

SciExplore: Evaluating Autonomous Agents from Scientific Navigation to Information Integration

Yinhao Tang^{1,2*} Youqing Fang^{1,2*} Yanan Sun^{2†}
Wenran Liu² Weiming Zhang¹ Bin Liu¹
Kuikun Liu² Wenwei Zhang² Kai Chen^{2†}

¹University of Science and Technology of China ²Shanghai AI Laboratory
{tangyinhao, fangyq}@mail.ustc.edu.cn, {sunyanan, chenkai}@pjlab.org.cn

Abstract

Scientific research involves complex information-seeking and reasoning workflows across heterogeneous sources. However, existing benchmarks primarily emphasize general-domain retrieval or static scientific question answering, and therefore fail to assess key capabilities required in realistic scientific research workflows. We introduce SciExplore, a benchmark designed to evaluate scientific information-seeking and reasoning capabilities of LLMs and agents. SciExplore comprises four task types covering 103 expert-curated tasks across more than ten scientific disciplines: scientific database navigation, ambiguous literature retrieval, missing reference completion, and cross-source structured knowledge synthesis, which probe progressively higher-level abilities from entity-level reasoning and document-level identification to evidence-level grounding and domain-level synthesis. We evaluate over ten state-of-the-art LLMs and autonomous agents on SciExplore, revealing substantial performance gaps with performance degrading sharply as task complexity increases and extremely low accuracy on the most challenging structured synthesis tasks. These results highlight significant limitations of current models and agents in realistic scientific information-seeking scenarios.

1 Introduction

As scientific discovery becomes increasingly data-intensive and interdisciplinary, Large Language Models (LLMs) are envisioned as autonomous scientific assistants (Schmidgall et al., 2025; Baek et al., 2025; Chai et al., 2025; Li et al., 2025a; Wu et al., 2025a; Li et al., 2025c; Bran et al., 2023; Li et al., 2025b; Huang et al., 2025) capable of supporting complex research workflows. These workflows extend well beyond fact lookup (Yang

et al., 2018; Wei et al., 2025; Zhou et al., 2025a) or question answering (Saikh et al., 2022; Wang et al., 2023; Yue et al., 2024), encompassing tasks such as navigating specialized databases, identifying relevant literature under ambiguous descriptions, grounding claims in appropriate evidence, and synthesizing structured knowledge across heterogeneous sources. A central challenge for the community, therefore, is to rigorously evaluate whether current LLMs or agents can reliably perform these research behaviors in realistic settings.

Despite rapid progress, existing evaluation paradigms remain poorly aligned with the demands of authentic scientific research. On one hand, deep search benchmarks (Wei et al., 2025; Zhou et al., 2025a; Wu et al., 2025b; Wong et al., 2025) primarily target general-domain information retrieval and open-web exploration, overlooking the structured databases, domain-specific conventions, and methodological reasoning intrinsic to scientific inquiry. Deep Research benchmarks (Du et al., 2025a) extend this paradigm by evaluating agents through long-form research report generation, with primary emphasis on the completeness and fluency of written reports rather than fine-grained evidence grounding or structured knowledge synthesis. On the other hand, science-oriented QA benchmarks (Saikh et al., 2022; Wang et al., 2023; Yue et al., 2024; He et al., 2024; Guo et al., 2023; Cui et al., 2025; Du et al., 2025b; Zhou et al., 2025b; Wang et al., 2022) typically reduce evaluation to static question answering in closed environments, abstracting away crucial steps such as evidence discovery, literature disambiguation, and cross-source integration. As a result, neither paradigm adequately measures the capabilities required of LLMs as research assistants—namely, evidence-driven, multi-step reasoning grounded in real, noisy, and heterogeneous scientific information environments.

To address this gap, we introduce SCIEXPLORE,

* Equal contribution. † Corresponding author.

