A), we construct ANABENCH across different data types, data formats, data sources, and data domains, thereby contributing to different data and analysis complexities (§C.2).

Starting from 9 broad scientific domains, we systematically delve into each domain to compile scientific analysis samples across 170 fine-grained subdomains (Tab. 6). Additionally, we define the task curriculum of ANABENCH according to the seven complexity levels of each task (§C.2.3).

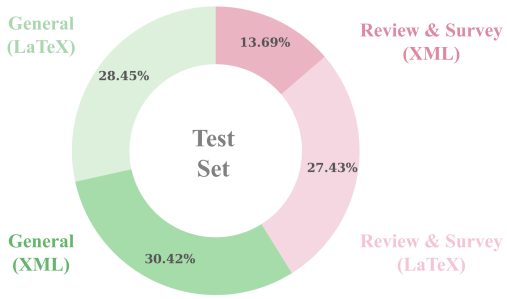


Figure 16: **Evaluation Set Data Distribution.**

Through quantitative analysis, we summarize the statistics of our benchmark in Fig.15. Our dataset construction follows real-world distributions of data types, formats, and domains, preserving both natural data distribution imbalance and inherent complexity curriculum (Fig. 2 & §2).

C.4 Data Preprocessing

For training and evaluation, we implement additional filtering and downsampling to ensure both data quality and computation efficiency.

Data Filtering. We add the first length filtering step to exclude samples with overly short ground-truth analysis to ensure the effectiveness of model learning, and then apply the second filtering step to filter out samples with overlong contexts or overlong analysis to preserve efficient training with affordable computational resources.

Data Downsampling. The large computation resources required by RL are intensified by the long-context inputs and long-analysis output. To this end, we further downsample a subset of the filtered dataset through random sampling with more narrowed thresholds.

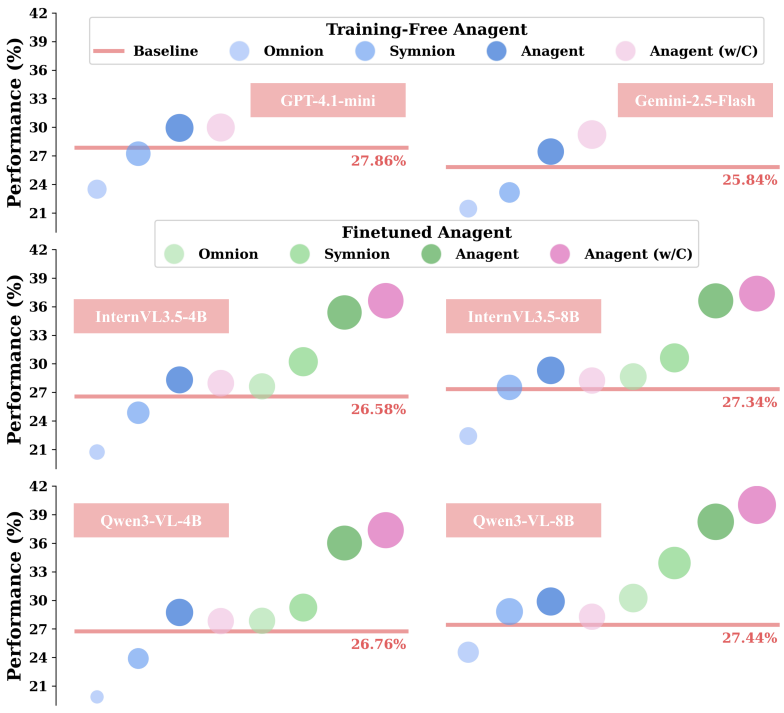


Figure 17: **Ablation Studies On ANAGENT Variants.**

Training Data. During the SFT stage, we leverage the filtered dataset with 31,350 samples. To explore the effects of data size, we extend our ablation studies to cover several variations of data sizes (Tab. 7). During RL optimization, we further downsample a smaller subset tailored for each agent's sub-goals (§3.4 & D.4).

Evaluation Data. Aiming for comprehensive evaluation, we randomly downsample the test set from ANABENCH in a dimension-wise manner to cover all challenge dimensions (Fig. 1). As such, our test set (Fig. 16) consists of 7,319 test samples across different data types, formats, modalities, sources, and domains, with tasks in varying analysis depths, widths, and writing categories. To mitigate data contamination (§C), our evaluation set consists of instances derived exclusively from source papers published in 2025 (Fig. 11). We visualize the domain distribution of the evaluation dataset in Fig. 14.

D ANAGENT: Multi-Agent Collaborative Scientific Analysis

D.1 ANAGENT Variants

We summarize five variants of ANAGENT in Tab. 9, which differ in their agent components and tool availability. To investigate the contribution of these design choices to scientific analysis generation, we conduct additional ablation studies using the same evaluation dataset and metrics (§E). As shown in Fig. 17, agents' performance exhibits a clear ordering: ANAGENT consistently outperforms *Symnion* variants, which in turn outperform *Omnion* variants. In the training-free setting, *Omnion* underperforms all baselines, indicating that equipping a standalone SOLVER with diverse tools is insufficient and can even overwhelm the agent, preventing effective interaction with task contexts and environments. *Symnion* alleviates this issue to some extent by introducing EXPERT to assist with tool invocation and contextual interaction. While this generally improves performance over baselines, results remain unstable and occasionally fall below baseline levels, suggesting that tool orchestration alone is not enough for robust scientific reasoning. In contrast, ANAGENT integrates PLANNER for high-level task decomposition, EXPERT for contextual and domain-specific information retrieval, SOLVER for context-aware problem solving, and CRITIC for reflection and refinement. This structured multi-agent design leads to consistently superior performance across all MLLM backbones, demonstrating the significance of explicit planning, contextual grounding, and context-aware problem-solving in complex scientific analysis. Comparing training-free and finetuned variants, all finetuned models achieve substantial gains over their training-free counterparts. Notably, even finetuned *Omnion* surpasses baseline methods, highlighting the critical role of targeted finetuning in enabling agents to effectively leverage tools and interact with scientific environments. Overall, these results underscore the complementary benefits of principled multi-agent architecture design and task-specific finetuning for reliable and high-performing scientific reasoning systems.

Table 9: **ANAGENT Variants.** Overview of ANAGENT variants.

ANAGENT Variants		Tools	Agent Component			
			PLANNER	EXPERT	SOLVER	CRITIC
Baselines	–	✗	✗	✗	✗	✗
Omnion	–	✓	✗	✗	✓	✗
Symnion	–	✓	✗	✓	✓	✗
ANAGENT	–	✓	✓	✓	✓	✗
	w/ CRITIC	✓	✓	✓	✓	✓

D.2 Multi-Agent Collaboration

Implementing collaborative scientific analysis, ANAGENT decomposes the end-to-end reasoning and writing process into four specialized agents with complementary roles (§3). As inspired by our preliminary exploration (§A.2), rather than relying on a single monolithic agent to simultaneously plan, retrieve, reason, and reflect, our design explicitly separates these responsibilities to reduce error propagation and

encourage iterative refinement (§1). Accordingly, each agent is guided by a tailored prompt that defines its task, objective, and interaction protocol, enabling structured collaboration across different stages of scientific analysis (§3.1). Fig. 3 illustrates our proposed multi-agent scientific analysis workflow (with example data from (Guo et al., 2025)), as motivated by Fig. 9 (with example data from (Wang et al., 2025b)).

PLANNER is responsible for high-level task decomposition and strategic guidance. Given a scientific analysis task, **PLANNER** identifies the core objectives, determines the required analytical depth and breadth, and outlines a step-by-step plan to guide downstream agents. By explicitly structuring the reasoning process before content generation, **PLANNER** mitigates common failures such as incomplete coverage, misaligned objectives, and shallow analysis. The task prompt for **PLANNER** is shown in Fig. 45.

EXPERT focuses on task-oriented exploration, retrieval, and domain-specific clarification. Under the guidance of **PLANNER**, **EXPERT** gathers relevant contextual knowledge, resolves ambiguities in terminology or context, and provides structured evidence or references across diverse modalities, formats, and domains. This separation allows **ANAGENT** to better handle domain-specific nuances and reduces errors arising from insufficient or incorrect contextual understanding. The detailed task prompt for **EXPERT** is presented in Fig. 46.

SOLVER performs the core scientific analysis writing. Guided by the problem-solving plan generated by **PLANNER** and the supporting information provided by **EXPERT**, **SOLVER** synthesizes semantically coherent, logically structured, and scientifically grounded analysis. **SOLVER**'s prompt (Fig. 47) emphasizes systematic integration of provided information and rigorous scientific reasoning, maintaining close alignment between retrieved knowledge and analysis objectives.

CRITIC conducts self-reflection and post-hoc verification of the generated analysis. By systematically reviewing the **SOLVER**'s solution, it identifies overlooked observations and findings, logical inconsistencies and reasoning flaws, groundless claims and hallucinated contents, as well as formatting errors and analysis inaccuracies, proposing targeted revisions for enhanced scientific analysis. By explicitly modeling critique and reflection as an explicit step, **ANAGENT** improves analytical reliability and scientific rigor. By applying the five-dimensional evaluation protocol (§A.2 & E.2), **CRITIC**'s prompt defines both evaluation criteria and critique objectives for high-quality scientific analysis (Fig. 48).

D.3 Scientific Toolkits

In scientific research, human researchers rely on a diverse set of skills to observe, analyze, and reason over complex scientific materials. These skills include *reading and comprehending scientific documents*, *retrieving targeted knowledge*, *searching for related literature and information*, *analyzing multimodal data with dissimilar structures and formats*, and *performing interactive or computational explorations*. To develop AI agents into AI scientists, we take inspiration from human scientific research, equipping AI agents with five specialized toolkits (Tab. 13) to support scientific reasoning and analysis.

D.4 Modular Optimization

To enhance the scientific analysis performance of **ANAGENT**, we implement modular optimization, training each agent on its specialized task using GRPO (Shao et al., 2024).

Concretely, by constructing four RL training datasets respectively designed for four agents' specialized tasks, each agent $a \in \{\textbf{PLANNER}, \textbf{EXPERT}, \textbf{SOLVER}, \textbf{CRITIC}\}$ is optimized through a specialized reward function tailored to its own task objectives. As shown in Eq. 8, the reward for agent a is decomposed into weighted components with $\sum_m \lambda_{a,m} = 1$ (§3.4).

D.4.1 PLANNER Optimization

PLANNER is optimized through selecting optimal problem-solving strategies from multiple candidate plans for each scientific analysis task We formulate this as a multi-choice preference selection task, where the agent is asked to identify the most effective decomposition strategy for the given analysis task.

Data. We construct the PLANNER's RL preference dataset by generating different versions of planning strategies for each scientific analysis instance using two models, where `Qwen3-VL-8B` acts as the baseline PLANNER and `Gemini-2.5-Flash` as the reference PLANNER. For each input (x, s, q), we collect multiple candidate plans from both models. We then filter the generated problem-solving plans by executing the complete ANAGENT workflow with each reference plan and retaining only those `Gemini-2.5-Flash` plans that can improve end-to-end performance over the `Qwen3-VL-8B` baseline:

$$\forall \Delta S \in \{\Delta S_{\text{LEX}}(y^*, y),\ \Delta S_{\text{SEM}}(y^*, y),\ \Delta S_{\text{AVG}}(y^*, y)\}\ (\Delta S > 0) \tag{9}$$

As shown in Eq. 9, $S_{\text{LEX}}(y^*, y)$ (Eq. 25), $S_{\text{SEM}}(y^*, y)$ (Eq. 26), and $S_{\text{AVG}}(y^*, y)\}$ are the final accuracy scores in our rule-based evaluation (§E.1). With `Qwen3-VL-8B` plans serving as baseline options, the performance-validated `Gemini-2.5-Flash` plans are designated as ground-truth preferred choices. This approach ensures PLANNER to learn to select problem-solving strategies that lead to measurably better scientific analysis quality.

Reward. Optimized to make strategic decisions from predefined option sets, PLANNER's reward function combines format compliance and answer accuracy:

$$R_{\text{PLANNER}} = \lambda_{\text{Pf}} \cdot r_{\text{Pf}}(z_a) + \lambda_{\text{Pacc}} \cdot r_{\text{Pacc}}(z_a, z_a^*) \tag{10}$$

where $r_{\text{Pf}}(z_a)$ validates structural correctness, and $r_{\text{Pacc}}(z_a, z_a^*)$ measures answer accuracy. For multi-choice selections, accuracy is computed as:

$$r_{\text{Pacc}}(z_a, z_a^*) = \begin{cases} 1 & \text{if } \mathcal{O}(z_a) = \mathcal{O}(z_a^*) \\ \frac{2P_{\mathcal{O}}R_{\mathcal{O}}}{P_{\mathcal{O}}+R_{\mathcal{O}}} & \text{otherwise} \end{cases} \tag{11}$$

where $\mathcal{O}(\cdot)$ extracts the set of selected options, $P_{\mathcal{O}} = \frac{|\mathcal{O}(z_a) \cap \mathcal{O}(z_a^*)|}{|\mathcal{O}(z_a)|}$ is precision, and $R_{\mathcal{O}} = \frac{|\mathcal{O}(z_a) \cap \mathcal{O}(z_a^*)|}{|\mathcal{O}(z_a^*)|}$ is recall.

D.4.2 EXPERT Optimization

As we observe notable inaccurate and hallucinated tool calls in domain-specific knowledge retrieval failures, EXPERT is optimized for task-oriented tool calling and execution through GRPO. Unlike general-purpose tool-use benchmarks, our optimization focuses on domain-specific information retrieval tools tailored to scientific analysis tasks (Tab. 13), with the specialized RL dataset built on downsampled scientific analysis instances (Tab. 8).

Data. We construct EXPERT's RL dataset by pairing each candidate tool with tool-specific queries and formats. Each training instance consists of tool prefix, current knowledge state $\mathcal{K}_{i-1}$, and the ground-truth tool invocation z^*_{expert}. The dataset emphasizes correct tool selection, proper parameter formatting, and contextually appropriate query aligned with the specific analysis objectives.

Reward. EXPERT performs tool-based information retrieval. Its reward evaluates both format validity and tool execution correctness:

$$R_{\text{EXPERT}} = \lambda_{\text{Ef}} \cdot r_{\text{Ef}}(z_a) + \lambda_{\text{Eacc}} \cdot r_{\text{Eacc}}(z_a, z_a^*) \tag{12}$$

where $r_{\text{Ef}}(z_a)$ validates that the action z_a conforms to the expected tool specification query and format. The accuracy component r_{Eacc} verifies tool selection and parameter correctness through:

$$r_{\text{Eacc}}(z_a, z_a^*) = \mathbb{I}[\mathcal{T}(z_a) = \mathcal{T}(z_a^*)] \cdot \rho(z_a, z_a^*) \tag{13}$$

where $\mathcal{T}(\cdot)$ extracts the tool type, $\mathbb{I}(\cdot)$ is the indicator function, and $\rho(z_a, z_a^*)$ measures parameter correctness using tool-specific validation rules.

D.4.3 Solver

SOLVER is optimized to generate high-quality scientific analysis. Its RL optimization objective, therefore, focuses on synthesizing retrieved knowledge with input data to produce coherent, accurate, and contextually appropriate analysis writing.

Data. SOLVER's RL dataset consists of instances $(x, s, q, \mathcal{K}_n, y^*)$, where y^* represents the ground-truth analysis. We use SciBERT (Beltagy et al., 2019) as the reward model to evaluate semantic quality, guiding SOLVER's scientific analysis generation with improved scientific accuracy, terminology usage, and writing style. SOLVER's reward function incorporates format compliance, length appropriateness (§G.5), and semantic similarity to guide SOLVER toward generating well-structured and high-quality analysis.

Reward. As discussed above, SOLVER's reward combines format compliance, length appropriateness, and semantic quality (Eq. 14):

$$R_{\text{SOLVER}} = \lambda_{\text{Sf}} \cdot r_{\text{Sf}}(z_a) + \lambda_{\text{Slen}} \cdot r_{\text{Slen}}(z_a, z_a^*) + \lambda_{\text{Sacc}} \cdot r_{\text{Sacc}}(z_a, z_a^*) \tag{14}$$

where $r_{\text{Slen}}(z_a, z_a^*)$ penalizes outputs with overlong or overshort length relative to the ground truth (Eq. 15):

$$r_{\text{Slen}}(z_a, z_a^*) = \mathbb{I}[0.5|z_a^*| \leq |z_a| \leq 1.5|z_a^*|] \tag{15}$$

and $r_{\text{Sacc}}(z_a, z_a^*)$ computes semantic similarity using SciBERT token-level embeddings (Eq. 16):

$$r_{\text{Sacc}}(z_a, z_a^*) = \frac{2P_{\text{emb}}R_{\text{emb}}}{P_{\text{emb}} + R_{\text{emb}}} \tag{16}$$

where P_{emb} and R_{emb} are computed from the maximum token-level cosine similarities between SciBERT embeddings of z_a and z_a^*.

D.4.4 CRITIC Optimization

Similar to PLANNER, CRITIC is optimized through multi-choice solution preference selection, but focuses on assessing analysis quality and providing constructive feedback for analysis refinement.

Data. We follow the same data construction methodology as PLANNER (§D.4.1). Concretely, for each scientific analysis y_i generated by SOLVER, we collect critique feedback from both Qwen2-VL-8B (serving as baseline CRITIC) and Gemini-2.5-Flash (serving as reference CRITIC). Following PLANNER data filtering, we filter the feedback by evaluating whether applying the suggested revisions leads to improved analysis quality (Eq. 9). As such, only those Gemini-2.5-Flash critiques that can enhance SOLVER's analysis writing are retained as ground-truth preferred feedback options, while Qwen2-VL-8B critiques serve as baseline options. This helps CRITIC to learn to identify key quality deficiencies and provide actionable improvement suggestions across multiple evaluation dimensions (Fig. 48).

Reward. Similar to PLANNER, CRITIC's reward also consists of format compliance and answer accuracy (Eq. 17):

$$R_{\text{CRITIC}} = \lambda_{\text{Cf}} \cdot r_{\text{Cf}}(z_a) + \lambda_{\text{Cacc}} \cdot r_{\text{Cacc}}(z_a, z_a^*) \tag{17}$$

where $r_{\text{Cf}}(z_a)$ validates evaluation formatting correctness, and $r_{\text{Cacc}}(z_a, z_a^*)$ measures answer accuracy. Same to PLANNER, answer accuracy is computed as (Eq. 18):

$$r_{\text{Cacc}}(z_a, z_a^*) = \begin{cases} 1 & \text{if } \mathcal{O}(z_a) = \mathcal{O}(z_a^*) \\ \frac{2P_{\mathcal{O}}R_{\mathcal{O}}}{P_{\mathcal{O}}+R_{\mathcal{O}}} & \text{otherwise} \end{cases} \tag{18}$$

E Scientific Analysis Evaluation

E.1 Rule-Based Evaluation

We design rule-based evaluation in both lexical and semantic dimensions. Lexical evaluation consists of ROUGE-L (Eq. 19) (Lin, 2004), BLEU (Eq. 20) (Papineni et al., 2002), and word overlap (Eq. 21) metrics. For semantic evaluation, we employ cosine similarity (Eq. 22), SciBERT (Eq. 23) (Beltagy et al., 2019), and METEOR (Eq. 24) (Banerjee and Lavie, 2005) metrics to calculate the semantic assessment scores. In addition to the lexical score S_{LEX} (Eq. 25) and semantic score S_{SEM} (Eq. 26) that are calculated as the mean of their three metrics, respectively. The overall score S_{AVG} (Eq. 27) is averaged across all six metrics.

Concretely, given the model-generated analysis y and ground-truth analysis y^*, the lexical and semantic evaluation scores are calculated as follows:

ROUGE-L measures the longest common subsequence (LCS) between y and y^*:

$$\text{ROUGE-L}(y^*, y) = \frac{\text{LCS}(y^*, y)}{\max(|y^*|, |y|)} \tag{19}$$

where $\text{LCS}(y^*, y)$ computes the length of the longest common subsequence, and $|\cdot|$ denotes sequence length.

BLEU calculates n-gram precision with a brevity penalty (BP):

$$\text{BLEU}(y^*, y) = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{20}$$

where p_n is the n-gram precision, w_n is the weight for each n-gram (typically $w_n = 1/N$), and BP is the brevity penalty to penalize short predictions.

Word Overlap measures the Jaccard similarity between word sets:

$$\text{WORD}(y^*, y) = \frac{|W(y^*) \cap W(y)|}{|W(y^*) \cup W(y)|} \tag{21}$$

where $W(\cdot)$ extracts the set of unique words from a text.

Cosine Similarity measures the angle between sentence embeddings:

$$\text{COSINE}(y^*, y) = \frac{\mathbf{e}(y^*) \cdot \mathbf{e}(y)}{\|\mathbf{e}(y^*)\| \|\mathbf{e}(y)\|} \tag{22}$$

where $\mathbf{e}(\cdot)$ is a sentence embedding function that maps text to a dense vector representation.

SciBERT Score computes token-level semantic similarity using SciBERT embeddings:

$$\text{SciBERT}(y^*, y) = \frac{2 \cdot P_{\text{SciBERT}} \cdot R_{\text{SciBERT}}}{P_{\text{SciBERT}} + R_{\text{SciBERT}}} \tag{23}$$

where P_{SciBERT} and R_{SciBERT} are precision and recall computed from token-level cosine similarities between SciBERT embeddings of y^* and y.

METEOR evaluates based on unigram matching with synonym and paraphrase support:

$$\text{METEOR}(y^*, y) = F_{\text{mean}} \cdot (1 - \text{PEN}) \tag{24}$$

where F_{mean} is the harmonic mean of unigram precision and recall, and PEN is a fragmentation penalty based on chunk count.