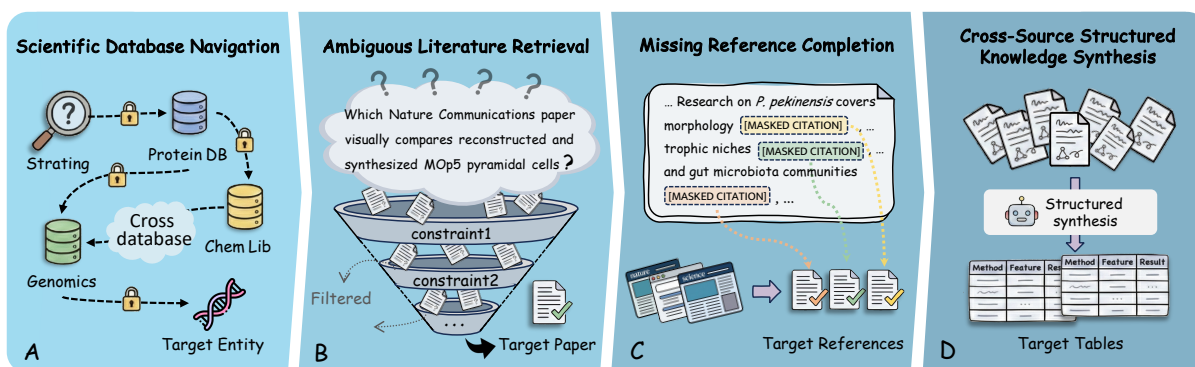


Figure 1: **Overview of SCIEXPLORE.** The benchmark evaluates scientific information seeking as a progressive cognitive process, spanning four task types that advance from (A) entity-level reasoning, to (B) document-level literature identification, (C) evidence-level reference grounding, and finally (D) domain-level knowledge synthesis.

an expert-curated benchmark for evaluating LLMs or agents in authentic scientific information-seeking and reasoning workflows. SCIEXPLORE is designed based on the observation that scientific information seeking is a progressive cognitive process where researchers incrementally move from locating individual entities to identifying relevant documents, grounding claims in evidence, and synthesizing structured knowledge at the domain level.

Accordingly, SCIEXPLORE comprises four task types, each targeting a distinct and essential scientific capability, as illustrated in Fig. 1. **(T1) Scientific Database Navigation** evaluates the foundational ability to traverse structured scientific databases, requiring multi-step reasoning to locate precise entities and attributes across interconnected records. **(T2) Ambiguous Literature Retrieval** builds on this foundation by shifting from databases to the scientific literature, challenging agents to identify a specific key paper based on vague, incomplete, or noisy methodological descriptions, without relying on explicit keywords. **(T3) Missing Reference Completion** further increases difficulty by moving from single-document identification to evidence-level grounding: given a scientific paragraph with missing citations, agents must correctly align multiple claims with their appropriate supporting references. **(T4) Cross-Source Structured Knowledge Synthesis** represents the highest level of abstraction, requiring agents to aggregate evidence from a broad set of papers and synthesize these findings into a structured comparison table that captures a domain-level overview of a research topic.

This study outlines a closed-loop evaluation of scientific information seeking, moving systematic-

cally from entity retrieval to knowledge synthesis, involving 103 PhD-level tasks curated by experts across various disciplines. The platform, SCIEXPLORE, maintains strict difficulty levels and ensures answer uniqueness, requiring deep domain understanding and multi-step reasoning instead of simple memorization.

We assessed 12 advanced LLMs and search agents on SCIEXPLORE, revealing significant performance gaps: most achieved success rates under 50%, with a sharp decline in effectiveness as task complexity increased. Notably, the toughest task—Cross-Source Structured Knowledge Synthesis—showed even top agents with accuracy lower than 20%. These results indicate that while current systems manage basic retrieval and shallow reasoning, they significantly falter in complex scientific research that requires comprehensive evidence, precise grounding, and accurate synthesis.

In summary, our contributions are threefold:

- We introduce SCIEXPLORE, a novel benchmark that evaluates agents along a hierarchy of scientific reasoning capabilities, spanning entity-level database navigation, document-level literature identification, evidence-level reference grounding, and domain-level structured knowledge synthesis.
- We propose a rigorous, expert-driven data construction and validation framework that enforces high reasoning depth, strong domain specificity, and answer uniqueness, effectively preventing shortcut-based or memorization-driven solutions.
- Through extensive evaluation of state-of-the-art agents, we expose substantial limitations

Table 1: Comparison of SCIEXPLORE with existing benchmarks. Abbreviations: **Sci. K. & R.**: Scientific Knowledge and Reasoning; **Sci. DB & Lit.**: Scientific Database and Literature; **ST**: Short Text; **MC**: Multi Choice; **Ref.**: Bibliographic Reference.

Benchmark	Domain	Sci. K. & R.	Search Depth	Search Width	Search Source	Output Format
<i>Deep Search and Deep Research Benchmarks</i>						
HotpotQA (Yang et al., 2018)	General	Low	Low	Low	Wikipedia	ST
BrowseComp (Wei et al., 2025)	General	Low	High	Low	Open Web	ST
WideSearch (Wong et al., 2025)	General	Low	Low	High	Open Web	Table
DeepResearchBench (Du et al., 2025a)	General	Low	Middle	Middle	Open Web	Report
<i>Scientific QA Benchmarks</i>						
ScienceQA (Saikh et al., 2022)	Science	Low	Low	Low	Open Web	MC
SciBench (Wang et al., 2023)	Science	Middle	Middle	Low	Open Web	ST
SuperGPQA (Du et al., 2025b)	Science	High	High	Low	Open Web	ST
HLE (Phan et al., 2025)	Math & Science	High	High	Low	Open Web	ST
<i>Our proposed</i>						
SCIEXPLORE	Science	High	High	High	Sci. DB & Lit.	ST & Ref. & Table

in current systems when deployed in realistic scientific research scenarios, particularly their inability to reliably integrate multi-source evidence into precise, structured outputs.

2 Related Works

Deep Search and Deep Research Benchmarks. Recent benchmarks for evaluating search agents in the general domain have progressively moved beyond single-hop fact retrieval toward deeper multi-step search, planning, and synthesis. Early QA datasets (Kwiatkowski et al., 2019; Yang et al., 2018) focus on limited multi-hop reasoning, while later benchmarks explicitly introduce tool use and long-horizon search behaviors (Mialon et al., 2023; Chen et al., 2025). More recent work expands the evaluation scope along complementary dimensions. BrowseComp-EN/ZH (Wei et al., 2025; Zhou et al., 2025a) emphasize search depth and robustness to noise and entity ambiguity, whereas WideSearch (Wong et al., 2025) stresses search breadth and structured aggregation from large-scale web sources. DeepResearchBench (Du et al., 2025a) further pushes this paradigm by requiring agents to synthesize multi-source evidence into long-form research-style reports.

Despite these advances, existing deep search benchmarks remain largely confined to the general web domain, relying on open-web sources and common-sense fact aggregation, and thus failing to evaluate core scientific assistant capabilities such as database navigation, interpretation of ambiguous methodological contexts, and evidence-level grounding. Moreover, synthesis is typically assessed through free-form text, avoiding the stricter requirement of producing precise, structured scien-

tific artifacts with high-fidelity evidence alignment (Table 1), leaving a gap in evaluating agents in authentic scientific information-seeking workflows.

Scientific QA Benchmarks. Scientific QA benchmarks primarily evaluate models’ parametric scientific reasoning under closed-book or minimally grounded settings. Early benchmarks target high-school or undergraduate-level science questions (Saikh et al., 2022; Hendrycks et al., 2020), while more recent datasets such as SciBench (Wang et al., 2023) introduce college-level problems involving multi-step calculations. Advanced benchmarks including GPQA (Rein et al., 2024), SuperGPQA (Du et al., 2025b), and HLE (Phan et al., 2025) further push difficulty toward expert-level and frontier scientific reasoning.

While effective at measuring reasoning depth, existing scientific QA benchmarks do not explicitly evaluate information seeking, evidence discovery, or cross-source validation, as high performance can often be achieved using parametric knowledge alone. In contrast, SCIEXPLORE is constructed to favor problems whose resolution requires substantial retrieval and filtering from external scientific sources, encouraging agents to acquire missing knowledge and ground their reasoning in externally obtained evidence for structured scientific synthesis.