Final Scores include the lexical score S_{LEX} (Eq. 25) and semantic score S_{SEM} (Eq. 26) that are computed as:

$$S_{\text{LEX}}(y^*, y) = \frac{\text{ROUGE-L} + \text{BLEU} + \text{WORD}}{3} \tag{25}$$

$$S_{\text{SEM}}(y^*, y) = \frac{\text{COSINE} + \text{SciBERT} + \text{METEOR}}{3} \tag{26}$$

The overall evaluation score S_{AVG} (Eq. 27) is the average across all six metrics:

$$S_{\text{AVG}}(y^*, y) = \frac{1}{6} \sum_{i=1}^{6} m_i(y^*, y) \tag{27}$$

where m_i denotes each of the six individual metrics.

E.2 Five-Dimensional Evaluation Protocol

We apply our five-dimensional evaluation protocol to both MLLM-As-Judge and human expert assessment (§2.4 & A.2). Fig. 44 shows the evaluation prompt for assessing scientific analysis quality. Accordingly, we prompt and finetune CRITIC for self- reflection and correction with the same five criteria (Fig. 48).

E.3 Performance Difference

To quantify the performance differences between two methods, we calculate *absolute* performance difference ($\Delta_{abs}\%$) and *relative* performance difference ($\Delta_{rel}\%$) across metrics.

Absolute Performance Difference. The absolute performance difference Δ_{abs} (Eq. 28) measures the direct performance gap between our method and the baseline:

$$\Delta_{abs} = (S_{ours} - S_{baseline})\% \tag{28}$$

where S denotes a metric score (*e.g.*, SciBERT, BLEU, etc.), and the Δ_{abs} result is expressed in percentage points.

Relative Performance Difference. The relative performance difference Δ_{rel} (Eq. 29) measures the proportional improvement or degradation with respect to the baseline:

$$\Delta_{rel} = \frac{S_{ours} - S_{baseline}}{S_{baseline}} \times 100\% \tag{29}$$

where S denotes a metric score, and the result represents the percentage change relative to the baseline performance.

F Implementation Details

To complement the implementation details presented in §4, we further summarize our experiment configurations and computation overhead in Tabs. 11-10. As can be seen in Tab. 10, RL demands significantly more computation resources for smaller MLLMs with model sizes of 3B–4B parameters. Therefore, we finetuned only Qwen2.5-VL-3B and Qwen3-VL-4B models for computation efficiency. Their performance (Tab. 2) further substantiates the effectiveness of combining supervised finetuning and reinforcement learning, delivering remarkable cumulative optimization benefits even for small-size MLLMs.

Table 10: **Computation Overhead.** This table summarizes the computation overhead during training and evaluation. n_{MLLM} denotes the number of different backbone MLLMs.

	n_{MLLM}			
	1	**2**	**3**	**4**
Evaluation	A100 80GB × 1	A100 80GB × 2	A100 80GB × 4	A100 80GB × 4
SFT	A100 80GB × 1	–	–	–
RL	A100 80GB × 8	–	–	–

Table 11: **Implementation Details.** This table summarizes the key configuration settings during training and evaluation. M_p, M_e, M_s, and M_c denote the maximum number of turns allowed for Planner, Expert, Solver, and Critic, respectively. The maximum depth of context searching tools is set to $k = 1$ across all settings. The maximum and minimum pixel values are denoted by max_{pixel} and min_{pixel}. The learning rate is represented by lr. $n_{rollout}$ denotes the number of rollouts during RL, and n_{epoch} denotes the number of training epochs.

	Agent				**Configuration**					
	PLANNER	EXPERT	SOLVER	CRITIC	(All Agents)					
	M_p	M_e	M_s	M_c	k	max_{pixel}	min_{pixel}	lr	$n_{rollout}$	n_{epoch}
Evaluation	1	5	2	1	1	1024 × 1024	128 × 128	–	–	–
SFT	1	1	1	1	1	512 × 512	128 × 128	1×10^{-4}	–	1
RL	1	1	1	1	1	512 × 512	128 × 128	1×10^{-6}	4	1

Table 12: **MLLM-As-Judge Evaluation.** Evaluation of scientific table & figure analysis across baselines, training-free AGENT, and finetuned AGENT, through five-dimensional evaluation protocol (§E.2). Compared with baselines, ***relative performance differences*** (Eq. 29) are shown as *positive* ↑ $\Delta_{rel}\%$ or *negative* ↓ $\Delta_{rel}\%$.

Model	**Size**	**Five-Dimensional Evaluation (%)**					**Overall Accuracy (%)**
		S_{ACC}	$S_{COMPLETE}$	S_{FORMAT}	$S_{CLARITY}$	S_{FAITH}	S_{MLLM}
Baselines							
InternVL-3.5	**4B**	42.00	27.17	69.67	64.33	55.33	51.70
	8B	52.67	30.50	71.67	68.50	60.83	56.83
Qwen2.5-VL	**3B**	43.17	24.88	64.50	64.83	56.00	50.68
	7B	50.83	**32.17**	73.50	67.50	60.33	56.87
Qwen3-VL	**4B**	46.50	28.50	68.33	67.17	57.00	53.50
	8B	**52.83**	31.17	**74.17**	**69.50**	**61.17**	**57.77**
AGENT (Training-Free)							
InternVL-3.5	**4B**	44.17	30.83	70.33	66.00	59.50	↑**4.78%**
	8B	**55.67**	35.17	71.83	71.67	**65.50**	↑**5.53%**
Qwen2.5-VL	**3B**	47.67	25.13	67.83	66.17	56.50	↑**3.91%**
	7B	52.17	37.50	74.17	69.83	63.00	↑**4.33%**
Qwen3-VL	**4B**	51.67	31.83	68.50	67.50	60.83	↑**4.80%**
	8B	54.17	**39.50**	**77.50**	**72.83**	64.50	↑**6.80%**
AGENT (SFT)							
InternVL-3.5	**4B**	50.83	47.50	77.83	69.33	64.67	↑**19.98%**
	8B	56.33	52.33	79.00	**75.17**	72.17	↑**17.90%**
Qwen2.5-VL	**3B**	50.33	48.33	76.50	71.17	65.67	↑**23.13%**
	7B	54.33	53.67	80.83	73.00	**74.50**	↑**18.29%**
Qwen3-VL	**4B**	53.50	50.17	76.17	72.17	65.33	↑**18.64%**
	8B	**57.17**	**54.83**	**83.83**	74.67	73.83	↑**19.21%**
AGENT (SFT + RL)							
Qwen2.5-VL	**3B**	54.00	55.17	78.83	72.17	67.33	↑**29.24%**
Qwen3-VL	**4B**	**56.50**	**57.17**	**79.33**	**73.33**	**68.50**	↑**25.18%**

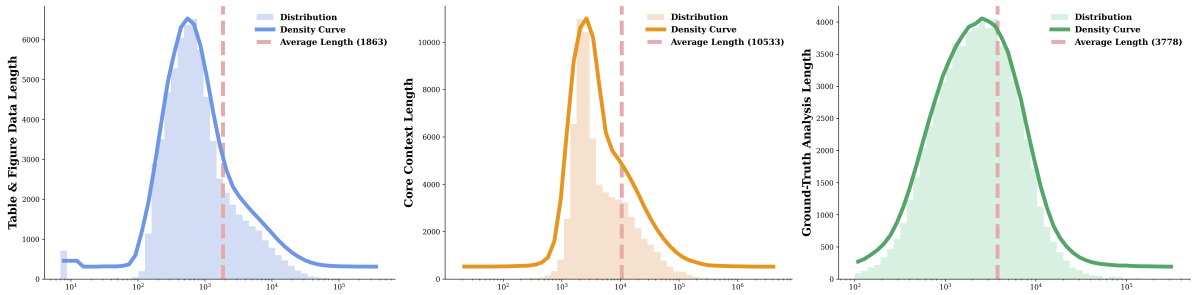


Figure 18: **Length Distribution.** Visualization of Data Length in ANABENCH. (1) Left: Input data length. (2) Center: Core context length. (3) Right: Ground-truth analysis length.

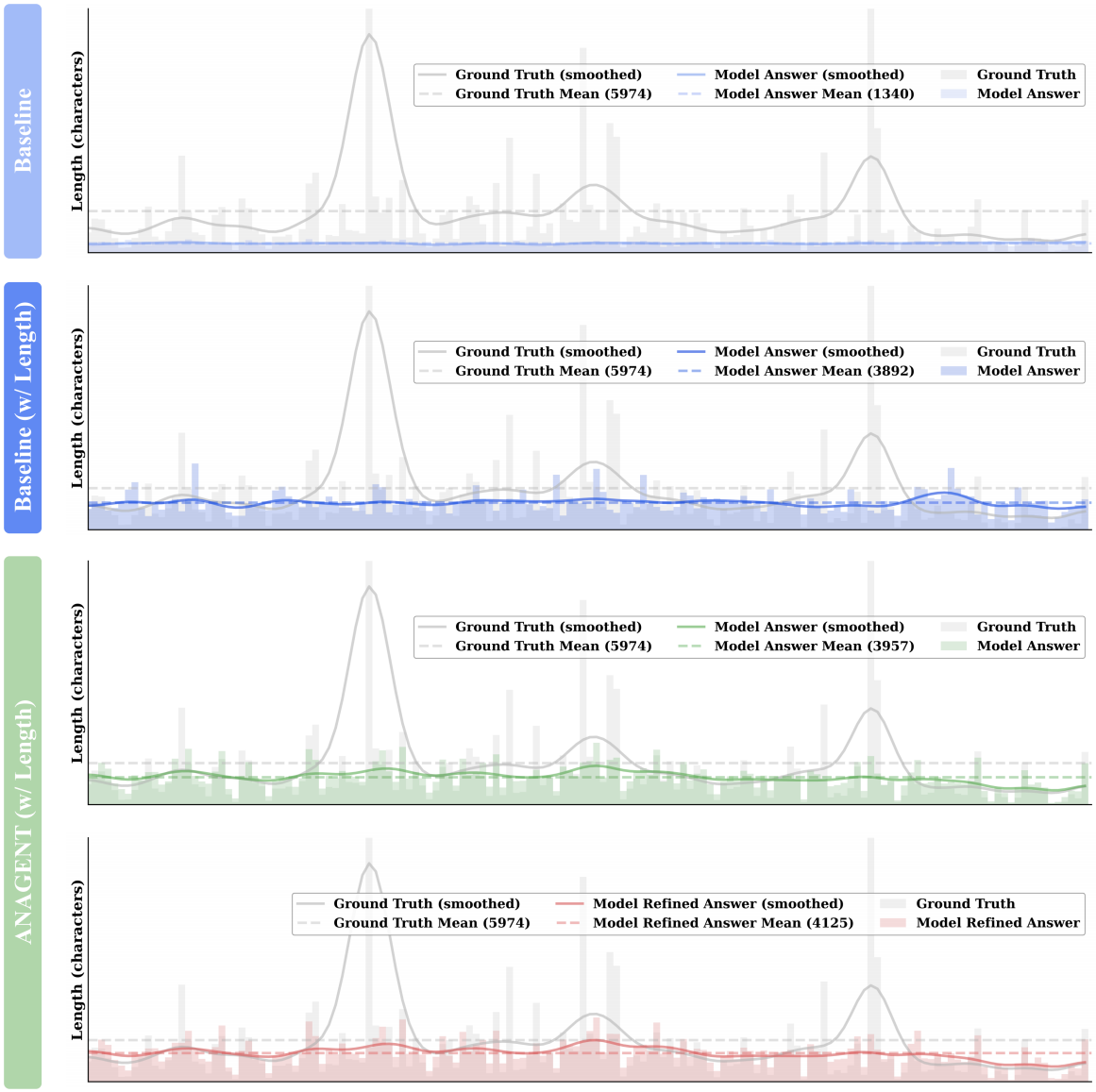


Figure 19: **Length Distribution of GPT-4.1-mini.** Using GPT-4.1-mini as the base model, this figure visualizes the length distribution of 150 randomly sampled agent-generated analysis versus ground-truth analysis length.

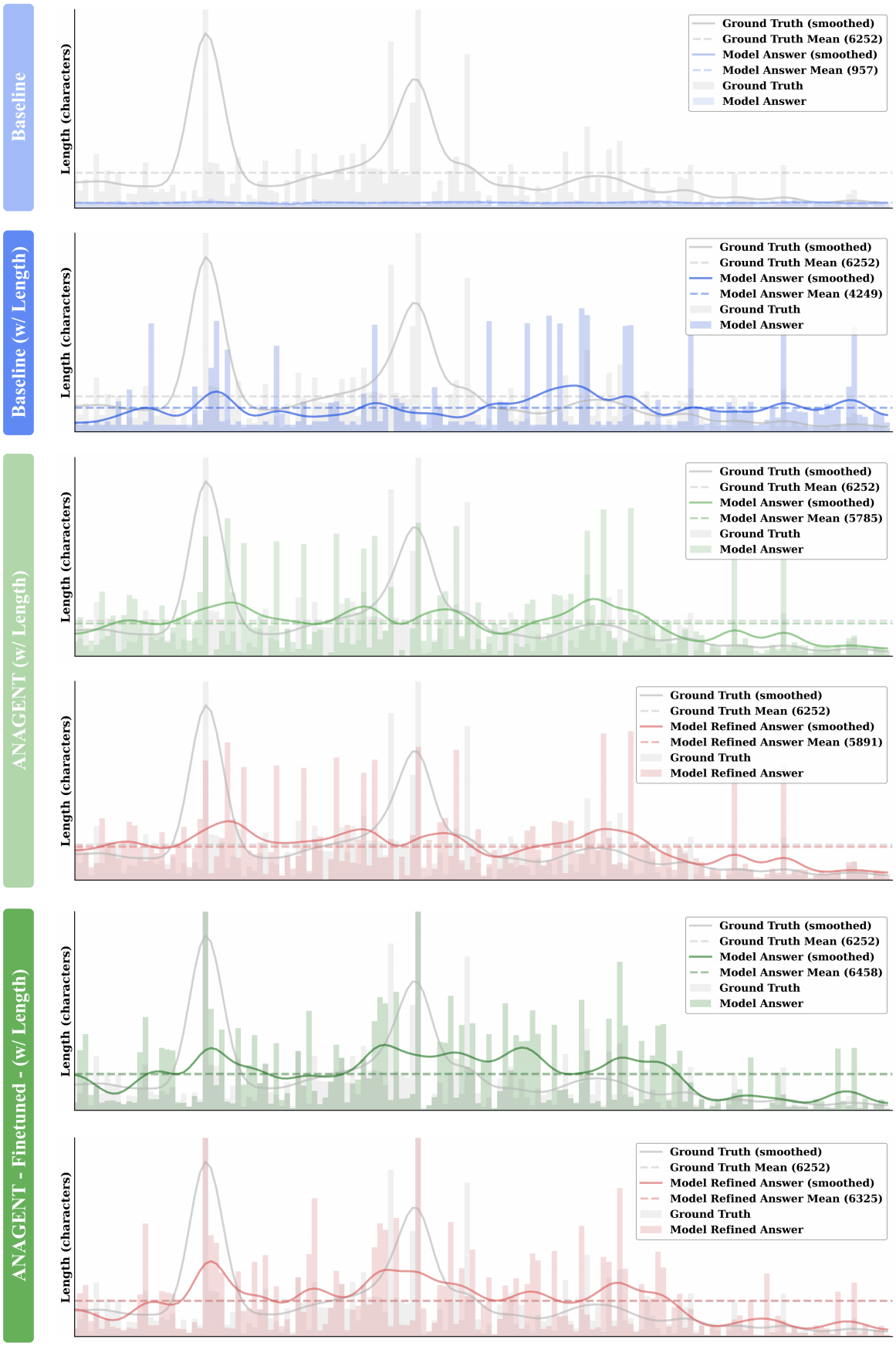


Figure 20: **Length Distribution of Qwen3-VL-8B.** Leveraging Qwen3-VL-8B, this figure visualizes the length distribution of 150 randomly sampled agent-generated analysis versus ground-truth analysis length.

Table 13: **Five Specialized Toolkits of ANAGENT.** To enhance specialized scientific research capabilities, ANAGENT is equipped with five complementary toolkits that respectively support document comprehension, knowledge retrieval, targeted searching, multimodal understanding, and autonomous coding exploration.

Specialized Tools	Supported Inputs & Formats	Functions
Document Toolkit		
Online Fetcher	URL (arXiv, Semantic Scholar, PubMed, general web), Source Type & Preferred Format (HTML, LaTeX, XML, PDF), DOI, Paper ID, Search Query	Fetch and parse documents from online sources, including arXiv papers (LaTeX/PDF), Semantic Scholar (search/metadata), PubMed Central (XML), and general URLs (HTML/PDF/XML)
PDF Parser	URL, Local Path, Bytes	Parse PDF documents to extract text, metadata, hierarchical contents with optional image saving
XML Parser	URL, Local Path, XML Strings, Bytes	Parse XML documents and extract metadata, hierarchical contents with configurable detail levels
Knowledge Toolkit		
Abstract Collector	Title, Author, Year, URL, DOI, arXiv ID, PMID, Local Path	Search and extract paper abstracts from multiple academic sources (e.g., arXiv, PubMed, Semantic Scholar, CrossRef)
Information Localizer	Search String (e.g., keyword, phrase, caption, title, equation), URL, Local Path	Localize and extract the complete section (broad meaning, e.g., section, subsection, figure, table, etc.) according to the input query
Context Finder	Search String (e.g., keyword, phrase, caption, title, equation), URL, Local Path	Localize a search string in a document and extract all bidirectional citation contexts with multi-level traversal support
Section Extractor	Section Identifier (e.g., keywords, numbers), URL	Extract the complete sections (broad meaning, e.g., section, subsection, figure, table, etc.) according to the input query
Search Toolkit		
arXiv Searcher	URL, Title, Keywords, Author, Category, Sort, Max Results, and Other arXiv Search Queries and Filters	Search and retrieve arXiv preprint scientific literature across diverse domains
PebMed Searcher	URL, Title, Keywords, Author, Time Range, PMID, Sort, Time Range, Max Results, and Other PubMed Search Queries and Filters	Search and retrieve biomedical scientific literature in PubMed
Semantic Scholar Searcher	URL, Title, Year, Field, Venue, Citations, Max Results, and Other Semantic Scholar Search Queries and Filters	Search and retrieve scientific literature via Semantic Scholar API
Web Searcher	Search String, Keywords, Temporal Filters, Max Results, Date, Language, Search Level, and Other Web Search Queries and Filters	Search and retrieve web information via Google Search API
Wikipedia Searcher	Search String, Title, Keywords, Search Mode, and Other Wikipedia Search Queries and Filters	Search and retrieve Wikipedia concepts, terminologies, and articles with intelligent routing across five MediaWiki APIs
Vision Toolkit		
OCR Extractor	Image Source, Language, Bounding Boxes, Threshold	Enhance visual perception through OCR
Figure Parser	Image Source, Query, Contexts	Parse and extract visual information from scientific figures
Image Analyzer	Image Source, Query, Focus, Detail Level, Contexts	Analyze images with interactive query-driven exploration
Sandbox Toolkit		
Sandbox Explorer	Python Code, Environment Dependencies	Execute Python code in isolated Docker/Apptainer sandbox with package installation and task data access for autonomous tool development and task exploration

G A Deeper Dive Beneath the Results

G.1 MLLM-As-Judge Evaluation On Scientific Analysis

Leveraging `GPT-4.1-mini` (OpenAI, 2025) as the MLLM judger, we conduct MLLM-as-judge evaluation following the five-dimensional evaluation protocol (§E.2). Due to cost constraints (approximately $\tilde{\$}45$/100 instances), for each agent, we evaluate performance on a randomly sampled subset of 300 instances drawn from the complete evaluation set (§C.4). Results are summarized in Tab. 12. Across five evaluated dimensions and the averaged S_{MLLM}, results demonstrate a monotonic improvement from baselines to training-free **ANAGENT**, with the finetuned version achieving the highest performance. Comparing between **ANAGENT** finetuned through SFT and the combination of SFT+RL, the latter **ANAGENT** consistently outperforms SFT-only counterparts, revealing the significance of specialized ability development in multi-agent systems. Moreover, while the absolute evaluation scores differ, our MLLM-as-judge assessment yields performance trends that are consistent with those observed under rule-based evaluation (Tabs. 1-2). **ANAGENT** that perform well on rule-based metrics also achieve higher rankings under MLLM-based judgment, supporting the use of rule-based evaluation and MLLM-as-judge for scientific analysis assessment.

G.2 Specialized RL as a Double-Edged Sword for Enhancing Scientific Table & Figure Analysis

Trained on tailored datasets (§3.4), agents benefit substantially from specialized RL optimization, achieving notable performance improvements (Tab. 2). In particular, RL finetuning yields consistent gains ($\Delta_{\text{AVG}} \geq 10.12\%$), while combining SFT with RL further amplifies performance, achieving even larger gains ($\Delta_{\text{AVG}} \geq 33.10$). Nevertheless, these gains come at a high computational cost. As shown in Tab. 10, specialized RL introduces substantially higher computation overhead as compared to SFT alone. When jointly considering performance optimization and computation efficiency, SFT demonstrates a more favorable trade-off for scientific table & figure analysis tasks that demand long-context understanding and long-output generation (§5.1).

Overall, while RL is highly effective at optimizing specialized agents for specialized tasks, it functions as a double-edged sword: delivering stronger task-specific optimization at the expense of markedly increased computational cost, which may limit its practicality when finetuning larger models in resource-intensive scientific settings.