3 SCIEXPLORE

To operationalize the desiderata above, SCIEXPLORE instantiates scientific information-seeking as four complementary tasks, curated through a fully expert-driven construction and verification pipeline. We formalize this task taxonomy in Section 3.1, detail the corresponding data construction

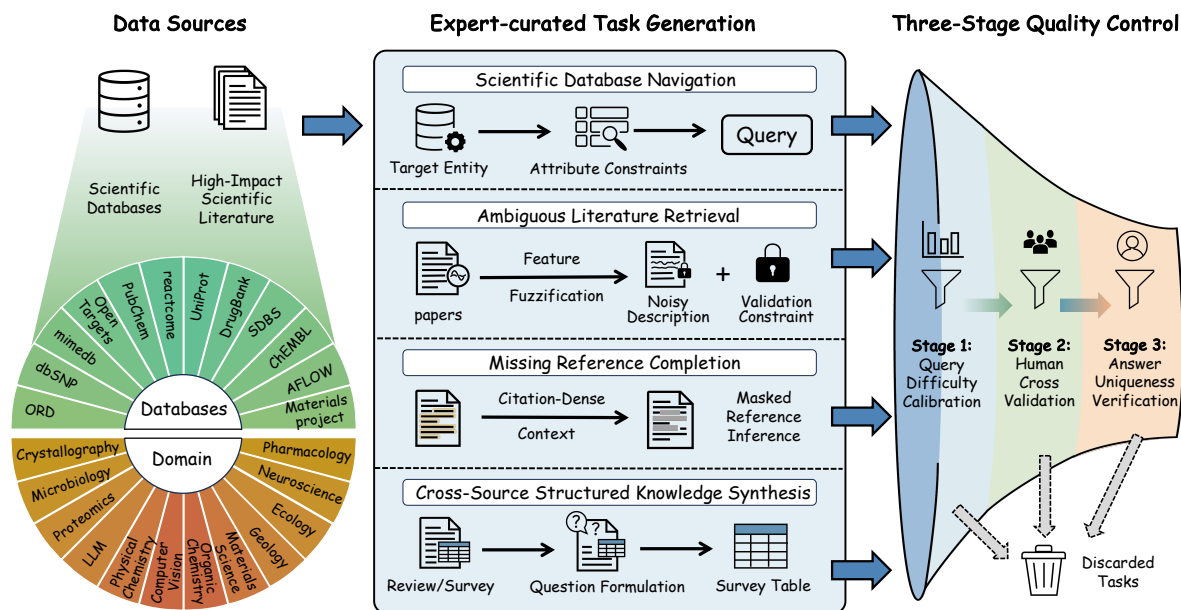


Figure 2: **Dataset construction pipeline of SCIEXPLORE.** The pipeline illustrates expert-driven task design, multi-stage validation, and rigorous quality control across T1–T4.

procedure in Section 3.2, and describe our quality control mechanisms in Section 3.3. Detailed data statistics and evaluation metrics are further provided in Appendix A.3 and Appendix A.5, respectively.

3.1 Task Definition

Table 2 elaborates on the taxonomy of tasks in SCIEXPLORE, which spans a progression from entity-level retrieval to cross-source structured synthesis and is designed to jointly cover the core information-seeking competencies required in authentic scientific research workflows.

Specifically, T1, Scientific Database Navigation, helps researchers identify, filter, and disambiguate relevant entities in complex databases, which is crucial for efficiently locating specific data. T2, Ambiguous Literature Retrieval, aids users in disambiguating intent and discriminating among documents to eliminate irrelevant ones. Together, these tasks assess the ability of LLMs or agents to locate precise information in dense scientific environments. T3, Missing Reference Completion, ensures the integrity of scientific claims. This task includes claim abstraction, aligning claims with supporting evidence, and validating relevant evidence, thereby enriching the quality of research by ensuring adequate support for claims. Finally, T4, Cross-source Structured Knowledge Synthesis, represents a higher-order capability, involving domain abstraction, information extraction, and the

integration of information from multiple sources for structured synthesis. Collectively, T3 and T4 probe the evidence-grounding and cross-source integration competencies that go beyond mere information location, complementing T1–T2 and together forming a comprehensive evaluation suite that mirrors the end-to-end workflow of a scientific research assistant.

3.2 Dataset Construction

SCIEXPLORE is constructed through a fully expert-driven process to ensure realism, answer uniqueness, and resistance to shortcut retrieval. All task instances are manually curated by domain experts and designed to reflect authentic scientific research behaviors, rather than synthetic or template-based generation, as shown in Fig. 2.