G.3 How Training Data Recipe Shapes Finetuning Outcomes?

Benefits of Increased Data Size. Fig. 10 shows the performance of **ANAGENT** finetuned on different subsets of the training set (§C.4), revealing the benefits of finetuning with increased data sizes. While Fig. 10(a) isolates the effects of data size by comparing finetuning on a 30K-instance subset against the full dataset, Fig. 10(b)-(d) evaluate subsets constructed along different data types, domains, and formats, respectively. As shown in Fig. 10(a), finetuning on 30K instances leads to consistent performance degradation across all four MLLM agents as compared to training on the full set ($\Delta_{rel} \geq \downarrow 10.27\%$), demonstrating the clear benefits of increased data size. Results in Fig. 10(b)-(d) lend further evidence to this advantage on subsets with varying data sizes and compositions ($\Delta_{rel} \geq \downarrow 9.29\%$).

Domain-Specific Learning Leads To Decreased Generalizability. Fig. 10(b) examines the impact of domain-specific finetuning on agent performance by comparing models trained exclusively on computer science (CS), biology, and a mixture of all nine domains. Across all four MLLM agents, training on the full set of nine domains consistently yields the strongest performance, while domain-specific training leads to notable degradation ($\Delta_{rel} \geq \downarrow 26.55\%$). In particular, agents finetuned solely on biology exhibit the lowest performance ($S_{\text{AVG}} \leq 29.79\%$), followed by those trained only on CS ($S_{\text{AVG}} \leq 31.92\%$). This trend reveals that restricting training data to a single domain substantially limits agents' ability to generalize beyond domain-specific patterns. Moreover, the larger performance gap observed for biology-only (*25 subdomains*) training compared to CS (*40 subdomains*) indicates that narrower or more specialized domains impose stronger constraints on cross-domain transfer, further emphasizing the significance of domain diversity for robust general-purpose scientific reasoning.

Single-Format Learning Limits Cross-Format Adaptability. Fig. 10(c) assesses the effects of data format on agent performance by comparing models finetuned exclusively on LaTeX, exclusively on XML, and on both. Across all four MLLM agents, training on LaTeX ($S_{\text{AVG}} \geq 31.98\%$) consistently outperforms training on XML ($S_{\text{AVG}} \geq 26.13\%$), indicating that LaTeX-based supervision provides more effective signals for scientific reasoning. However, agents trained on either single format exhibit degraded performance compared to those trained on both ($\Delta_{rel} \geq \downarrow 8.85\%$). These results suggest that restricting training to a single data format limits agents' adaptability to heterogeneous input representations, while exposure to multiple formats enhances robustness and cross-format generalization.

Single-Type Learning Limits Multi-Type Adaptability. Fig. 10(d) further analyzes the impact of data type by comparing agents finetuned exclusively on tables, exclusively on figures, and on both. Across all four MLLM agents, models trained only on tables ($S_{\text{AVG}} \geq 28.29\%$) consistently underperform those trained only on figures & tables ($S_{\text{AVG}} \geq 30.57\%$), revealing limited generalization from structured tabular inputs to multimodal reasoning. More importantly, agents trained on either single type exhibit substantial performance degradation relative to those trained on both tables and figures ($\Delta_{rel} \geq \downarrow 9.53\%$), indicating a pronounced loss of adaptability when exposure is restricted to a single data type. These trends highlight that learning from a single data type constrains agents' ability to handle heterogeneous layouts and modalities, while maintaining data heterogeneity in training is critical for enhanced scientific analysis performance.

G.4 Significance of Contextual Information & Domain Knowledge In Scientific Analysis

Contextual information and domain-specific knowledge play a critical role in scientific analysis, supporting both accurate observations and well-grounded inference. To examine the role of contextual information and domain knowledge in scientific analysis, we conduct a comparative evaluation across baselines and **ANAGENT**. Specifically, we consider baselines and training-free **ANAGENT** under two settings: (1) with gold contextual information and domain-specific knowledge included in the input, and (2) with them excluded from the input. All models are evaluated on the same evaluation split of **ANABENCH** (§C.4), ensuring evaluation consistency across experimental conditions.

The provided contexts in setting (1) are gold contexts extracted from highly-relevant components of the source scientific papers of each instance, such as *table and figure captions*, *sections*, *citations*, *etc.*, that both refer to and are referred by the task input ($d = 1$). These contexts represent the most relevant and effectively optimal information for the task, as simulating what human researchers naturally attend to during scientific analysis. They provide a highly informative signal that agents would otherwise search for and extract themselves, often at greater computational cost and with lower accuracy.

Fig. 21 visualizes the quantitative results of this comparison and highlights consistent performance improvements when additional contextual information is provided. Across both baseline methods and **ANAGENT**, enriching the input with supplementary contexts and domain-specific knowledge yields marked gains, with relative improvements $\Delta_{rel} \geq 4.42\%$ on scientific analysis tasks. These improvements are observed consistently across evaluation settings, indicating that the benefit is not model-specific but instead reflects a general reliance on contextual grounding for complex scientific reasoning. In particular, domain-specific context enables more accurate interpretation of technical concepts, experimental setups, and implicit assumptions, which are often under-specified in isolated inputs. The results empirically validate the significance of contextual and domain knowledge in complex scientific analysis.

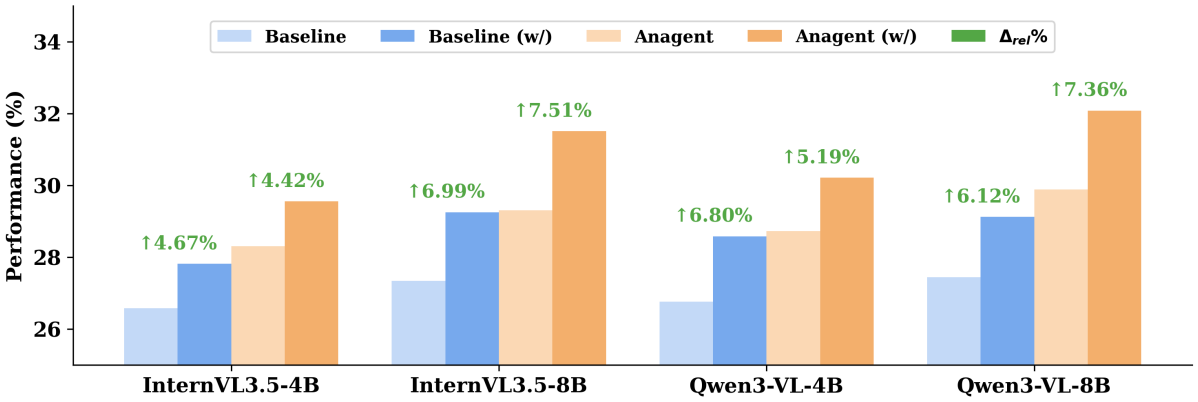


Figure 21: **Ablation Studies On Additional Contexts & Domain Knowledge.** Comparison between with ("w/" in the figure) and without *gold contextual information and domain-specific knowledge*.

G.5 Can AI Agents Deliver What Is Asked?

In *human–AI co-discovery*, AI agents are expected to follow human-specified requirements, especially in terms of the *depth*, *breadth*, and *length* of writing in our scientific analysis tasks. This raises a central question: ***Can AI agents reliably meet stated expectations?***

Motivated by this question, we conduct an ablation study examining the effects of requirement prompting using three types of explicit statements: *analysis length*, *analysis width*, and *analysis depth*. Particularly, each prompt specifies the expected *writing length*, *contextual breadth*, or *analytical depth* of the scientific analysis writing. Relative to the baseline setting in which no requirements are specified, we evaluate agent performance under five conditions: (1) no requirement, (2) length requirement, (3) width requirement, (4) depth requirement, and (5) all three requirements combined. To maintain valid comparisons, we leverage samples with *medium* analysis width and *in-depth* analysis depth as the test set (Fig. 15 & §C.3), while remaining the heterogeneity of the dataset with varying data and analysis complexties (§C.2.1 & §C.2.2).

Comparing between baselines and **ANAGENT** (Fig. 9), results in Fig. 22 demonstrate consistent effects of explicit analysis requirements on MLLM agents' performance. As indicated by the red bars, imposing writing constraints on analysis length leads to a noticeable performance degradation, suggesting current AI agents' lack of sufficient awareness of how to effectively organize and prioritize their reasoning under stated space limitations. In contrast, explicitly specifying requirements on analysis width or depth significantly improves performance, as such guidance helps agents better understand the expected analytical scope and level of detail for each task. Notably, combining analysis length, width, and depth together yields the largest performance gains. This indicates a strong complementary effect among these dimensions, highlighting the importance of holistic analysis constraints in enhancing AI agents' reasoning capabilities and informing the design of future AI research assistants and human-AI co-discovery systems.

To further investigate AI agents' awareness of generation length and the implications of explicitly stated length expectations, we extend our study to compare the lengths of agent-generated analyses with those of ground-truth analyses. Fig. 19 presents the distribution of analysis lengths for three settings using agents powered by GPT-4.1-mini: (1) Baseline agent *without* explicit analysis length requirement, (2) Baseline agent *with* explicit analysis length requirement, and (3) **ANAGENT** *with* explicit analysis length requirement. As shown in Fig. 19, the baseline agent without analysis length expectations consistently generates short analyses, in stark contrast to the substantial variability observed in ground-truth analyses colored in gray. On the other hand, explicitly instructing the agent to generate analyses around the specified length not only increases the overall length of the generated outputs but also substantially enriches the variance in analysis length. **ANAGENT** with length expectations exhibit variation more consistent with the ground-truth analyses. Building upon the observations from GPT-4.1-mini (Fig. 19), we employ Qwen3-VL-8B to power agents under four settings: (1) Baseline agent *without* explicit analysis length requirement; (2) Baseline agent *with* explicit analysis length requirement; (3) **ANAGENT** *with* explicit

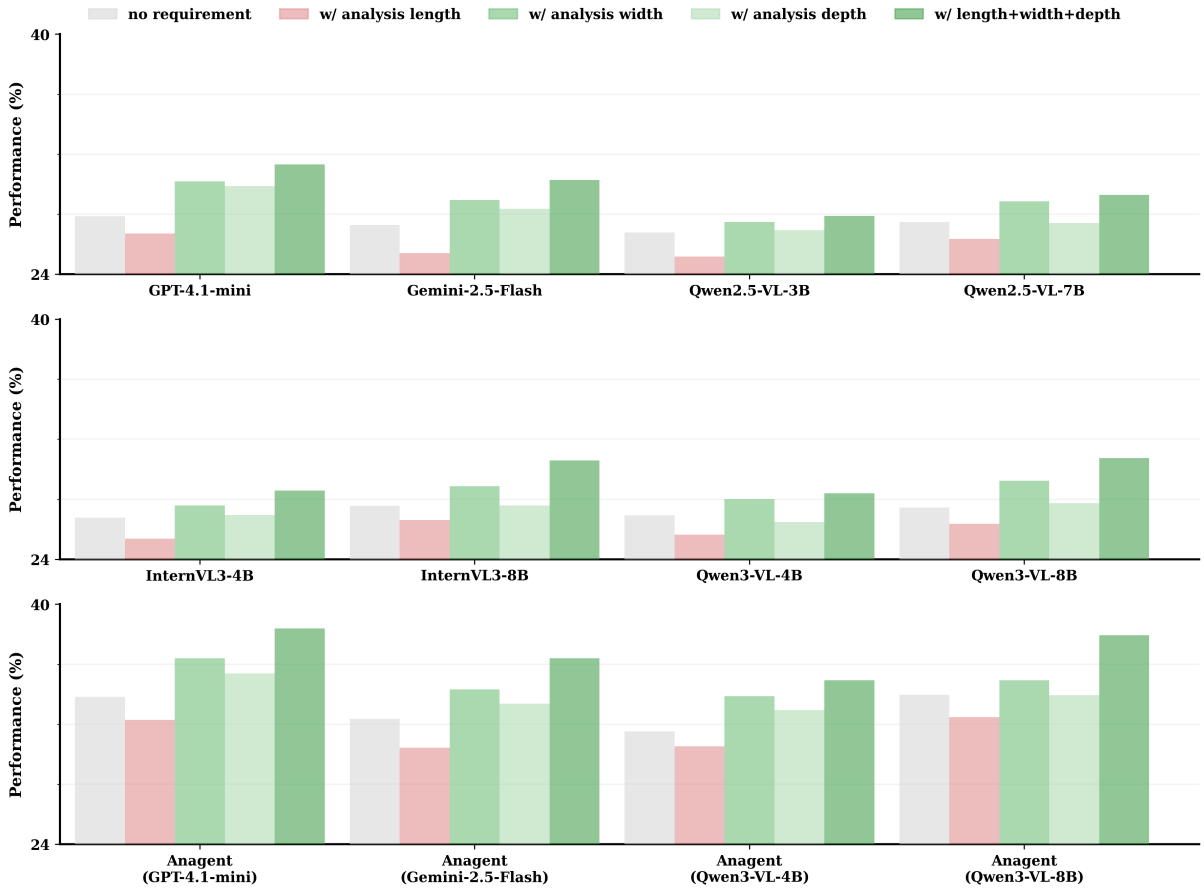


Figure 22: **When AI Scientists Meet Stated Analysis Expectations.** Our experiments investigate AI agents' capabilities in properly comprehending and implementing scientific analysis with explicit expectations.

analysis length requirement; and (4) **ANAGENT**, powered by finetuned Qwen3-VL-8B, *with* explicit analysis length requirement. The upper three groups in Fig. 20 exhibit patterns similar to those observed with GPT-4.1-mini: The baseline agent without explicit length constraints consistently produces short analyses, whereas introducing length requirements increases both the average length and the variance of the generated analyses. On the contrary, **ANAGENT** finetuned through the combination of SFT+RL shows markedly higher variance in analysis length, accompanied by notably improved performance (Tab. 2). Collectively, Figs. 19-20 unveil that current AI agents fail to reliably interpret explicitly stated analysis lengths, which function not only as generation constraints but also as implicit signals of the expected analytical coverage scope.

G.6 Scientific Analysis Across Data Complexities

The complexity of each scientific analysis task is fundamentally shaped by the nature of the input data itself. Across real-world scientific writing, analyses are conducted over heterogeneous data that vary in *type*, *source*, *format*, and *domain* (§C.2.1 & Fig. 15), each of which introduces distinct interpretive and integration challenges. For example, tables and figures differ substantially in how information is structured and accessed, while variations in data domains shape what contextual information and domain-specific knowledge are needed for accurate reasoning and analysis. These variations give rise to different levels of data complexity that directly affect MLLM agents' abilities to search, perceive, extract, and comprehend the related evidence. To systematically examine these challenges across varying data complexities, we evaluate scientific analysis with different data *types*, *sources*, *formats*, and *domains*. Beyond contrasting individual data types, we further extend our investigation to how models handle varying *input scales*, where evidence is distributed across multiple tables, multiple figures, or multimodal combinations of different data types. Results reveal how multiple facets of data complexity influence scientific table &

figure analysis, suggesting potential directions for future improvement.

Tabular Structures Are More Challenging Than Multimodal Figure Reasoning. Although *figures* require multimodal reasoning, *tables* pose distinct and often greater challenges due to the need to accurately parse and reason over complex tabular structures. As shown in Fig. 23, both the baselines and **ANAGENT** consistently achieve higher performance on scientific figures than on tables, with absolute improvements of $\Delta_{abs} \geq 2.40\%$ for baselines, $\Delta_{abs} \geq 1.47\%$ for **ANAGENT**, $\Delta_{abs} \geq 4.83\%$ for finetuned **ANAGENT**. Augmented with task-oriented **EXPERT**, **ANAGENT** narrows the performance gap to 1.47%. Nonetheless, the heterogeneous layouts and diverse semantic intents of scientific tables (*e.g.*, reporting empirical results versus comparing methods) continue to present substantial challenges for reliable scientific analysis.

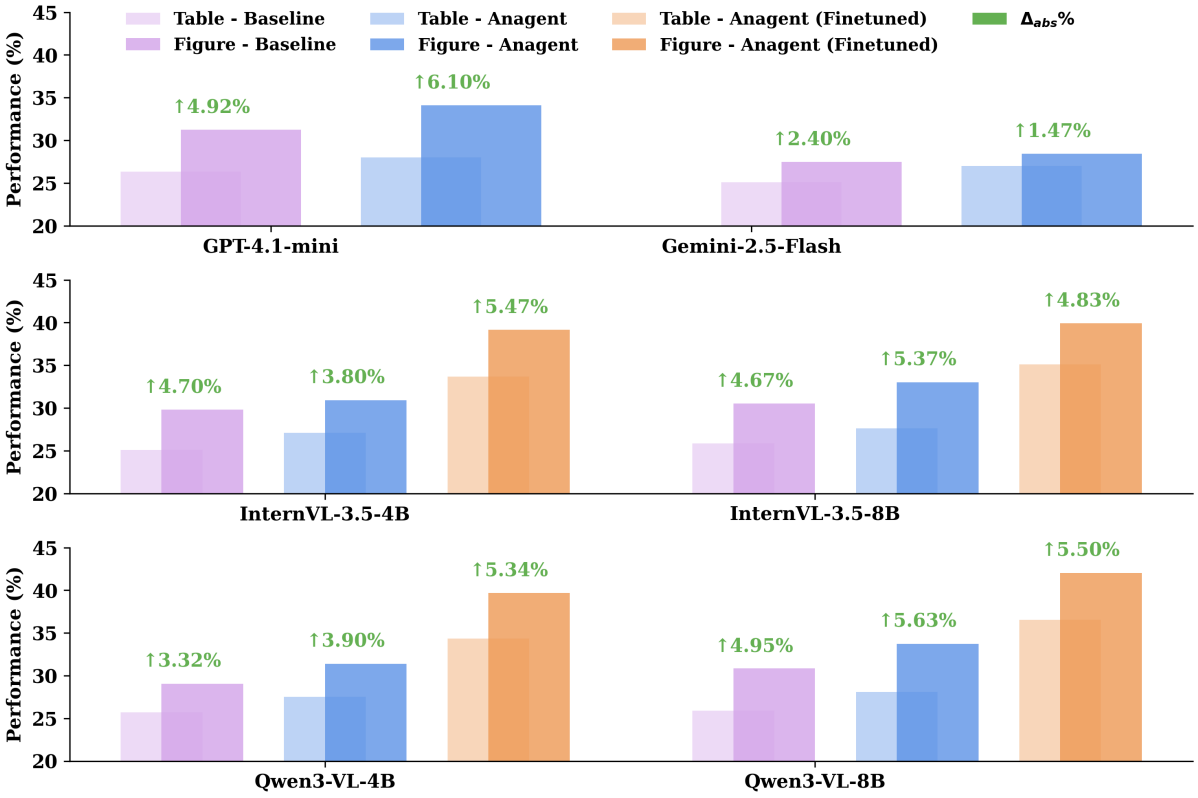


Figure 23: **Ablation Study on Data Type.** Evaluation of agent performance across different data types (§C.2.1).

LaTeX Is Easier to Understand Than XML. We observe a clear and consistent performance advantage when scientific contents are represented in `LaTeX` rather than `XML`. As shown in Fig. 24, all baselines as well as **ANAGENT** achieve higher accuracy on LaTeX data ($S_{\text{AVG}} \geq 27.15$) as compared to XML ($S_{\text{AVG}} \geq 23.44$), indicating that `XML` introduces additional challenges for MLLM-powered agents. Across both data formats, performance improves monotonically from baselines to **ANAGENT** and further to the finetuned **ANAGENT**, demonstrating the robustness of our approach regardless of representation. Nevertheless, the overall performance gap between `LaTeX` and `XML` persists ($\Delta_{abs} \geq 1.07$), suggesting that `XML`'s verbose and nested structure hinders effective reasoning. A possible reason is that `LaTeX` is the dominant format for scientific writing and is therefore more prevalent in model pretraining corpora, leading to stronger prior familiarity. In contrast, XML often emphasizes structural markup over semantic clarity, requiring agents to interpret scientific meaning from less explicit cues and references, which further exacerbates the difficulty.

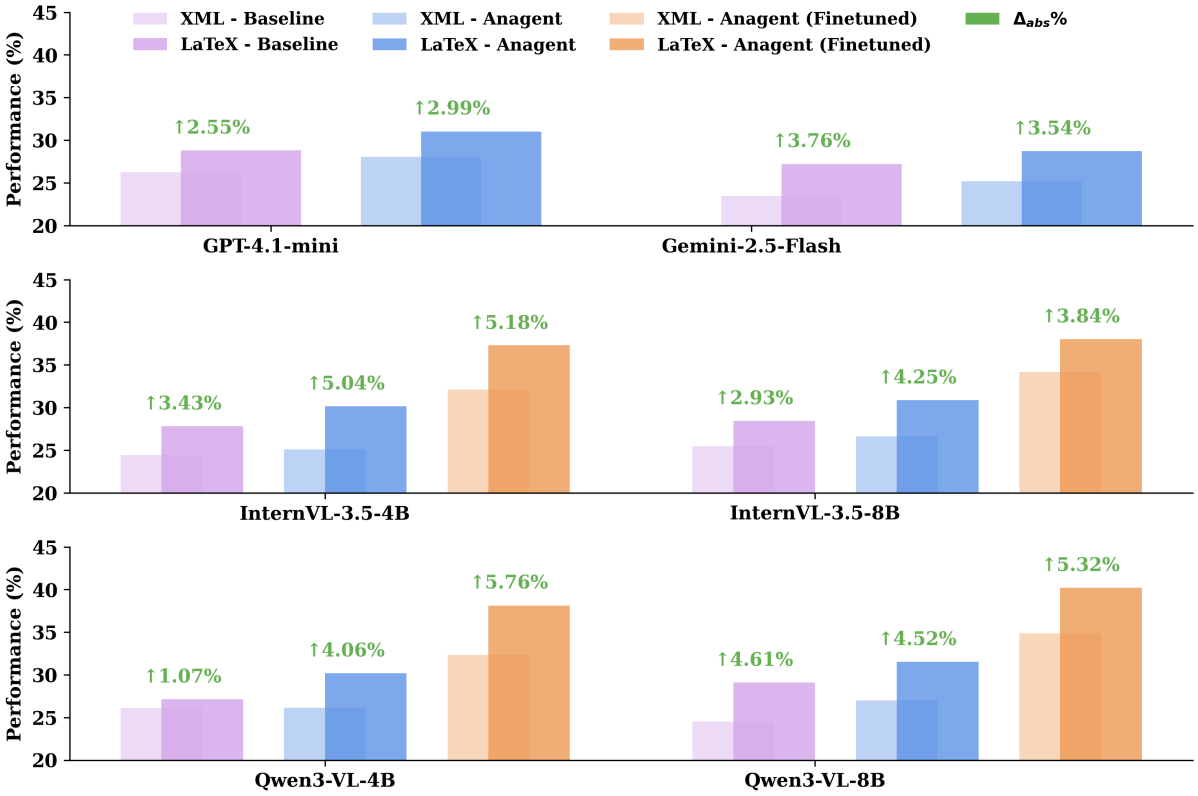


Figure 24: **Ablation Study on Data Format.** Evaluation of agent performance across different data formats (§C.2.1).

Review and Survey Papers Are More Challenging Than General Scientific Literature. We observe that both baselines and ANAGENT consistently achieve higher performance on general scientific paper analysis tasks than on review and survey papers. As illustrated in Fig. 25, the absolute performance gap between these two data sources remains substantial across all models ($\Delta_{abs} \geq 3.51\%$), reaching up to 8.07% on the InternVL3.5-8B baseline. Despite this difficulty, performance improves monotonically from baselines to ANAGENT and further to the finetuned ANAGENT on both paper sources, indicating the effectiveness of our method across varying document sources and types. Notably, in most cases, ANAGENT and finetuned ANAGENT can narrow the gap between general papers and review/survey papers, indicating the effectiveness of specialized tools for information searching and retrieval. Through manual checking, we also observe that scientific analysis sections in review and survey papers are significantly less likely to be self-contained as compared with general papers. This supports our potential thinking that attributes the increased difficulty of review and survey papers to their heavy reliance on extensive internal and external references, cross-paper comparisons, and high-level synthesis, which makes their scientific analysis less self-contained than that of general research articles.

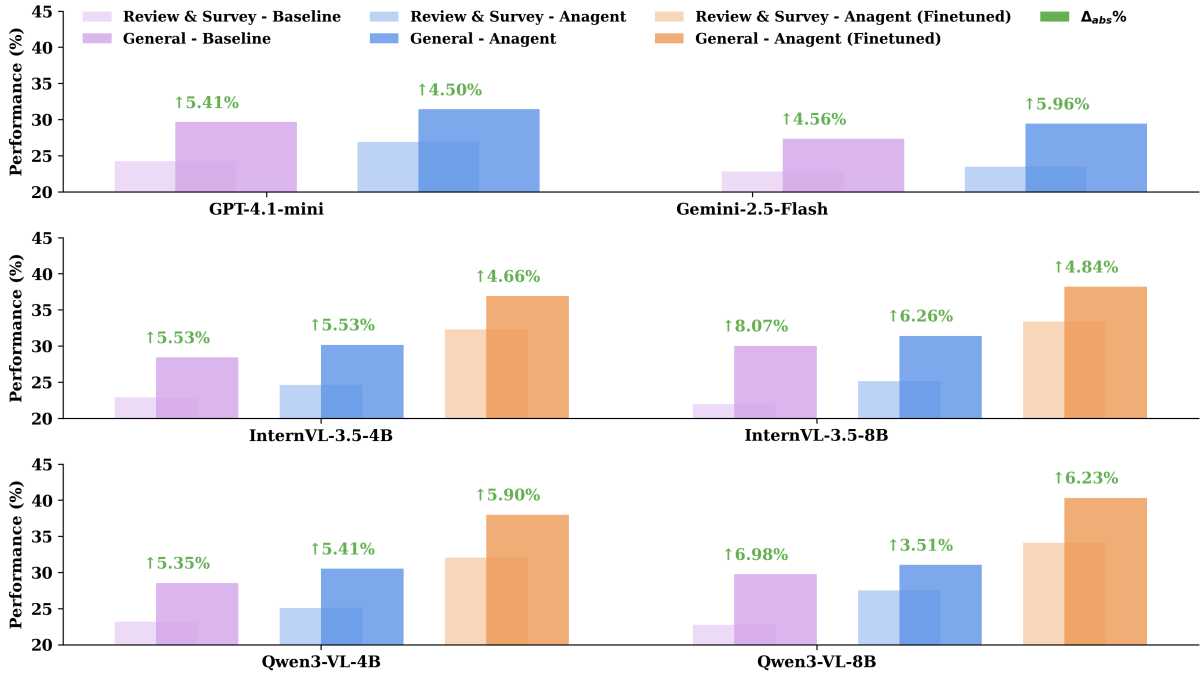


Figure 25: **Ablation Study on Data Source.** Evaluation of agent performance across different data sources (§C.2.1).

Enhancing Scientific Analysis Across Domains. We observe inconsistent variability in scientific analysis performance across domains. As shown in Fig. 26, baselines exhibit pronounced domain variability in scientific analysis performance. In particular, the lowest baseline performance is observed on *Statistics* domain ($S_{\text{AVG}} = 17.08\%$), followed by *Economics* ($S_{\text{AVG}} = 20.28\%$) and *Quantitative Finance* ($S_{\text{AVG}} = 21.87\%$). In contrast, the highest baseline performance is achieved on *Quantitative Biology* ($S_{\text{AVG}} = 33.96\%$). This disparity reveals that baseline agents struggle to generalize scientific reasoning capabilities across domains with varying levels of mathematical abstraction, formalism, and domain-specific assumptions. In contrast, both **ANAGENT** and finetuned **ANAGENT** consistently improve performance across all domains ($\Delta_{rel} \geq 4.06\%$), effectively elevating even the most challenging *Statistics* domain from the lowest baseline performance to $S_{\text{AVG}} = 36.84\%$ under finetuned **ANAGENT**. Moreover, when comparing the performance distribution across domains, baseline agents show pronounced inter-domain performance gaps, while **ANAGENT** significantly reduces such variation, and finetuned **ANAGENT** further presents an even more uniform performance distribution across domains. This trend reveals the effectiveness of **ANAGENT** in mitigating domain-specific reasoning bottlenecks, leading to more robust and consistent scientific analysis across diverse research areas.

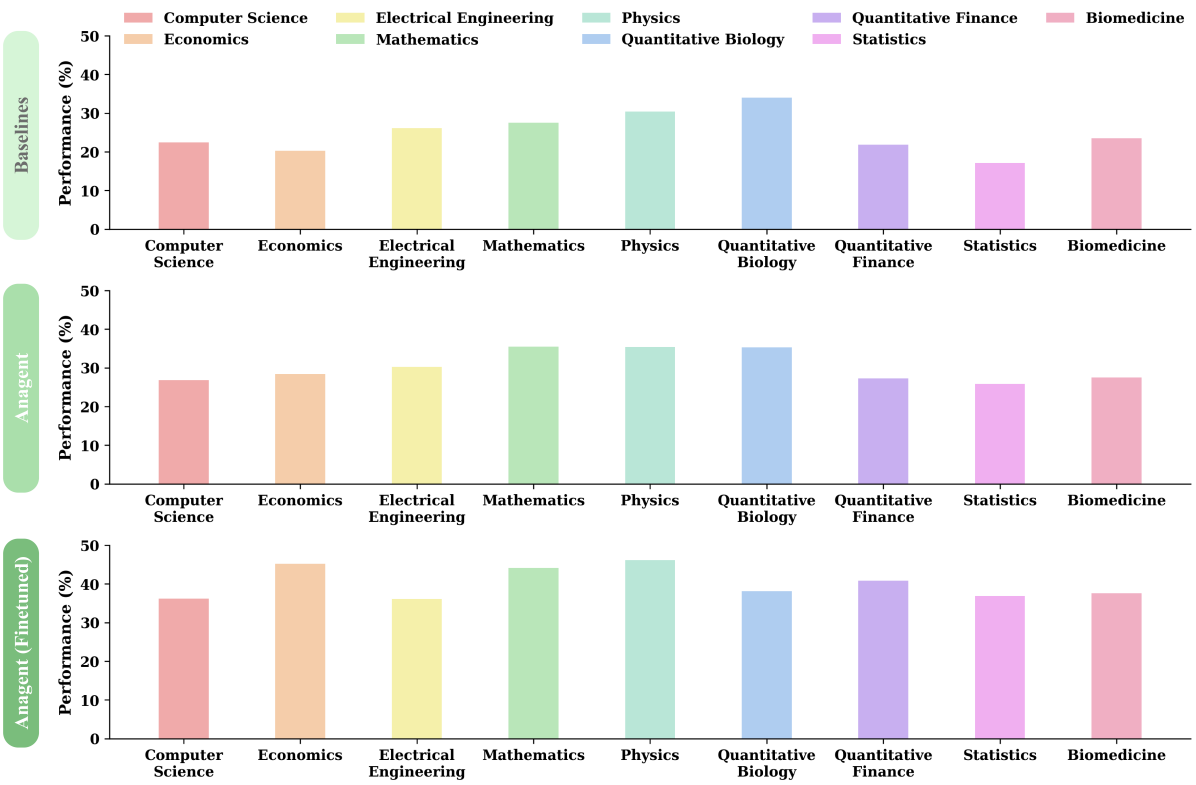


Figure 26: **Ablation Study on Data Domain.** Evaluation of agent performance across nine broad domains (§C.2.1).

Scaling Inputs Furnishes Additional Contexts While Simplifying Data Complexity. In addition to the width, depth, and objectives of an analysis task, the nature of the input data plays a fundamental role in determining the complexity of scientific analysis. Input data may take the form of a single table or figure, multiple tables or figures, or a combination of both. To this end, we conduct an additional study examining input coverage across inputs that vary in layout, modality, and quantity. As shown in Fig. 27, expanding input coverage generally leads to more reliable and higher-quality analyses, with our method consistently outperforming baseline counterparts across all settings. This trend suggests that broader data coverage provides complementary contextual cues and domain knowledge that help agents better interpret experimental evidence and articulate scientific insights. Notably, agents tend to benefit from multi-input settings over single-input ones (i.e., *darker* bars higher than *lighter* bars in most cases), indicating that aggregating information across multiple sources often facilitates reasoning by grounding analysis in richer contextual support. The benefits of increased coverage are particularly pronounced for multimodal inputs. MLLM-powered agents exhibit stronger reasoning abilities when figures are present, either alone or in combination with tables, as compared to tabular-only layouts. In some cases (*e.g.*, Qwen2.5-VL baselines), single-figure inputs can yield performance comparable to, or even exceeding, that of other input coverage settings, underscoring the advantages of visual representations for improved scientific reasoning. While scaling inputs typically enhances performance, these gains are not unbounded. Agents with limited capacity may struggle to effectively integrate excessive information, resulting in diminished returns (*e.g.*, Qwen3-VL-4B gains higher performance on `S-T&S-F` over `M-T&M-F`). Overall, evaluation results across six input coverage settings highlight a nuanced trade-off in input scaling, where richer coverage can reduce effective data complexity and improve analysis, as long as it aligns with the agent's reasoning capacity.

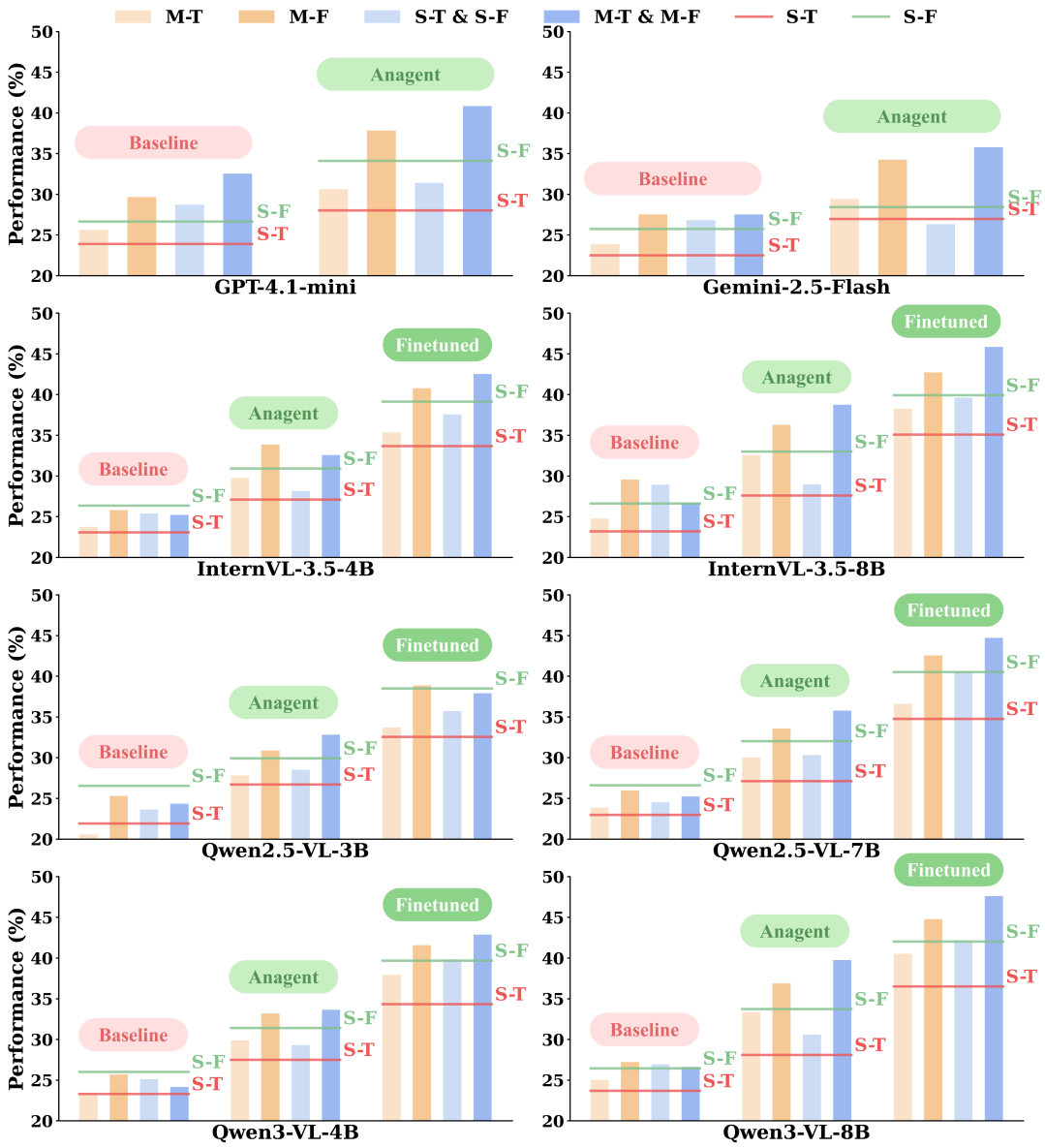


Figure 27: **Ablation Study on Input Coverage.** Evaluation of agent performance across six different input coverage settings. `S-T` represents single-table input, `S-F` represents single-figure input, `M-T` represents multi-table input, and `M-F` represents multi-figure input.

G.7 Scientific Analysis Across Analysis Complexities

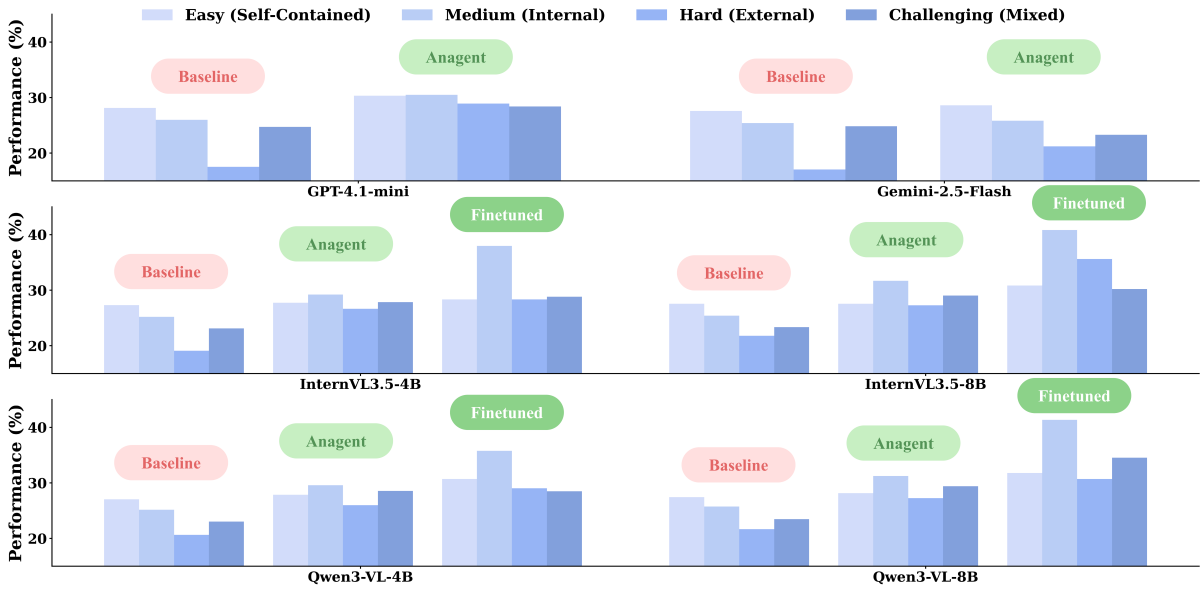


Figure 28: **Ablation Study on Analysis Width.** Evaluation of agent performance across four analysis width curriculum (§C.2.2).

Variations in analysis *depths*, *widths*, and *objectives* contribute to tiered curriculum complexities (§C.2.2 & Fig. 15) that reflect the level of detail, breadth of coverage, and overall focus in real-world scientific analysis. With varying writing widths, scientific analysis may either *focus narrowly on the individual table or figure*, such as summarizing and interpreting it in isolation to draw specific findings; or *adopt a broader perspective*, comparing across multiple tables and figures to identify patterns and synthesize insights. The combination of analysis width and depth further accounts for how much of an analysis section is devoted to the targeted input. Some analyses may dedicate only a small portion to a specific table or figure, while others may use it as the central basis for extensive discussion and conclusions. Moreover, targeting methodology interpretation or experimental demonstration, analysis objectives fundamentally shape how an analysis is formulated. By evaluating the challenges introduced by varying analysis widths, depths, and objectives, results reveal their impact on scientific table & figure analysis across different agents.

The Core Challenge Of Analysis Width Comes From External References. As suggested by the challenges (Fig. 1) and the curriculum of *(d) Analysis Width* (Fig. 15), we implement the four-level complexity curriculum for analysis width, including: ***easy*** tasks for self-contained analysis writing, ***moderate*** tasks with internal references within the source paper, ***hard*** tasks with external references aside from the source paper, and ***challenging*** tasks that contain both internal and external references. Results in Fig. 28 reveal how increasing analysis width affects scientific table and figure analysis writing. Specifically, across all complexity levels, **ANAGENT** variants consistently outperform baselines, with further improvements observed after finetuning, unveiling the effectiveness of our multi-agent collaboration design. The reduced variations among four analysis width curricula further demonstrate the robustness of collaborative analysis writing under varying context widths. Notably, task difficulty does not monotonically increase with the curriculum level. For baselines, the highest performance is consistently achieved on self-contained tasks and degrades substantially when external references are involved, reflecting agents' limited capabilities to adaptively comprehend and incorporate additional contextual knowledge. In contrast, for both **ANAGENT** and its finetuned variant, ***moderate*** tasks achieve the strongest performance, suggesting that internal references, such as paper-specific definitions, methodologies, experiments, domain concepts, etc., provide the most effective contextual grounding for scientific analysis writing. Performance declines markedly on ***hard*** and ***challenging*** tasks that involve external references, underscoring external knowledge integration as the primary bottleneck in scientific analysis. Interestingly, in some cases, agents exhibit

lower performance on ***hard*** tasks than on ***challenging*** ones, despite the latter containing the broader combination of both internal and external references. This observation reveals that current MLLM-powered agents struggle to selectively identify, interpret, and adapt externally referenced knowledge that is closely connected to the internal context, whereas the presence of internal references in challenging tasks may partially anchor the use of external information. Overall, these results highlight the critical role of highly relevant internal context in supporting scientific analysis, as it supplies essential concepts, core methodologies, and domain-specific cues that facilitate coherent reasoning. Conversely, self-contained tasks place greater emphasis on precise interpretation of the input tables and figures themselves, while external-reference-heavy tasks pose the greatest challenge due to the need for adaptive comprehension and task-specific integration of externally sourced knowledge.