Scientific Database Navigation. For T1, we design tasks that require genuine multi-step database navigation using a *Reverse Trajectory Construction Strategy*, where explicit entity identifiers are progressively replaced by nested attribute-based constraints. This design enforces structured reasoning over interconnected database records and prevents direct lookup.

Ambiguous Literature Retrieval. In T2, experts apply *Feature Denoising and Fuzzification* to deliberately obscure surface-level cues, while preserving answer uniqueness through *Validation Constraint Injection*, which enables post-retrieval veri-

Table 2: Task taxonomy of SCIEXPLORE, detailing the task types, corresponding capability levels, atomic capabilities, task counts, and associated metrics.

No.	Task Type	Capability Level	Atomic Capabilities	Tasks	Metric
T1	Scientific Database Navigation	Entity-Level Reasoning	Entity identification, filtering, disambiguation, multi-constraint navigation	39	Accuracy
T2	Ambiguous Literature Retrieval	Document-Level Identification	Intent disambiguation, document-level discrimination and elimination	32	Accuracy
T3	Missing Reference Completion	Evidence-Level Grounding	Claim abstraction, claim–evidence alignment, evidence discovery and validation	14	Accuracy
T4	Cross-source Structured Knowledge Synthesis	Domain-Level Synthesis	Domain abstraction, information extraction, cross-source integration, structured synthesis	18	Recall

fication without guiding the search process.

Missing Reference Completion. For T3 tasks, we construct citation reconstruction tasks that require aligning scientific claims with their appropriate supporting references. Surface-level textual overlap is minimized, ensuring that successful prediction depends on understanding claim–evidence relationships rather than keyword matching.

Cross-Source Structured Knowledge Synthesis. Tasks in T4 are derived from expert-curated comparison schema and require agents to integrate heterogeneous evidence across multiple papers into a strictly formatted table. This setting emphasizes cross-source consistency, schema compliance, and high-precision synthesis.

A detailed description of the construction procedure and validation criteria for each task is provided in Appendix A.4.

3.3 Quality Control

To ensure the rigor of SCIEXPLORE, we implemented a streamlined three-stage quality control process: query difficulty calibration, human cross-validation and answer uniqueness verification.

Query Difficulty Calibration. To minimize parametric memory reliance, we strictly enforce two principles: *Explicit Index Blocking* (removing unique identifiers to compel semantic matching) and *Long-tail Entity Preference* (prioritizing obscure entities). This ensures tasks require deep retrieval reasoning rather than simple knowledge recall.

Human Cross-Validation. Annotators independently attempt to solve each question using search engines. Questions are retained only if no solution can be found within a strict 10-minute limit, ensuring sufficient difficulty for the target models.

Answer Uniqueness Verification. We utilize search-enabled LLMs (e.g., Gemini-3-Pro, o3) to generate candidate solutions, which are then re-

viewed by human annotators. If any alternative answer is satisfied, the task is deemed ambiguous and discarded.

4 Experiments

4.1 Experimental Setup

We evaluate a wide range of systems on SCIEXPLORE to assess the reliability of AI assistants in realistic scientific workflows, spanning three paradigms: Foundation LLMs, LLMs with Search Tools, and Specialized Deep Research Agents. **Foundation LLMs** are evaluated to measure intrinsic scientific reasoning and parametric knowledge. These include open-source models DeepSeek-V3.2 (Liu et al., 2025), Qwen2.5-72B-Instruct (Team, 2024), Qwen3-30B-A3B-Thinking, and Qwen3-235B-A22B-Thinking (Yang et al., 2025), as well as closed-source models GPT-5.1 (OpenAI, 2025), Gemini-2.5-Pro (Google Gemini, 2025), and Gemini-3-Pro (Google, 2025b). To examine the effect of retrieval augmentation, we further evaluate **LLMs with Search Tools** by enabling search API or web browsing for GPT-5.1 and Gemini-3-Pro. Finally, we benchmark **Deep Research Agents** designed for autonomous long-horizon scientific inquiry, including Tongyi-DeepResearch-30B-A3B, Gemini Deep Research (Google, 2025a), and OpenAI Deep Research (OpenAI, 2024). Together, this setup isolates the roles of parametric knowledge, retrieval augmentation, and agentic planning in multi-step scientific problem solving.