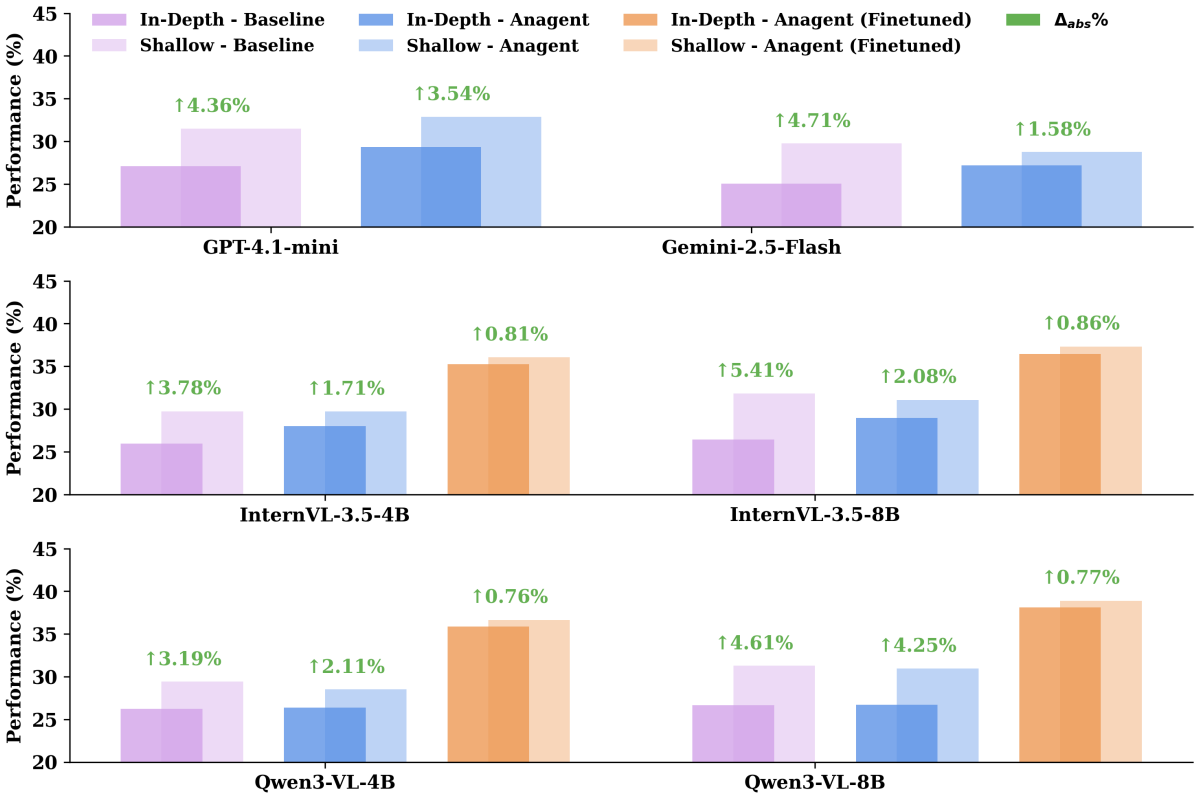


Figure 29: **Ablation Study on Analysis Depth.** Evaluation of agent performance across two analysis depth: *shallow analysis* and *in-depth analysis* (§C.2.2).

Increased Analysis Depth Leads to More Challenging Tasks. Following the curriculum design of *(c) Analysis Depth* in Fig. 15, we instantiate two levels of analysis depth as illustrated in Fig. 1: (1) ***shallow*** analysis tasks, which primarily involve surface-level description and direct interpretation of tables and figures, and (2) ***in-depth*** analysis tasks that require reasoning beyond evidence, causal and logical interpretation, as well as synthesis of underlying conclusions or experimental implications. Ablation results are presented in Fig. 29. Across all evaluated methods, performance on ***in-depth*** analysis tasks is consistently lower than that on ***shallow*** tasks, confirming that increased analysis depth introduces substantially greater difficulty. This performance gap suggests that deep scientific reasoning, such as drawing non-trivial inferences, explaining observed trends, or connecting empirical results to broader methodological or theoretical considerations, remains challenging for current MLLM agents. Despite this increased difficulty, our approach demonstrates consistent improvements over baselines at both depth levels, with additional gains obtained through finetuning. Importantly, the performance gap between shallow and in-depth analysis is noticeably reduced for **ANAGENT** ($\Delta_{abs} \leq 4.25\%$) and its finetuned variant ($\Delta_{abs} \leq 0.86\%$) as compared to baselines ($\Delta_{abs} \leq 5.41\%$). This reduction indicates that collaborative analysis writing is effective at supporting not only shallow but also deeper reasoning, enabling

agents to better decompose complex analytical requirements and progressively refine interpretations. Overall, these results suggest that, while increased analysis depth substantially raises task difficulty, structured multi-agent collaboration provides a meaningful means of handling deeper scientific reasoning and analysis.

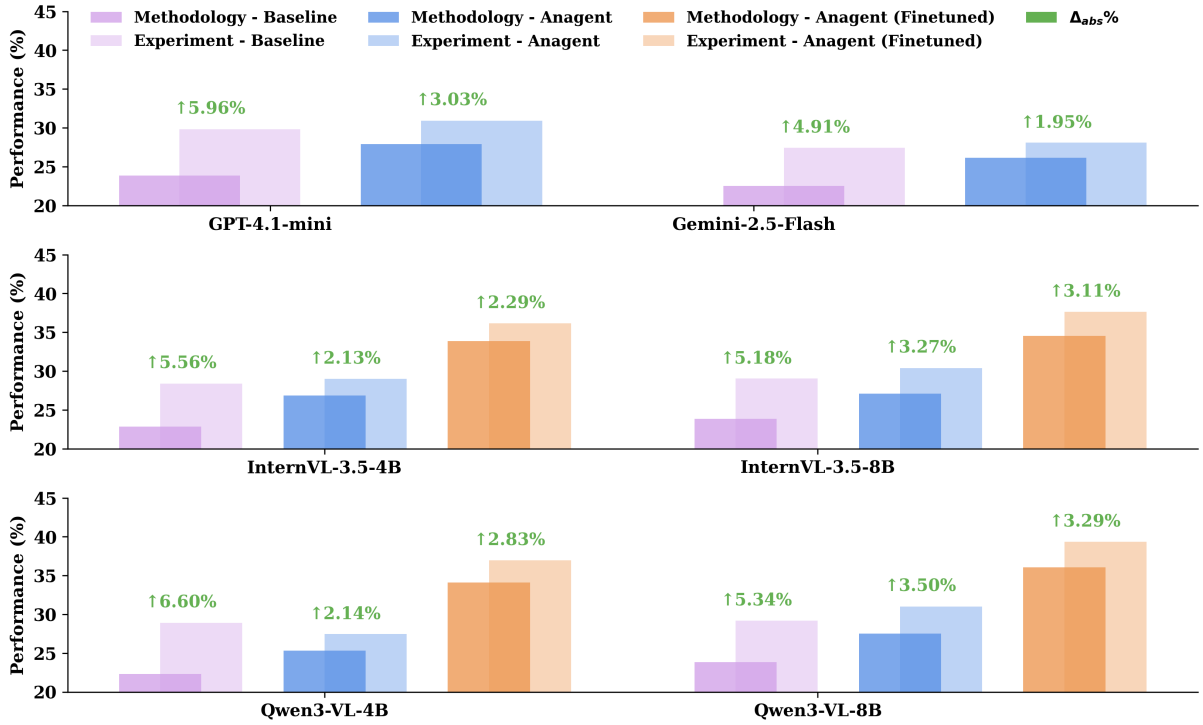


Figure 30: **Ablation Study on Analysis Objective.** Evaluation of agent performance across two analysis objectives: *methodology-oriented* analysis and *experiment-oriented* analysis (§C.2.2).

Methodology-Oriented Analysis Is More Challenging Than Experiment-Oriented Analysis. Scientific analysis tasks are typically driven by different analytical objectives. Accordingly, we categorize scientific analysis into two categories according to the analysis objective of the task: (1) ***methodology-oriented*** analysis, which aims to explain, interpret, and reason about methodological designs, theoretical formulations, algorithmic principles, and so on; and (2) ***experiment-oriented*** analysis, which focuses on analyzing empirical results, experimental settings, and observed trends presented in tables or figures. As shown in Fig. 30, the experimental results demonstrate a clear and consistent performance gap between these two analysis objectives across all evaluated agents. The phenomenon where agents in general perform worse on ***methodology-oriented*** analysis than on ***experiment-oriented*** analysis ($\Delta_{\text{ABS}} \geq\downarrow 1.95\%$) indicates that accurately understanding and explaining methodological concepts and theoretical objectives is essentially more challenging than analyzing empirical evidence. This gap is observed consistently across different agent variants, suggesting a general limitation of current MLLM agents in methodology-level reasoning. Despite this challenge, our approach improves performance on both methodology-oriented and experiment-oriented analysis tasks, with further gains achieved through finetuning. More importantly, comparing to baselines ($\Delta_{\text{ABS}} \geq 4.91\%$), the performance gap between the two analysis objectives is notably reduced for **ANAGENT** ($\Delta_{\text{ABS}} \leq 3.60\%$) and finetuned **ANAGENT** ($\Delta_{\text{ABS}} \leq 3.29\%$). This observation suggests that the extended task-specific context and domain-specific knowledge collectively enable agents to better understand fundamental methods, clarify analytical objectives, and perform more coherent and objective-driven scientific analysis. Overall, these results indicate that, while methodology-oriented analysis remains inherently more difficult than experiment-oriented analysis, providing richer contextual grounding and structured collaborative reasoning can enhance agents' abilities in reasoning about both

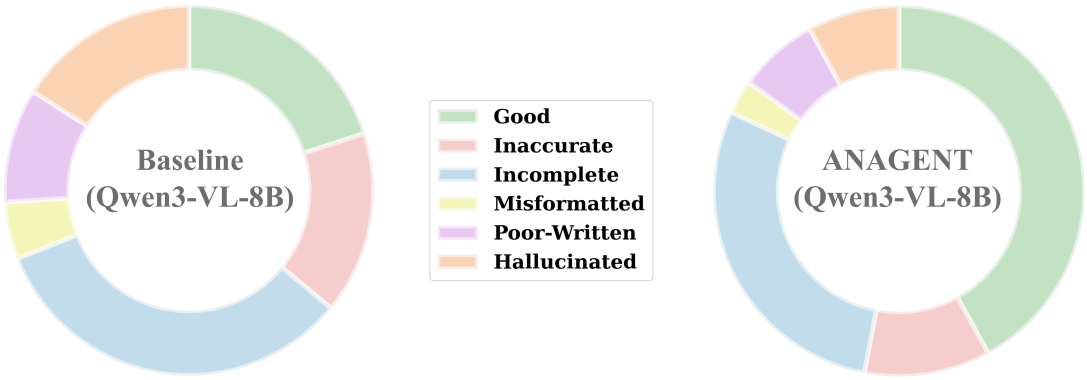


Figure 31: **Domain Evaluation.** Assessing 100 samples manually for each model on each domain, we compare the evaluation results between baseline and ANAGENT using the same base model (Qwen3-VL-8B).

methodological contents and experimental evidence. The narrowed performance gap further underscores the significance of objective-aware contextual support for advancing scientific analysis capabilities beyond empirical result interpretation.

G.8 Human Expert Evaluation of Domain Analysis

To more comprehensively assess the quality of scientific analysis writing, we incorporate an additional evaluation conducted by domain experts. Specifically, human evaluators with expertise in computer science manually assess 100 agent-generated analyses in their respective domains. Each analysis is generated based on explicit human requirements (§22) and evaluated through the same five-dimensional evaluation protocol (§E.2). Results in Fig. 31 show human assessment on two models in *computer science* domain (Tab. 6). The notable increase of "Good" analysis demonstrates significant performance gain achieved by ANAGENT. Meanwhile, the decrease of five-dimensional errors further reveal the effective improvements on scientific analysis writing through systematic planning, reasoning, problem-solving, and reflective refinement (§3).

Dimension	MLLM Judge (%)		Human Expert (%)			Agreement & Significance	
	Baseline	ANAGENT	Baseline	ANAGENT	Gold	Krippendorff's α	Wilcoxon p
Five-Dimensional Evaluation Protocol							
Analysis Consistency	48.23	59.34	52.57	64.52	92.78	0.92	< 0.001
Query-Analysis Alignment	29.32	33.03	32.63	37.42	90.28	0.87	< 0.001
Knowledge Utilization	67.63	78.54	71.24	83.15	93.72	0.88	< 0.001
Format Correctness	70.22	76.36	72.18	80.60	**99.37**	**0.97**	< 0.001
Grounding Accuracy	55.37	67.91	58.52	72.23	92.74	0.82	< 0.001
Overall	**54.15**	**63.04**	**57.43**	**67.58**	**93.78**	**0.89**	$\mathbf{< 0.001}$

Table 14: **Multi-Domain Human Expert Evaluation.** Three experts per domain rate agent-generated and gold analyses across six domains under the five-dimensional protocol (§E.2). **Gold** shows expert ratings of ANABENCH's author-written reference analyses under the identical protocol.

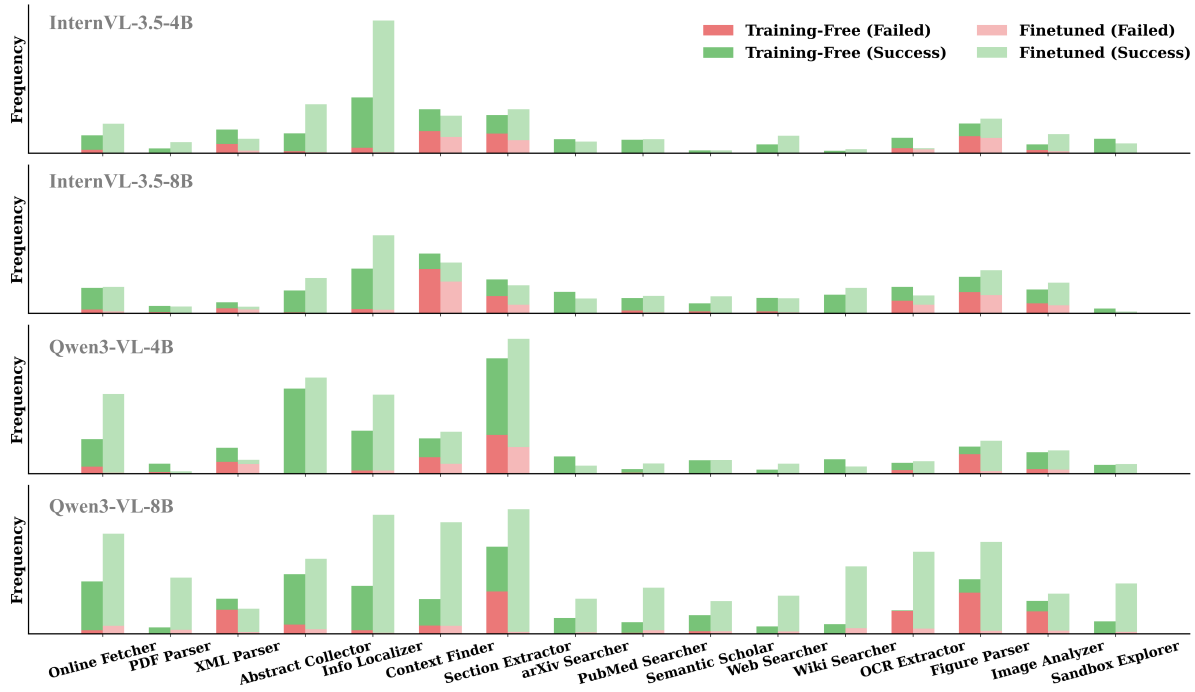


Figure 32: **ANAGENT Tool Utilization.** Investigation on tool utilization of training-free **ANAGENT** and finetuned **ANAGENT**.

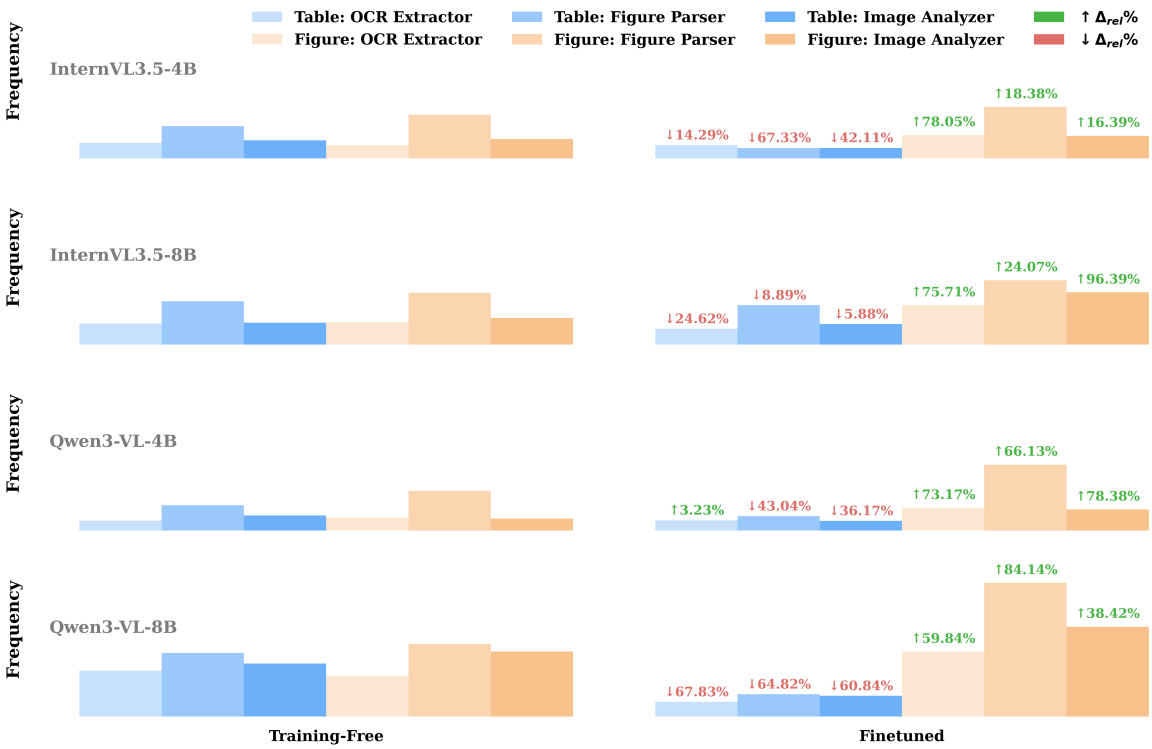


Figure 33: **ANAGENT Tool Utilization Strategy Optimization Through Finetuning.** Optimization of tool utilization strategy through finetuning.

To assess whether the improvements in Fig. 31 generalize beyond a single discipline, and to verify the quality of ANABENCH's references directly, we extend the expert study to six domains: computer science, electronic engineering, biomedicine, mathematics, statistics, and economics (Tab. 14). In each domain, three experts independently rate 50 agent-generated and 50 gold analyses under the same five-dimensional protocol used by the MLLM judge (§E.2), giving 600 rated analyses in total. Analyses are produced by Qwen3-VL-4B, and the MLLM columns are scored by GPT-4.1-mini. As experts rate gold analyses under the identical protocol, model outputs and references are directly comparable on a single scale.

As shown in Tab. 14, ANAGENT improves over the baseline on every dimension under expert rating, with the largest gains on grounding accuracy and analysis consistency, and every difference is significant at $p < 0.001$ after Holm correction. In addition, inter-expert agreement is high throughout ($\alpha \geq 0.82$), so the ratings do not reflect individual annotator preference. Also, experts rate the gold analyses at 93.78% overall, which empirically confirms the quality of ANABENCH's author-written references and quantifies the remaining headroom between ANAGENT and human-written scientific analysis.

G.9 Tool Utilization

Equipped with five toolkits comprising sixteen tools in total (Tab. 13), ANAGENT extends its contextual awareness and domain expertise through autonomous tool invocation. However, effective tool utilization requires accurate alignment between tool functionalities and task-specific execution objectives. Without such ability, agents exhibit substantially degraded performance due to non-strategic or failed tool calls (Fig. 32).

Fig. 32 presents a comparative analysis between finetuned ANAGENT (lighter colors) and training-free ANAGENT (darker colors) across four backbone MLLMs. Overall, finetuned ANAGENT demonstrates a pronounced increase in the usage frequency of critical information-retrieval tools from the *search toolkit*, such as *abstract collector* ($\Delta_{rel} \geq \uparrow 12.31\%$), *information localizer* ($\Delta_{rel} \geq \uparrow 61.27\%$), etc. This trend suggests that finetuning enables the model to more reliably identify when additional knowledge acquisition is necessary and to select appropriate tools accordingly. Beyond increased tool engagement, finetuned ANAGENT consistently achieves a markedly higher tool invocation success rate, accompanied by a substantial reduction in failed or invalid tool calls. These observations reveal that finetuning not only improves ANAGENT 's awareness of *which* tools to use, but also enhances ANAGENT 's ability to correctly format and execute tool calls, thereby reducing execution-level errors and facilitating analysis generation. More importantly, finetuned ANAGENT exhibits clearly more *strategic* and task-aware tool utilization. As illustrated by Fig. 33, in figure-centric scientific analysis tasks, finetuned ANAGENT invokes tools from the *vision toolkit* $\Delta_{rel} \geq \uparrow 16.39\%$ more frequently than its training-free counterpart. This behavior reflects a stronger alignment between task objectives and tool selection, highlighting the effectiveness of finetuning in fostering objective-oriented reasoning and adaptive tool-use policies. Taken together, these findings demonstrate that finetuning effectively enhances both the reliability and strategic capacity of tool utilization. The improved tool-call success rate and task-aware tool selection jointly contribute to the superior scientific analysis performance of ANAGENT, meanwhile underscoring the critical role of objective-driven tool orchestration in multimodal agent systems.

G.10 Controlled Experiments: Test-Time Scaling & Compute Accounting

Two mechanisms could in principle explain ANAGENT's gains without any contribution from multi-agent collaboration: the additional information retrieved by EXPERT, and the larger inference-time compute consumed by multi-turn execution. §G.4 addresses the first by supplying baselines with oracle context. Here we address the second directly, comparing ANAGENT against single-model test-time scaling strategies whose call budgets match or exceed its own.

Controls. We evaluate three standard single-model test-time scaling strategies on the same Qwen3-VL-4B backbone and the same evaluation set (§G.4). *Self-Consistency* (SC@N) (Wang et al., 2023) samples N candidate analyses and returns the one with the highest mean pairwise similarity to the remaining $N-1$. *Best-of-N* (BoN@N) samples N candidates and returns the one with the highest S_{AVG}. *Self-Refine* (SR@N) (Madaan et al., 2023) has a single CRITIC iteratively critique and revise its own analysis for N

Method	MLLM Calls	Tokens/Call	Latency	S_{AVG}	Δ_{REL}
Single-Model Test-Time Scaling					
Direct Generation	1.00	1.00×	1.00×	26.76	–
Self-Consistency (SC@10)	10.00	1.18×	7.09×	26.82	↑0.22%
Best-of-N (BoN@10)†	10.00	1.07×	7.11×	27.03	↑1.01%
Self-Refine (SR@10)	10.00	4.06×	7.47×	26.79	↑0.11%
Multi-Agent (Training-Free)					
Omnion	5.48	4.87×	7.26×	22.87	↓14.54%
ANAGENT	6.64	4.52×	**6.24**×	28.73	↑7.36%
Multi-Agent (Modular Training)					
ANAGENT (SFT)	6.79	4.38×	6.36×	**36.02**	↑34.60%

Table 15: **Controlled Test-Time Scaling & Compute Accounting.** All experiments use the same Qwen3-VL-4B backbone and the same evaluation set (§G.4). MLLM calls, tokens per call, and latency are averaged per instance, with tokens per call and latency calculated relative to single-pass direct generation. *Relative differences* (Eq. 29) are shown as *positive* ↑ $\Delta_{rel}\%$ or *negative* ↓ $\Delta_{rel}\%$. †BoN@10 selects the candidate with the highest S_{AVG} against the reference analysis, and is therefore an *oracle upper bound* rather than a deployable selection rule.

rounds. We set $N = 10$ so that every control issues more MLLM calls than ANAGENT's per-instance average of 6.64. Because BoN selects using the evaluation metric itself, it is an oracle upper bound on best-of-N selection rather than a deployable method. All methods are measured on the same hardware.

Additional inference compute alone does not produce gains. SC, BoN, and SR each issue $N = 10$ MLLM calls, more than ANAGENT's 6.64. However, they improve over single-pass generation by only ↑ 0.11% to ↑ 1.01%, against ANAGENT's ↑ 7.36% training-free and ↑ 34.60% after modular training. This holds even though BoN is granted oracle selection, which upper-bounds what any best-of-N rule could achieve at that budget. Spending more test-time compute on the base model, without structural change, does not reproduce the gains.

What tokens are spent on matters more than how many. SR is the largest token consumer among the controls, at 4.06× tokens across 10 calls, and still gains only ↑ 0.11%. *Omnion*, ANAGENT, and ANAGENT (SFT) all operate in a comparable per-call regime of 4.38× to 4.87×, yet span ↓ 14.54% to ↑ 34.60%. SR is compute-neutral because resampling and self-critique redraw only from the model's own distribution and cannot introduce information the backbone lacks, while *Omnion* shows that tool access without orchestration is actively harmful. Only orchestrated collaboration converts the budget into gains.

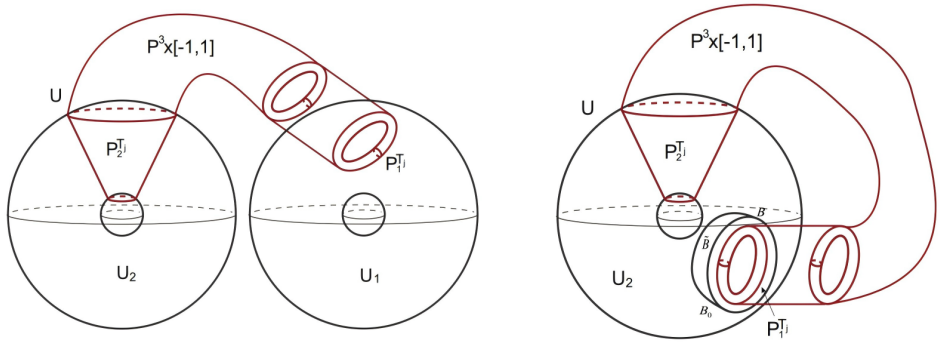
Training is orthogonal to inference cost. ANAGENT (SFT) runs at essentially the same inference cost as its training-free counterpart, 6.79 calls at 4.38× tokens against 6.64 calls at 4.52×, in fact slightly cheaper per call. This effectively reflects more reliable and more targeted tool calls after finetuning (§G.9). It nonetheless raises the relative gain from ↑ 7.36% to ↑ 34.60%. Together with the modular versus end-to-end comparison at identical inference cost (Tab. 3), this indicates that the lever is the method rather than the inference budget.

H Example: ANAGENT For Scientific Table & Figure Scientific Analysis

Here is an end-to-end scientific analysis writing example of ANAGENT:

Example of A Good Scientific Analysis

Input:



ANAGENT Output:

Planner

- Find, retrieve, distill, and summarize all the related contexts from the paper, especially the definitions of P^4, $P^3\times[-1,1]$, P_T^1, P_T^2, B_0, $\tilde{B}$, and the surgery construction to ensure precise topological interpretation.
- Map the geometric relationships in the figure to the abstract topological operations described in Lemma \ref{UiL} and Lemma \ref{decoo}, particularly how the manifold M_j is constructed via surgery along tori and how boundaries Q_j relate to E-irreducible components.
- Analyze the topological significance of the solid tori P_T^1, P_T^2 and the 3-manifold B_0 in the context of the proof of Lemma \ref{UiL}, case 2), and how they constrain the topology of M_j to be homeomorphic to P^4.
- Explicitly connect the surgery operation depicted in the figure to the abstract formula \eqref{P123} and the decomposition \eqref{rouhan} to justify why the resulting manifold M_j must be homeomorphic to P^4.
- Synthesize the above into a single cohesive paragraph of scientific analysis, ensuring the paragraph is approximately 902 tokens, written in plain LaTeX, and avoids merely describing the figure while maintaining precise, professional, and academically rigorous language.

Expert

Expert

- * **Objective:** Analyze Lemma (\ref{UiL}), Case 2), showing that surgery along an embedded 2-torus (T_j \subset M_j) preserves the homeomorphism type of the manifold within a filtration decomposition.
- * **Model manifold:** (M_j) is assumed (or previously shown) to be homeomorphic to (P^4 = \mathbb{S}^2 \times \mathbb{S}^1), with boundary (\partial P^4 = \mathbb{S}^2 \times \mathbb{S}^1).
- * **Surgery locus:** (T_j) is a 2-torus embedded in (M_j) with a tubular neighborhood (N(T_j)).
- * **Gluing data:** The surgery is defined by an embedding (\varphi_j : P^3 \times \mathbb{S}^0 \to \partial N(T_j)), specifying how new pieces are attached.
- * **Replacement pieces:** Two solid tori (P_T^1) and (P_T^2) are glued in after removing (N(T_j)).
- * **Construction:** The surgically modified manifold is (\bar{M}*j /*{\varphi_j}), denoted (M_j^T), obtained by replacing (T_j) with (P_T^1 \cup P_T^2).
- * **Key topological insight:** This is a standard torus surgery that does not alter the global topological type of the manifold.
- * **Main result:** The post-surgery manifold satisfies (M_j^T \cong P^4), hence (M_j^T) is homeomorphic to the original (M_j).
- * **Evidence:** The boundary identifications and solid-torus replacements respect the existing product structure and introduce no new topological features.
- * **Domain context:** Uses classical results from 3- and 4-manifold topology on torus surgery and cut-and-paste constructions.

Solver

```
\\section{Topology of the carrying manifold}
In this section, we will prove Theorem \\ref{s3s1s}. To do this, note that in the
filtration \\eqref{fil}, $M^4\\setminus{ int}\\,M_{k-1}\\cong P^4$. Then, by
\\cite{Max}), it suffices to show that $M_{k-1}\\cong P^4$. This fact will follow
from the following lemma.\\begin{lemma} Each connected component of the manifold
$M_j,\\,j\\in\\{1,\\dots,k-1\\}$ in the filtration \\eqref{fil} is homeomorphic to
$P^4$.\\end{lemma}\\begin{proof} From the decomposition \\eqref{fil} it follows that
$M_j$ consists of $j$ copies of the manifold $P^4$ for $j\\in\\{1,\\dots,k_0\\}$.
From formula \\eqref{rouhan} it follows that for $j\\in\\{k_0+1,\\dots,k-1\\}$, the
manifold $M_j$ is obtained from the manifold $M_{j-1}$ in the following way. Let
$\\bar M_{j-1}=M_{j-1}\\sqcup (P^3\\times\\mathbb B^1)$ and
$\\varphi_j:P^3\\times\\mathbb S^0\\to Q_{j-1}$ be an embedding such that
$\\varphi_j(P^3\\times\\{-1\\})=P^{T_j}_1,\\varphi_j(P^3\\times\\{1\\})=P^{T_j}_2$
are the solid tori associated with $T_j$. Then3. Additional InformationThis analysis
focuses on Lemma \\ref{UiL}, Case 2), which addresses the topological structure of a
4-manifold $M_j$ after surgery along a 2-torus $T_j$ embedded in $M_j$, specifically
within the context of a filtration manifold decomposition. The key topological
objects involved are $P^4 = \\mathbb{S}^2 \\times \\mathbb{S}^1$, $P^3 \\times
[-1,1]$, and $B_0$, with $P_T^1$ and $P_T^2$ being the solid tori associated with
$T_j$. The surgery construction $\\bar{M}_j /_{\\varphi_j}$ is defined via the
embedding $\\varphi_j: P^3 \\times \\mathbb{S}^0 \\to \\partial N(T_j)$, which
identifies the boundaries of the solid tori $P_T^1$ and $P_T^2$ with the boundary
torus $T_j$. The resulting manifold $M_j^T$ inherits the topological structure of the
original manifold $M_j$ but with the torus $T_j$ replaced by the solid tori $P_T^1$
and $P_T^2$. [...]
```

Critic

Critic

```
<accuracy>0</accuracy>
<completeness>0</completeness>
<format>0</format>
<writing>0</writing>
<faithfulness>0</faithfulness>
<feedback>
- The answer is not 1-3 cohesive paragraphs as required; it is fragmented with
repeated, redundant text and incomplete sentences.
- It fails to analyze the topological surgery operation or the role of $P_T^1$,
$P_T^2$, and $B_0$ in Case 2 of Lemma \ref{UiL}.
- It does not connect the figure to the abstract formula \eqref{P123} or explain why
$M_j \cong P^4$.
- It contains no original analysis, only regurgitated text from the context and
summary.
- The LaTeX formatting is broken with missing commands and incomplete structures.
[...]
</feedback>
```

Solver

```
\\section{Topology of the carrying manifold}
In this section, we will prove Theorem \\ref{s3s1s}. To do this, note that in the
filtration \\eqref{fil}, $M^4\\setminus{ int}\\,M_{k-1}\\cong P^4$. Then, by
\\cite{Max}), it suffices to show that $M_{k-1}\\cong P^4$. Since $M_{k_0}$ contains
$k_0$ copies of $P^4$, the base case holds. Now suppose each connected component of
$M_{k_0+i} = M_{j-1}$ is homeomorphic to $P^4$. For the manifold $M_j$, its connected
components are either homeomorphic to $P^4$, or arise from applying formula (11) to
connected components of $M_{j-1}$. Denote by $U_l$ (where $l \in \{1,2\}$) the
connected component of $M_{j-1}$ that contains the solid torus $P_l^{T_j}$, and set
$A_l = \partial U_l$. Lemma 4 gives us $A_l \cong E$. From Lemma 2, we know that
removing the interior of one solid torus (say $P_2^{T_j}$) leaves a complement
homeomorphic to a solid torus: $(A_2 \setminus \text{int}\, P_2^{T_j}) \cong P^3$. We
now distinguish two cases: 1) $U_1 \neq U_2$; 2) $U_1 = U_2$. For case 1), we have $U
\cong U_1 \cup (P^3 \times [-1,1]) \cup U_2$. The induction hypothesis ensures that
$U_1 \cong U_2 \cong P^4$, while $P^3 \times \{-1\} = P_1^{T_j}$ and $P^3 \times \
{1\} = P_2^{T_j}$ (refer to Fig. 5). Since $(A_2 \setminus \text{int}\, P_2^{T_j})
\cong P^3$, we obtain $U_2 \cong P^3 \times [1,2]$. This yields $(P^3 \times [-1,1])
\cup U_2 \cong P^3 \times [-1,2]$, which implies $U \cong U_1$. For case 2), we have
$U \cong (P^3 \times [-1,1]) \cup U_2$, where the induction hypothesis gives $U_2
\cong P^4$ and $P^3 \times \{-1\} = P_1^{T_j}$, $P^3 \times \{1\} = P_2^{T_j}$ (see
Fig. 6). From $(A_2 \setminus \text{int}\, P_2^{T_j}) \cong P^3$, we deduce that $U_2
\cong P^3 \times [1,2]$.
```

Figure 34: **Example of ANAGENT Scientific Analysis Writing.** This example ($S_{\text{AVG}} = 72.41\%$) shows a complete end-to-end scientific analysis process of **ANAGENT**, illustrating the full scientific analysis writing workflow from task planning, searching and executing, context-aware problem-solving, to iterative reflection and refinement. Input query and agent prompts are provided in §D.2 and omitted here for clarity.

I Failure Analysis

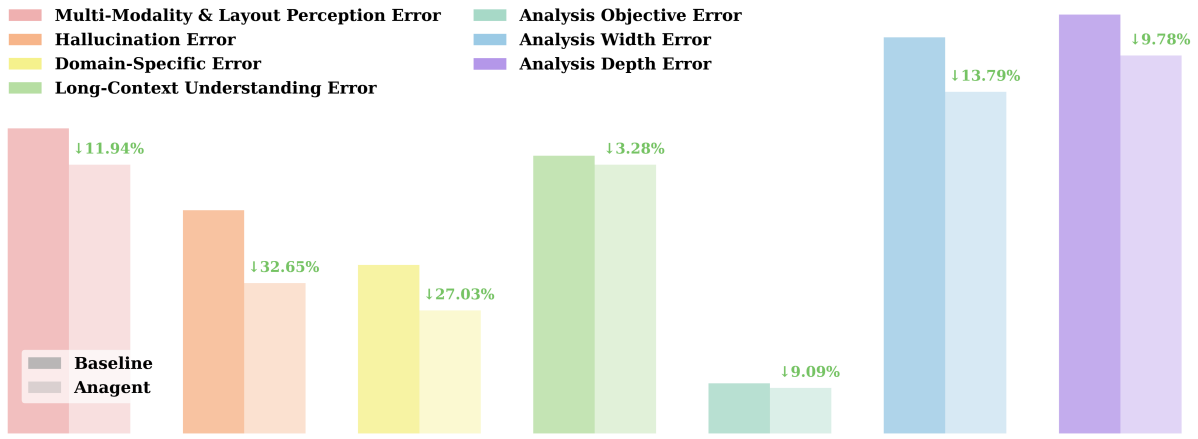


Figure 35: **Examination On Error Distribution.** Comparison of the seven key error patterns (Fig. 6) between baseline and **ANAGENT** on the same set. Similar to Fig. 6, the Y axis represents "Error Rate (%)" from low to high.

Following our preliminary exploration that reveals seven key error patterns in scientific table and figure analysis (Fig. 6), we extend the evaluation to **ANAGENT**, powered by the same backbone model `Qwen3-VL-8B` and tested on the identical 120-sample subset. Figure 35 compares and visualizes the distributions of the seven error patterns between baseline and **ANAGENT**. As can be seen from Fig. 35, **ANAGENT** consistently reduces error rates across all seven categories ($\Delta_{rel} \geq \downarrow 3.28\%$), demonstrating the effectiveness of integrating high-level planning with low-level knowledge acquisition and problem-solving through multi-agent collaboration for enhancing scientific table and figure analysis.

Building on these seven error patterns (Figs. 6-35), we conduct a systematic failure analysis and identify seven representative failure types corresponding to each error category, aiming to provide meaningful insights to inform future research.

I.1 Multi- Modality & Layout Perception Error

Perception errors are a major obstacle to accurate multimodal reasoning in MLLMs (Wang et al., 2025b). Although perception errors are notably decreased ($\Delta_{rel} = \downarrow 11.94\%$) by **ANAGENT**, there are still cases where **ANAGENT**'s scientific analysis receives low evaluation scores as a result of perception errors. Fig. 36 shows an example of a scientific analysis failure case ($S_{\text{AVG}} = 25.64\%$) with *multi- modality & layout error*.

I.2 Hallucination Error

Hallucination remains a known challenge in MLLMs and poses a significant source of errors for scientific table & figure analysis that requires faithful and reliable scientific generation. Although hallucination errors are remarkably reduced ($\Delta_{rel} = \downarrow 32.65\%$) by **ANAGENT**, there are still cases where **ANAGENT**'s scientific analysis writing receives low evaluation scores due to hallucination errors. Fig. 37 shows an example of a scientific analysis failure case ($S_{\text{AVG}} = 21.85\%$) with *hallucination error*.

I.3 Domain-Specific Error

Considering the domain-specific nature of scientific literature, acquiring accurate understanding of domain-specific tables and figures remains a significant challenge for MLLMs. While **ANAGENT** markedly reduces domain-specific errors ($\Delta_{rel} = \downarrow 27.03\%$), there are still cases where **ANAGENT**'s scientific analysis writing receives low evaluation scores due to limited domain knowledge or misunderstandings of domain-specific contents. Fig. 38 shows an example of a scientific analysis failure case ($S_{\text{AVG}} = 11.28\%$) with *domain-specific error*.

I.4 Long-Context Understanding Error

Scientific literature usually involves long-horizon contexts and cross-document references, posing significant challenges for MLLMs to effectively process, reason, and comprehend. While **ANAGENT** is able to reduce errors stemming from long-context understanding ($\Delta_{rel} =\downarrow 3.28\%$), it remains a high error rate (Fig. 35) for **ANAGENT**. For example, Fig. 39 shows a scientific analysis failure case ($S_{\text{AVG}} = 24.43\%$) as a result of *long-context understanding error*.

I.5 Analysis Width Error

Scientific literature contains heterogeneous elements, such as tables, figures, sections, citations, etc., that introduce varying scopes of information coverage. Although **ANAGENT** reduces errors stemming from analysis width ($\Delta_{rel} =\downarrow 13.79\%$), it still exhibits a relatively high error rate (Fig. 35). One representative failure case is shown in Fig. 40, where **ANAGENT** achieves a low scientific analysis score ($S_{\text{AVG}} = 25.96\%$) due to *analysis width errors*.

I.6 Analysis Depth Error

The roles and implications of tables and figures in scientific literature impose varying demands on the *depth* of analysis. Some only require shallow summarization directly derived from tabular or visual content, while others demand deeper reasoning, such as interpreting trends, drawing comparisons, or inferring conclusions supported by the presented evidence. Although **ANAGENT** reduces errors stemming from analysis depth ($\Delta_{rel} =\downarrow 9.78\%$), it still exhibits a relatively high error rate (Fig. 35). One representative failure case is shown in Fig. 41, where **ANAGENT** attains a low scientific analysis score ($S_{\text{AVG}} = 31.24\%$) due to *analysis depth errors*.

I.7 Analysis Objective Error

Tables and figures in scientific literature serve diverse analytical objectives. Some are designed to illustrate methodological designs, some to highlight benchmark innovations, and some to present empirical evidence in support of hypothetical claims, among a variety of scientific analysis objectives. Although **ANAGENT** reduces errors related to analysis objectives ($\Delta_{rel} =\downarrow 9.78\%$), correctly identifying and fulfilling these objectives remains challenging for MLLM agents (Fig. 35). A representative failure case is shown in Fig. 42, where **ANAGENT** achieves a low scientific analysis score ($S_{\text{AVG}} = 31.45\%$) due to *analysis objective errors*.

I.8 Other Errors Due To MLLM Backbone's Ability Constraints

The heterogeneity of multimodal content and the requirement for long-context understanding pose significant challenges for MLL agents. These challenges give rise to fundamental errors that profoundly impair scientific analysis, while stemming from the inherent ability constraints of the backbone MLLMs. Fig. I.8 illustrates two representative types of failures observed across different MLLMs: (a) *repetitive content* in **ANAGENT**'s output, and (b) incorrectly generated *intermediate solutions* by upstream agents prior to the scientific analysis writing performed by **SOLVER**. These phenomena reveal the significance of backbone MLLM's capacity in multi-agent systems for effective global context modeling, logical consistency, and accurate information consolidation.

J Limitations

By proposing **ANABENCH**, we introduce the task of scientific table & figure analysis, which is one of the core stages for scientific research. While **ANAGENT** demonstrates promising capabilities in scientific table & figure analysis, we acknowledge several limitations that we aim to address in our future work:

Computational Overhead. The multi-agent architecture with iterative execution and refinement introduces additional computational overhead as compared to single-pass generation. In practice, increasing the maximum number of iterations allowed for **ANAGENT** may consume more computational resources. To mitigate this, we implement engine pre-checking during the initialization of **ANAGENT** to avoid redundant MLLM engine initialization, and we plan to further optimize GPU utilization efficiency in future work.

Domain Coverage. Although ANAGENT spans nine scientific domains across 170 fine-grained disciplines, it primarily focuses on arXiv and PubMed publications. The generalization of ANAGENT to emerging scientific domains, non-English literature, and alternative dissertation platforms remains to be validated.

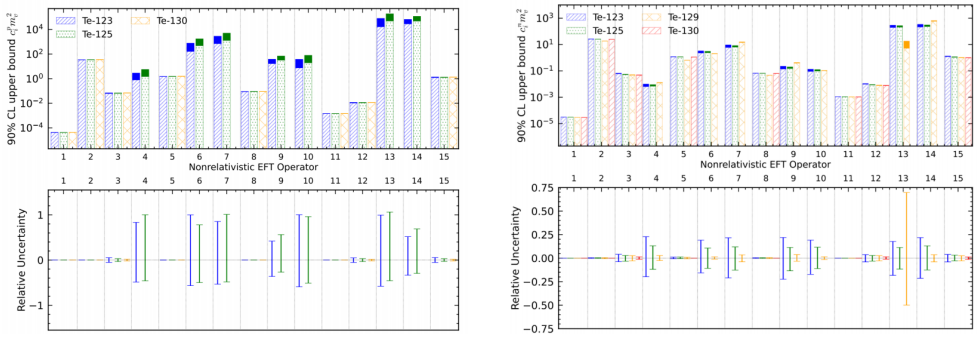
Tool Dependency. The effectiveness of ANAGENT is closely tied to the quality and reliability of provided scientific toolkits. Unsuccessful tool calls, as restrained by backbone MLLM reasoning capabilities or truncated tool execution outputs due to MLLM context window limits, can directly impact analysis quality. These challenges motivates us to explore tool utilization and optimization in our future work.

Evaluation Challenges. While we employ multi-faceted evaluation including rule-based metrics, MLLM-as-Judge, and human expert assessment, the cost of MLLM-as-Judge and human expert assessment can be very large that substantially limits evaluation efficiency. We aim to investigate systematic evaluation with low costs in future research.

Despite these limitations, we believe ANABENCH and ANAGENT represents a meaningful step toward more capable and reliable AI systems for scientific research, and we hope our work will inspire future investigations into scientific reasoning and analysis.

Error 1: Multi- Modality & Layout Perception Error

Input:



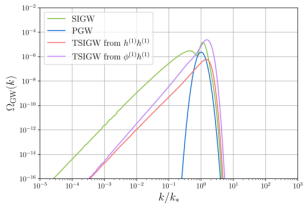
ANAGENT Output:

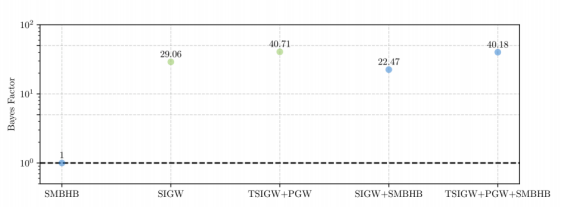
```
The results of our Gaussian ensemble uncertainty quantification calculation for \TeO are
shown in Figs.~\ref{te_p_uncerts} and \ref{te_n_uncerts}. We also plot the relative
uncertainty of each EFT operator on the bottom panel of Fig.~\ref{te_n_uncerts}. We note
that the relative uncertainty of each EFT operator is not simply the ratio of the
absolute uncertainty to the upper bound, but the ratio of the difference in the upper
bound from the nominal value to the upper bound. This is because the nominal value is
the upper bound that would be obtained if the nuclear shell model inputs were perfectly
known, whereas the absolute uncertainty is the difference between the upper bound
obtained with nuclear shell model inputs and the nominal value. We plot the relative
uncertainty rather than the absolute uncertainty because the absolute uncertainty is not
a meaningful measure of the uncertainty in the upper bound. For example, the upper bound
for $c^p_4$ in Te-125 is about 0.4, which is about 400.
```

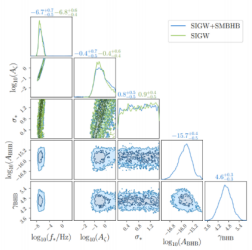
Figure 36: **Example of Multi- Modality & Layout Perception Error.** This example ($S_{\text{AVG}} = 25.64\%$) illustrates a scientific analysis failure case with *multi- modality & layout perception errors* (§I.1). Input query and agent prompts are provided in §D.2 and omitted here for clarity. Other intermediate outputs and contexts are also omitted for clear presentation.

Error 2: Hallucination Error

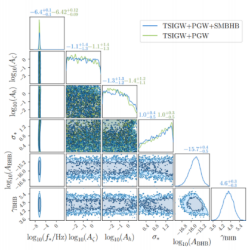
Input:



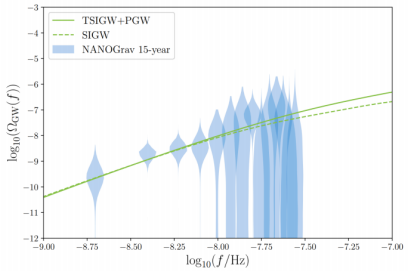




(a) SIGWs and SIGWs+ SMBHB.



(b) PGWs+TSIGWs and PGWs+TSIGWs+ SMBHB.



ANAGENT Output:

```
\section{Bayes factor}
The Bayesian evidence quantifies the relative likelihood of different models given the
observed data. A higher Bayes factor indicates stronger support for the compared model.
We compare five models: SMBHB (baseline), SIGW, TSIGW+PGW, SIGW+SMBHB, and
TSIGW+PGW+SMBHB. The vertical axis represents the Bayes factor of different models
relative to SMBHB, and the horizontal axis represents the different models. The green
dots are for models without SMBHB and the blue dots are for models in combination with
the SMBHB signal. As shown in Fig.~\ref{bayes}, the model combining TSIGW+PGW+SMBHB
achieves the highest Bayes factor of 40.18, followed by TSIGW+PGW at 40.71, indicating
that these models are strongly favored over SMBHB. In contrast, SIGW and SIGW+SMBHB have
lower Bayes factors (29.06 and 22.47, respectively), suggesting that SMBHB is less
likely to dominate PTA observations than the combined TSIGW+PGW signal. The inclusion of
SMBHB in the TSIGW+PGW model (TSIGW+PGW+SMBHB) does not diminish its evidence strength,
as the Bayes factor remains high (40.18), implying that the combined signal is robust
even when accounting for SMBHB. The posterior distributions shown in
Figs.~\ref{corner_SIGW} and~\ref{corner_TSIGW} further support this conclusion,
revealing that when large-amplitude PGW on small scales are considered, the median
amplitude $A_{\zeta}$ for the tensor-scalar perturbations is significantly reduced,
consistent with the high Bayes factors for the TSIGW+PGW models. This analysis is
grounded in Bayesian inference using the KDE representations of the NANOGrav 15-year PTA
data and the \textsc{bilby} package with \textsc{dynesty} nested sampling, ensuring
statistical rigor. The results align with the paper's broader conclusion that TSIGW+PGW
are more likely to dominate current PTA observations than SMBHB, highlighting the
importance of considering tensor-scalar induced GWs for future PTA parameter estimation.
The degeneracy between $A_{\zeta}$ and $A_h$ observed in the posterior distributions,
which is significant only for log-normal power spectra with large $\sigma_*$, further
underscores the need for multi-model Bayesian evidence to disentangle parameter
degeneracies and robustly constrain the primordial GW background.
```

Figure 37: **Example of Hallucination Error.** This example ($S_{\rm AVG} = 21.85\%$) illustrates a scientific analysis failure case with *hallucination errors* (§I.2). Input query and agent prompts are provided in §D.2 and omitted here for clarity. Other intermediate outputs and contexts are also omitted for clearer presentation.

Error 3: Domain-Specific Error

Input:

#unsolved / n	4	5	6	7	8	9	10	Total
Shapley value	0.00%	0.00%	0.00%	1.50%	1.75%	6.75%	3.50%	1.93%
Nucleolus	0.00%	0.00%	0.00%	0.50%	0.00%	0.00%	0.25%	0.11%
Benefit value	0.25%	0.25%	1.25%	0.25%	0.75%	0.75%	0.00%	0.50%
Contribution value	0.75%	0.75%	1.75%	1.00%	1.50%	0.50%	2.25%	1.21%
Banzhaf value	0.00%	0.75%	1.50%	1.75%	0.50%	4.25%	4.50%	1.89%
Tau value	0.00%	0.25%	0.00%	0.00%	0.00%	1.00%	0.00%	0.18%
Banzhaf* value	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.50%	0.07%
total: d1	0.00%	0.00%	0.00%	0.00%	0.00%	0.17%	0.00%	0.02%
total: lexmin	0.17%	0.33%	0.50%	1.33%	0.83%	4.17%	3.33%	1.52%
total: d1+c	0.00%	0.00%	0.00%	0.00%	0.00%	0.14%	0.00%	0.02%
total: lexmin+c	0.43%	0.86%	2.14%	1.71%	1.86%	3.71%	3.29%	2.00%

#unfinished / n	4	5	6	7	8	9	10	Total
Shapley value	0.00%	0.00%	0.00%	0.50%	0.00%	0.50%	2.00%	0.43%
Nucleolus	0.00%	0.00%	0.00%	0.00%	0.00%	0.75%	0.00%	0.11%
Benefit value	0.25%	0.00%	0.75%	0.00%	0.00%	0.25%	0.75%	0.29%
Contribution value	0.00%	0.00%	0.50%	0.00%	0.50%	0.50%	0.25%	0.25%
Banzhaf value	0.00%	1.00%	0.25%	0.25%	1.25%	0.50%	2.00%	0.75%
Tau value	0.00%	0.00%	0.00%	0.00%	0.25%	0.00%	0.00%	0.04%
Banzhaf* value	0.00%	1.00%	0.00%	0.00%	0.50%	0.50%	2.00%	0.57%
total: d1	0.00%	0.17%	0.00%	0.00%	0.00%	0.17%	0.00%	0.05%
total: lexmin	0.00%	0.17%	0.33%	0.33%	0.50%	0.50%	1.67%	0.50%
total: d1+c	0.00%	0.29%	0.14%	0.00%	0.14%	0.00%	0.00%	0.08%
total: lexmin+c	0.14%	0.29%	0.43%	0.14%	0.71%	1.00%	2.00%	0.67%

ANAGENT Output:

```
Table~\ref{table:not_solved} shows that overall incompletion rates are relatively low
across all solution concepts, with most remaining under 2\%. The Shapley value shows the
highest total incompletion rate at 1.93\%, followed closely by the Banzhaf value at
1.89\%. This suggests that the Shapley value requires significantly more computational
resources than other solution concepts, making it less practical for applications
requiring real-time coalition formation.

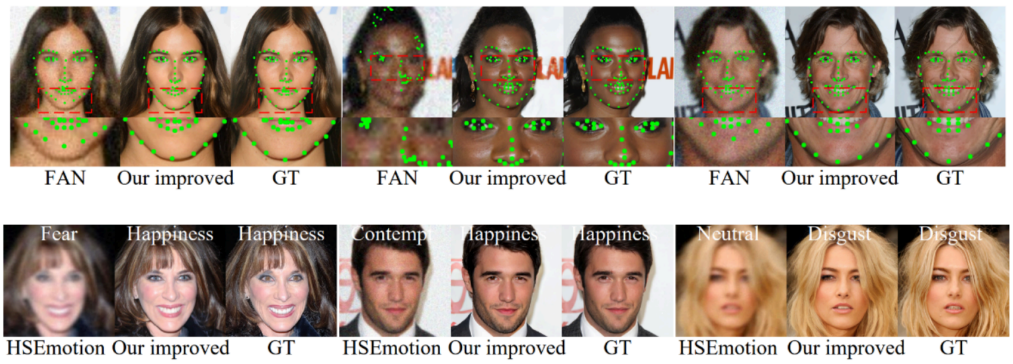
Examining the scenario-based totals at the bottom of the table, the lexmin+c scenario
exhibits the highest overall incompletion rate at 2.00\%. The d1 and d1+c scenarios
achieve nearly perfect completion at 0.02\%, indicating that these simpler allocation
mechanisms converge much faster than the more complex lexmin-based approaches. The plain
lexmin scenario shows a moderate 1.52\% incompletion rate, suggesting that adding the
coalition structure constraint (+c) increases computational difficulty by approximately
0.5 percentage points.
```

Figure 38: **Example of Domain-Specific Error.** This example ($S_{\text{AVG}} = 11.28\%$) illustrates a scientific analysis failure case with *multi- modality & layout perception errors* (§I.3). Input query and agent prompts are provided in §D.2 and omitted here for clarity. Other intermediate outputs and contexts are also omitted for clear presentation.

Error 4: Long-Context Understanding Error

Input:

	FAN	Our improved
DSR% ↑	99.73%	**100.00%**
NME% ↓	6.08%	**2.43%**



ANAGENT Output:

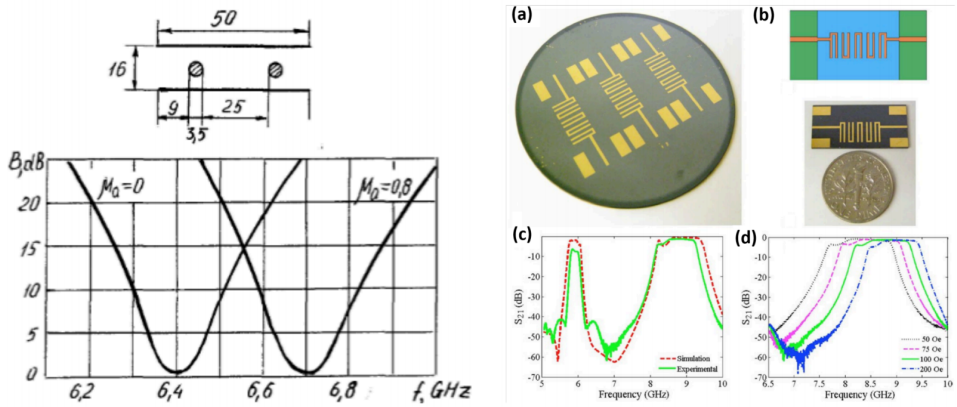
```
\subsubsection{Facial Landmark Detection}

We further evaluate our method on facial landmark detection. The results are shown in
Table~\ref{application_detection} and Fig.~\ref{fig_landmark}. The proposed method
demonstrates superior performance in detecting facial landmarks, achieving a detection
success rate of 100.
To evaluate our method in face emotion recognition, we employ the HSEmotion model
\\cite{hsemotion}. The results are shown in Fig.~\\ref{fig_emo_rec}. Our method
demonstrates superior performance in face emotion recognition, achieving an accuracy of
82. The face emotion recognition task evaluates the model's ability to accurately
classify facial expressions. The facial landmark detection task evaluates the model's
ability to accurately identify key facial features, such as eyes, nose, and mouth. These
results are consistent with our method's strong performance in face emotion recognition.
```

Figure 39: **Example of Long-Context Understanding Error.** This example ($S_{\text{AVG}} = 24.43\%$) illustrates a scientific analysis failure case due to *long-context understanding errors* (§I.4). Input query and agent prompts are provided in §D.2 and omitted here for clarity. Other intermediate outputs and contexts are also omitted for clear presentation.

Error 5: Analysis Width Error

Input:



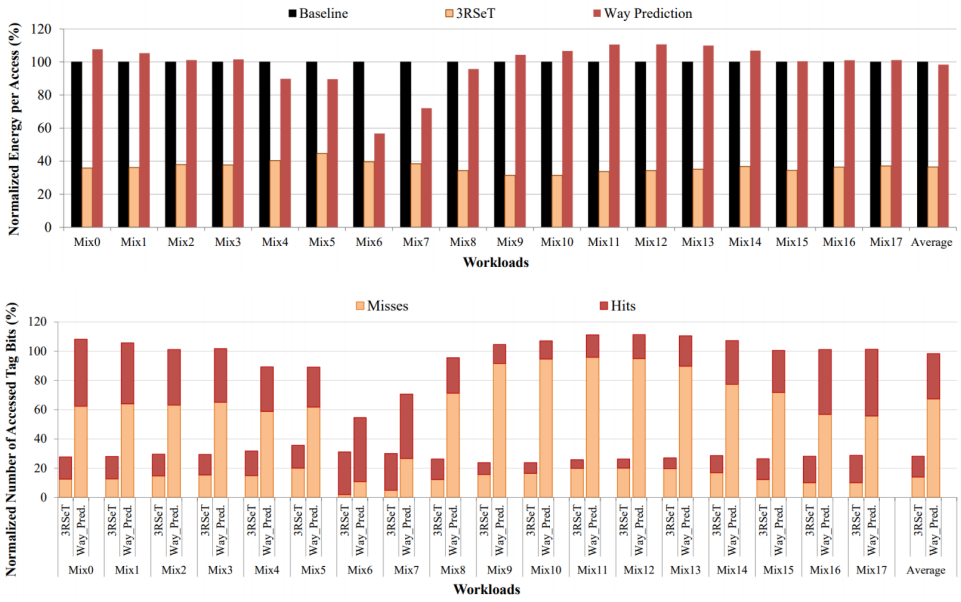
ANAGENT Output:

```
\section{MSSW filters}
The first application of YIG for RF filters was the two-order ferrite filter presented
in 1990~\cite{yu1990tunable}. This device, depicted in Fig.~\ref{7}, consists of two YIG
cylinders with a diameter of 3 mm and a height of 1.5 mm, placed symmetrically in the
waveguide. The filter response, characterized by insertion loss and VSWR, exhibits two
passbands, each centered at a specific frequency, with a minimum insertion loss of
approximately 0.5 dB. The tunability of the filter is achieved by applying a magnetic
bias field, which shifts the center frequency of the passbands. The filter operates with
a VSWR of 1.5 and maintains low insertion loss even at higher frequencies, making it
suitable for mobile relay equipment. The performance of ferrite filters is significantly
improved by using polycrystalline YIG, which enhances power handling by three orders of
magnitude compared to single-crystal YIG. This is attributed to the higher microwave
absorption in polycrystalline materials, which reduces the power density and minimizes
the risk of thermal breakdown, allowing the filter to handle higher power levels without
damage~\cite{Harris2012}.In 1996, Yamanouchi and Qureshi proposed the use of a YIG-based
surface acoustic wave (SAW) filter for the 5-15 GHz range~\cite{yamanouchi19975}. This
device, based on an electrode thickness difference type and a new reflector bank type of
unidirectional interdigital transducers, demonstrated a low insertion loss of 1.5 dB and
a high return loss of 30 dB. The filter's performance was attributed to the use of a YIG
layer on a quartz substrate, which provided a high electromechanical coupling
coefficient and enabled the device to operate efficiently at high frequencies. In 2010,
Hara et al. developed a super-high-frequency band filter using air-gap-type thin-film
bulk acoustic resonators (BAMR)~\cite{hara2010super}. The device, fabricated with a 3.3
mm-thick YIG film on a Si substrate and an air gap of 3 mm, achieved a high center
frequency of 3.5 GHz with a low insertion loss of 2.0 dB. The filter's performance was
further enhanced by the use of a 200 nm-thick gold film on the YIG surface, which
increased the electromechanical coupling coefficient and improved the device's
sensitivity to external magnetic fields.
```

Figure 40: **Example of Analysis Width Error.** This example ($S_{\text{AVG}} = 25.96\%$) illustrates a scientific analysis failure case due to *analysis width errors* (§I.5). Input query and agent prompts are provided in §D.2 and omitted here for clarity. Other intermediate outputs and contexts are also omitted for clear presentation.

Error 6: Analysis Depth Error

Input:



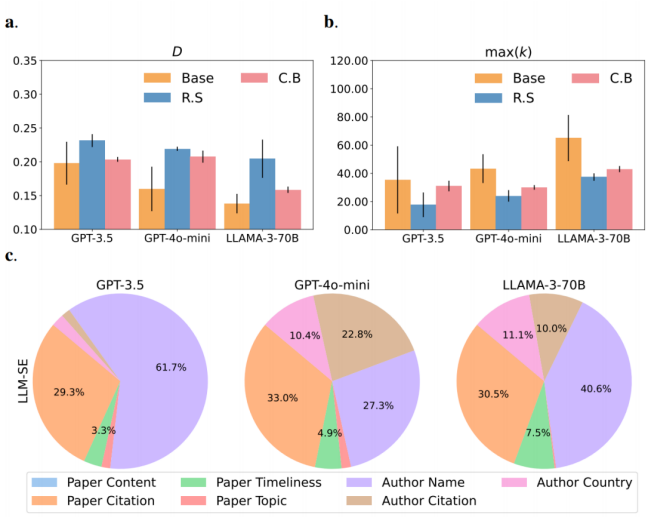
ANAGENT Output:

```
\subsubsection{Energy Consumption}
The energy consumption of the tag array is shown in Fig.~\ref{11}. The average energy
consumption of 3RSeT is 37.9. The energy consumption of way prediction is only 1.6. The
energy consumption reduction in 3RSeT is due to the significant reduction in the number
of tag accesses, as shown in Fig.~\ref{8}. On the other hand, the energy consumption of
way prediction is increased due to the high misprediction rate. The misprediction
penalty requires reading the entire tag array to find the correct way, which consumes
more energy than the baseline. Therefore, the energy consumption of way prediction is
not lower than the baseline, as shown in Fig.~\ref{11}. The energy consumption of way
prediction is 1.6. The energy consumption of way prediction is not lower than the
baseline because the energy consumption of way prediction is increased due to the high
misprediction rate. The misprediction penalty requires reading the entire tag array to
find the correct way, which consumes more energy than the baseline. Therefore, the
energy consumption of way prediction is not lower than the baseline, as shown in
Fig.~\ref{11}.
```

Figure 41: **Example of Analysis Depth Error.** This example ($S_{\text{AVG}} = 31.24\%$) illustrates a scientific analysis failure case due to *analysis depth errors* (§I.6). Input query and agent prompts are provided in §D.2 and omitted here for clarity. Other intermediate outputs and contexts are also omitted for clear presentation.

Error 7: Analysis Objective Error

Input:



ANAGENT Output:

```
\subsection{LLM-LE: LLMs generate citation networks with power-law distributions}
The first experiment, LLM-LE, aims to analyze the power-law distribution of citation
networks generated by LLMs. The power-law distribution is a common characteristic of
citation networks \cite{boerner2007network}. In citation networks, the power-law
distribution is often attributed to the Barabási-Albert (B-A) model
\cite{barabasi1999emergence}, which is based on the principle of preferential attachment.
Preferential attachment refers to the phenomenon where highly cited papers are more
likely to be referenced. The B-A model suggests that the probability of a paper being
cited is proportional to its current number of citations. This mechanism leads to a
power-law distribution in the citation network, where a few papers receive a large number
of citations, while the majority of papers receive only a few citations. The power-law
distribution of citation networks has been widely observed and studied, and it has
important implications for understanding the dynamics of citation behavior and the
structure of scientific knowledge. In the B-A model, the probability of a paper being
cited is proportional to its current number of citations. The power-law distribution of
citation networks has been widely observed and studied, and it has important implications
for understanding the dynamics of citation behavior and the structure of scientific
knowledge. Figure \ref{fig3}a shows the $D$ values for the generated citation networks
under all experimental conditions. The results demonstrate that all models exhibit a
power-law distribution in the generated citation networks. This finding is consistent
with the Barabási-Albert model and suggests that the power-law distribution in citation
networks may be influenced by the citation information that LLMs can access. Furthermore,
the results show that the power-law fitness of the generated citation networks is
significantly influenced by citation information visibility. The random paper
recommendation setting, which anonymizes the citation data of the papers, has a more
significant impact on the power-law fitness than the anonymizing citation data setting.
This suggests that the recommendation algorithm may play a more significant role in
shaping the citation network than the citation bias itself.
```

Figure 42: **Example of Analysis Objective Error.** This example ($S_{\text{AVG}} = 31.45\%$) illustrates a scientific analysis failure case due to *analysis objective errors* (§I.7). Input query and agent prompts are provided in §D.2 and omitted here for clarity. Other intermediate outputs and contexts are also omitted for clear presentation.

Error 8: MLLM Backbone Capability Constraints

(a) Repetitive Content

ANAGENT (Backbone MLLM: InternVL3.5-4B)

Solver outputs repetitive string "The 90":

```
\section{Analysis of Neutron Coupling Uncertainties in Tellurium Isotopes}
The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The
90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The
90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The
90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The
90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90\The 90 [...]
```

ANAGENT (Backbone MLLM: Qwen3-VL-4B)

Solver's output contains repetitive contents occasionally:

```
\subsection{Variable Rate Spraying with Constant Flow Rate}
This section is a direct continuation of the analysis presented in Fig.
\ref{Drone_Path_Hotspot_1_1} and Fig. \ref{Height_Map_1_1}. The figures depict a
drone's path over a farmland, with a focus on its flight altitude, which is adjusted
to optimize pesticide application. The drone is programmed to fly at a constant
altitude for areas with low disease intensity, ensuring a uniform spray rate.
However, when the drone approaches a high-probability disease hotspot, as identified
by the graph-based method, it descends to a lower altitude. This altitude adjustment
is a direct response to the hotspot detection, ensuring that the drone delivers a
higher concentration of pesticide at the target location, thereby optimizing resource
utilization and enhancing the efficacy of the spraying operation. The sharp rise in
altitude at approximately the 50-unit mark in Fig. \ref{Drone_Path_Hotspot_1_1} is a
clear indicator of the drone's adaptive response to disease intensity, reflecting the
core functionality of the SprayCraft method: variable rate spraying based on disease
hotspot detection. The constant altitude for the rest of the path, as shown in Fig.
\ref{Height_Map_1_1}, demonstrates the drone's ability to maintain efficiency in non-
hotspot areas, where a constant spray rate is sufficient, thereby conserving
resources and minimizing unnecessary pesticide application. This analysis is
consistent with the paper's methodology, which leverages a graph-based approach to
identify disease hotspots and compute an optimal drone path using the TSP. The
altitude profile, as depicted in the figures, is a direct outcome of this
methodology, demonstrating how the system dynamically adjusts to disease intensity,
ensuring that resources are used efficiently and effectively. The figures, therefore,
are not just representations of the drone's flight path but visualizations of the
system's intelligent response to disease intensity, showcasing the core innovation of
the SprayCraft approach: adaptive, precision-based spraying. The sharp rise in
altitude at the 50-unit mark is a direct consequence of the system's ability to
adjust to disease intensity, ensuring that the drone delivers the optimal amount of
pesticide at the hotspot, while maintaining efficiency in non-hotspot areas. The
constant altitude for the remainder of the path reflects the drone's ability to
maintain a standard spray rate in areas with lower disease intensity, thereby
conserving resources while maintaining the required level of coverage. [...] The
sharp rise in altitude at the 50-unit mark is a direct consequence of the system's
ability to adjust to disease intensity, ensuring that the drone delivers the optimal
amount of pesticide at the hotspot, while maintaining efficiency in non-hotspot
areas. The constant altitude for the remainder of the path reflects the drone's
ability to maintain a standard spray rate in areas with lower disease intensity,
thereby conserving resources while maintaining the required level of coverage. This
adaptive altitude profile is a key innovation of the SprayCraft method, allowing for
efficient, precision-based spraying that is responsive to disease intensity rather
than relying on a fixed spray rate across the entire farmland. [...]
```

(b) Intermediate Solution

ANAGENT (Backbone MLLM: Qwen2.5-VL-3B)

Planner generates final analysis solution, instead of problem-solving plans:

```
GHOST coronagraphic PSFs. Left: the PSF without any turbulence, and the DM set to be
flat. Right: the PSF with simulated 1-stage systems residual phase screens played on
SLM, and a flat DM. \jalo{The speckle at around 1 o'clock is a ghost in the
system.}}\end{figure}In Figure \ref{ghost} we show the coronagraphic PSFs of the
GHOST system with and without simulated turbulence. The speckle at around 1 o'clock
is a ghost in the system. Ghost speckles are a direct consequence of the residual
phase screens introduced by the 1st stage AO correction, which are not fully
corrected by the 2nd stage AO correction. Ghost speckles are an artifact of the
system and are a direct consequence of uncorrected phase errors and the actuator
structure of the DM. Ghost speckles are a critical indicator of residual wavefront
error and highlight the need for robust control strategies. The PO4AO method, trained
using a dynamics model and policy optimization, aims to suppress such artifacts by
learning to predict and correct for phase errors in real-time. The PSF shown in
Figure \ref{ghost} demonstrates the system's sensitivity to phase error propagation
and the importance of effective AO correction for exoplanet imaging. Table
\ref{table:po4ao_param} lists adjustable PO4AO parameters, including episode length
(500 frames), warm-up episodes (20), and planning horizon (4 frames), which are
critical for training and performance.
```

ANAGENT (Backbone MLLM: Qwen3-VL-4B)

Expert generates final analysis solution, instead of problem-solving plans:

```
The provided figure illustrates the radial spectral density $\rho(r)$ for increasing
matrix size $N$ in the class AI$^{\dagger}$ of complex symmetric random matrices, as
derived analytically and validated numerically. The exact expression for $\rho(r)$,
Eq.~(\ref{P(r)}), exhibits a characteristic behavior: near the origin, it scales as
$\rho(r) \sim \frac{N}{N+1}r$, peaks around $r \approx \sqrt{2}$, and decays
exponentially for $r \gg 1$. Numerical simulations over $10^6$ realizations for
Gaussian matrices (histograms) confirm this behavior, with the exact result (solid
lines) matching the simulations for small $N$ (e.g., $N=2,3,5$) and showing
convergence to the large-$N$ approximation (dashed line for $N=50$). The limiting
behavior as $N \to \infty$ is characterized by the triangular law for $\rho(r)$,
corresponding to the circular law for $\rho(r)/r$, a universal feature distinct from
the Ginibre ensemble, as evidenced by the asymptotic edge profile in Eq.~
(\ref{edgeprof}) and its deviation from Ginibre behavior \cite{fyodorov2018CMP}. The
convergence of the spectral density to this triangular law underscores the emergence
of a unique universality class for complex symmetric matrices, particularly at the
spectral edge, where both eigenvalue density and eigenvector statistics diverge from
standard Ginibre predictions. This behavior is further supported by numerical
evidence for Bernoulli random matrices, indicating potential universality beyond
Gaussian ensembles. The figure's structure—showing the transition from finite $N$ to
the large-$N$ limit—highlights the interplay between precise analytic results and
stochastic simulations, reinforcing the theoretical framework's predictive power in
the context of dissipative quantum systems and non-ergodic wave transport.",
```

Figure 43: **Example of Other Errors Due To MLLM Backbone Ability Constraints.** This figure shows two scientific analysis failure cases due to *MLLM backbone ability constraints* (§I.8): (a) Repetitive Content, (b) Intermediate Solution. Input query and agent prompts are provided in §D.2 and omitted here for clarity. Some contents denoted as `[...]` are omitted for presentation brevity.

Prompt for Five-Dimensional Evaluation Protocol

You are an expert evaluator for scientific {data_type} analysis writing tasks. Given a {data_type} and its ground-truth analysis, evaluate whether a model-generated analysis is of high quality.

Ground-truth Analysis
{gt_analysis}

Model Analysis
{model_analysis}

Evaluation Criteria
1. **Content Accuracy**: Does the model analysis accurately and correctly reflect the explicitly presented information in the {data_type}, such as reported values, trends, stated findings, and observed patterns, without distortion or omission?
2. **Analytical Completeness**: Does the model analysis sufficiently develop the derived analyses, interpretations, and conclusions implied by the {data_type}, capturing key insights without missing important analytical findings?
3. **Format Correctness**: Are there any format errors in model analysis that violate the task requirements, such as incorrect structure, typographical errors, or improper formatting?
4. **Clarity & Coherence**: Is the model analysis clearly articulated, properly structured, and logically coherent, with sound scientific writing throughout?
5. **Reliability & Faithfulness**: Are there any hallucinated information in model analysis that is never covered, analyzed, and inferred in the ground-truth analysis? NOTE that both referred contents (e.g., referring to other tables/figures/sections/equations/etc., within the paper or referring to other papers) or reasonably inferred contents are considered as grounded (CAREFUL: analysis writing SHOULD have inferred contents, as long as they are reasonable and inferrable from the given information in such as way that are covered by ground-truth analysis)

Instructions
- Compare the **Model Analysis** against the **Ground-truth Analysis**
- Focus on the quality of model analysis writing in the five dimensions of **Evaluation Criteria**
- Consider semantic similarity and information coverage, NOT exact word matching
- Rate the overall quality on a scale from 0-2 (where 2 is excellent)

Response Format

```
<think>your evaluation reasoning</think>
<accuracy>grade 0-2 for **Content Accuracy**: 0 = mostly inaccurate; 1 = partially accurate; 2 =
    fully accurate</accuracy>
<completeness>grade 0-2 for **Analytical Completeness**: 0 = most key analyses missing; 1 = some
    key analyses missing; 2 = all key analyses present</completeness>
<format>grade 0-2 for **Format Correctness**: 0 = many format errors; 1 = some format errors; 2 =
    no format error</format>
<writing>grade 0-2 for **Clarity & Coherence**: 0 = poorly written; 1 = moderately clear; 2 =
    clear and coherent</writing>
<faithfulness>grade 0-2 for **Reliability & Faithfulness**: 0 = mostly hallucinated; 1 = partially
     hallucinated; 2 = fully reasonable</faithfulness>
```

Figure 44: **Prompt For Five-Dimensional Evaluation Protocol.** During LLM-as-judge and human expert evaluation (§2.4), the evaluator is prompted to assess the quality of each model-generated scientific analysis in five dimensions, including *content accuracy*, *analytical completeness*, *format correctness*, *clarity & coherence*, and *reliability & faithfulness*.

Prompt for Planner Agent

You are a Task Planning agent specialized in analyzing and task problems, decomposing tasks into concrete subtasks, and generating high-level plans that can guide resolving the given tasks.

Task Problem
`{task_problem}`
Additional Contexts
`{task_context}`

What You Need To Do
1. Understand the **Task Problem** and its requirements, and analyze the **analysis width, depth, and objective** of this task
2. Analyze what contextual information and domain-specific knowledge are needed to resolve this task
3. Decompose complex tasks into clear, actionable subtasks
4. Create a concise and instructive **Problem-Solving Plan** that can effectively guide resolving the **Task Problem**

Requirements
- Be systematic, thorough, and strategic in your planning
- As contextual information and domain-specific knowledge are highly important, please ALWAYS highlight "*Find, retrieve, distill, and summarize all the related contexts*" at the very beginning of your plan
- Please specify finding and adding necessary citations and references to support the analysis completeness and format accuracy of the final answer
- Your **Problem-Solving Plan** should be around `{plan_limit}` characters, consisting several subtasks in bullet-point format using "*" WITHOUT indents
- Provide both your reasoning and planning in the following format with proper enclosure:

```
<think>your reasoning and analysis</think>
<plan>your **Problem-Solving Plan**</plan>
```

Figure 45: **Prompt For PLANNER Agent.** PLANNER directs task-oriented planning, providing systematic guidance for improved analysis accuracy and completeness.

Prompt for Expert Agent (PART I)

You are an Expert Scientist equipped with various tools for information retrieval and knowledge collection.
Your job is to iteratively gather task-specific information by:
1. Analyze the given **Task Problem** and the corresponding **Problem-Solving Plan**
2. Decide which ONE tool to use in each turn
3. Receive and summarize tool execution results
4. Repeat until you have gathered sufficient information to solve the given task
You can ONLY use ONE tool per turn. Think carefully about which tool can provide the most useful information based on what you already know. After gathering all the relevant contexts and domain knowledge by calling different tools, in the end, you should generate a concise task-specific **Knowledge Summary** with all the important information needed to answer the **Task Problem**.

Task Problem
`{task_problem}`

Prompt for Expert Agent (PART II)

Additional Contexts
`{task_context}`
Problem-Solving Plan
`{planner_plan}`
Available Tools
`{tool_info}`

What You Need To Do
You have {max_turns} turns to gather information. This is TURN 1. In each turn, you have two choices:
1. **Use A Tool** to gather more information:
   - Analyze the task and current state to decide which tool to use. You can ONLY use ONE tool in each turn.
   - To call a tool, please follow these steps: (a) First, provide your analysis and reasoning, enclosed in: `<think>your analysis of what information is needed, which specific tool you choose for current turn, and why</think>`; (b) Second, specify your selected tool for this turn, enclosed in: `<tool>tool_name</tool>`; (c) Third, provide your tool-specific query and options with proper enclosure as required in **Available Tools**.
2. **Stop and Summarize** if you believe your collected information is sufficient to resolve the **Task Problem**

Instructions
- Carefully review and analyze the **Task Problem** and its requirements, the **Problem-Solving Plan**, and all your collected information via tool calling
- DO NOT ANSWER THE **Task Problem**: You ONLY need to analyze and summarize all your collected information based on your reasoning, analysis, observations, key findings, and any other highly important information that you believe can support solving the **Task Problem**
- BE MINDFUL about the required depth of the **Task Problem**: Analyze carefully the **analysis width, depth, and objective** of this task, e.g., whether it asks about shallow description or in-depth analysis, academic writing or technical report, scientific discovery or creative generation, etc.
- Your **Knowledge Summary** should: (1) Be both concise and informative, consisting of around `{summary_len}` characters; (2) Use Markdown format but in free style that you believe best suits current task, e.g., plain text, bullet points, or a mix of both, etc.; (3) **Ensure The Accuracy and Quality of Your Summary:** Ensure the faithfulness of you **Knowledge Summary**, incorporating ONLY factual information of significance directly from the collected contexts and domain knowledge, avoiding self-generated, self-interpreted, self-inferred, or rephrased details; (4) Always use the correct format in line with the input table/figure. For example, if the input uses LaTeX, you SHOULD use Latex Bib and `\cite` for reference citations, and specify Latex `\ref` when referring to other tables/figures/sections/equations/etc.; (5) Organize a well-structured summary that includes both high-level insights and detailed information, analysis, observations, and key findings, etc. For example: `<think>your reasoning</think><summary>{summary_structure}</summary>`

Please generate your reasoning and **Knowledge Summary** enclosed in:

```
<think>your reasoning</think>
<summary>your knowledge summary</summary>
```

Figure 46: **Prompt For EXPERT Agent.** EXPERT leads information searching and retrieval to supply task-specific contextual and domain knowledge.

Prompt for Solver Agent (Initial Generation)

You are a Expert Scientist that writes high-quality scientific analysis according to the given **Task Problem**.

Task Problem
`{task_problem}`
Additional Contexts
`{task_context} # Expert knowledge summary included`
Problem-Solving Plan
`{planner_plan}`

What You Need To Do
1. Analyze the **Task Problem** and its requirements, and conclude the **analysis width, depth, and objective** of this task
2. Review the **Problem-Solving Plan** and **Additional Contexts**
3. Reason and analyze all the available information and knowledge systematically and thoroughly
4. Generate your solution enclosed in: `<answer>your solution</answer>`

Requirements
- Please ensure the accuracy, precision, completeness, profession, and quality of your scientific analysis
- Please carefully review all the given information and contexts, and generate semantically coherent, logically structured, and scientifically grounded analysis
- Always use the correct format in line with the input table/figure. For example, if the input uses LaTeX, you SHOULD use Latex Bib and `\cite` for reference citations, and specify Latex `\ref` when referring to other tables/figures/sections/equations/etc.
- Provide your reasoning and solution in the following format with proper enclosure:

  `<think>your reasoning</think>`
  `<answer>your scientific analysis</answer>`

Figure 47: **Prompt for the SOLVER Agent.** As supported by PLANNER and EXPERT, SOLVER integrates important contextual information and domain knowledge into task inputs to generate initial scientific analysis, and collaborates with the CRITIC to iteratively reflect on, correct, and refine the analysis. CRITIC feedback is omitted here for prompt presentation clarity, and is added to the prompt during iterative reflection and refinement.

Prompt for Critic Agent

You are a CRITIC ADVISOR responsible for evaluating and improving STUDENT's **STUDENT Answer** to the given **Task Problem** based on available **Additional Contexts**.

Task Problem
`{task_problem}`
Additional Contexts
`{task_context}`
STUDENT Problem-Solving Plan
`{planner_plan}`
STUDENT Answer
`{solver_solution}`

Evaluation Criteria
1. **Content Accuracy**: Does the model analysis accurately and correctly reflect the explicitly presented information in the `{data_type}`, such as reported values, trends, stated findings, and observed patterns, without distortion or omission?
2. **Analytical Completeness**: Does the model analysis sufficiently develop the derived analyses, interpretations, and conclusions implied by the `{data_type}`, capturing key insights without missing important analytical findings?
3. **Format Correctness**: Are there any format errors in model analysis that violate the task requirements, such as incorrect structure, typographical errors, or improper formatting?
4. **Clarity & Coherence**: Is the model analysis clearly articulated, properly structured, and logically coherent, with sound scientific writing throughout?
5. **Reliability & Faithfulness**: Are there any hallucinated information in model analysis that is never covered, analyzed, and inferred in the ground-truth analysis? NOTE that both referred contents (e.g., referring to other tables/figures/sections/equations/etc., within the paper or referring to other papers) or reasonably inferred contents are considered as grounded (CAREFUL: analysis writing SHOULD have inferred contents, as long as they are reasonable and inferrable from the given information in such as way that are covered by ground-truth analysis)

What You Need To Do
1. Analyze the **Task Problem** and its requirements, and conclude the **analysis width, depth, and objective** of this task
2. Review the **Additional Contexts**, reasoning and analyzing all the available information and knowledge systematically and thoroughly
3. Evaluate STUDENT's **STUDENT Answer** according to the five **Evaluation Criteria**
4. Generate your detailed improvement guidance feedback to STUDENT, which should at least cover the five evaluation dimensions above, and less than `{feedback_limit}` characters

Response Format

```
<think>your evaluation reasoning</think>
<accuracy>grade 0-2 for **Content Accuracy**: 0 = mostly inaccurate; 1 = partially accurate; 2 =
    fully accurate</accuracy>
<completeness>grade 0-2 for **Analytical Completeness**: 0 = most key analyses missing; 1 = some
    key analyses missing; 2 = all key analyses present</completeness>
<format>grade 0-2 for **Format Correctness**: 0 = many format errors; 1 = some format errors; 2 =
    no format error</format>
<writing>grade 0-2 for **Clarity & Coherence**: 0 = poorly written; 1 = moderately clear; 2 =
    clear and coherent</writing>
<faithfulness>grade 0-2 for **Reliability & Faithfulness**: 0 = mostly hallucinated; 1 = partially
     hallucinated; 2 = fully reasonable</faithfulness>
<feedback>your feedback</feedback>
```

Figure 48: **Prompt For CRITIC Agent.** Cooperating with SOLVER, CRITIC also employs the five-dimensional evaluation protocol (Fig. 44 & §A.2) for reflective correction and refinement.