4.2 Main results

Deep Research agents consistently outperform foundation LLMs and retrieval-augmented LLMs across all task types. As shown in Table 3, this advantage holds across both standalone LLMs and search-augmented baselines. Across model families, closed-source systems generally outper-

Table 3: Main results on the SCIEXPLORE. T4 is reported at two granularities: item- and row-level. All scores are percentages (%).

Model	Category	T1	T2	T3	T4		Overall
					Item	Row	
<i>Foundation LLMs</i>							
DeepSeek-V3.2	Open-Source	20.51	6.25	5.95	8.32	2.25	11.06
Qwen2.5-72B-Instruct	Open-Source	7.69	0.00	0.00	8.63	6.94	4.99
Qwen3-30B-A3B-Thinking	Open-Source	10.26	0.00	3.57	2.54	1.60	4.50
Qwen3-235B-A22B-Thinking	Open-Source	20.51	3.12	0.00	7.85	4.15	7.87
GPT-5.1	Closed-Source	17.95	9.38	9.66	8.76	4.83	12.83
Gemini-2.5-Pro	Closed-Source	20.51	6.25	4.17	10.57	3.78	11.25
Gemini-3-Pro	Closed-Source	23.08	12.50	13.69	16.23	7.29	19.08
<i>LLMs with Search Tools</i>							
DeepSeek-V3.2 w/ search	Open-Source	17.95	9.38	9.66	9.20	3.95	11.12
GPT-5.1 w/ search	Closed-Source	15.38	34.38	20.85	12.13	5.12	21.61
Gemini-3-Pro w/ search	Closed-Source	33.33	53.12	49.40	23.53	11.49	46.46
<i>Deep Research Agents</i>							
Tongyi-DeepResearch-30B-A3B	Open-Source	43.59	50.00	34.40	11.12	4.26	36.93
Gemini Deep Research	Closed-Source	35.90	56.25	36.56	15.61	7.35	39.30
OpenAI Deep Research	Closed-Source	23.08	43.75	59.91	30.14	18.59	49.39

form open-source models, with particularly pronounced advantages on tasks involving ambiguous retrieval, citation reasoning, and cross-source synthesis. At the entity level, Tongyi-DeepResearch-30B-A3B achieves the strongest performance on Scientific Database Navigation. For Ambiguous Literature Retrieval, Gemini Deep Research attains the highest accuracy. OpenAI Deep Research achieves the best results on Missing Reference Completion and Cross-source Structured Knowledge Synthesis, and obtains the highest overall score among all evaluated systems. In some cases, search augmentation degrades performance by amplifying hypothesis-first reasoning rather than correcting it, leading to lower accuracy than non-search baselines (see Appendix 12).

There remains a substantial gap before current agents can function as reliable autonomous scientific assistants. Despite the relative advantages of Deep Research agents observed above, evaluation results on SCIEXPLORE indicate that existing systems still far from reliably fulfilling the role of autonomous scientific assistants. As shown in Table 3, even the strongest system, OpenAI Deep Research, achieves an Overall Score of only 49.39%, while most evaluated systems remain below 20%. These limitations stem from the long and heterogeneous cognitive chain required by SCIEXPLORE, which extends far beyond conventional scientific question answering.

Domain-level structured synthesis reveals critical weaknesses in long-horizon reasoning and

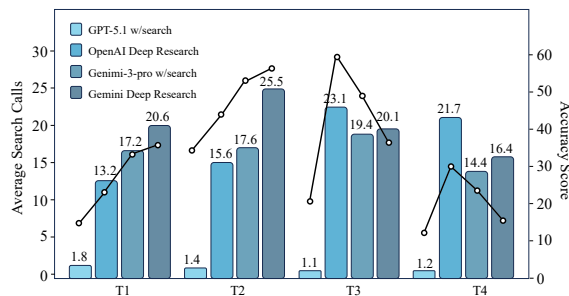


Figure 3: Average Search Calls per Type.

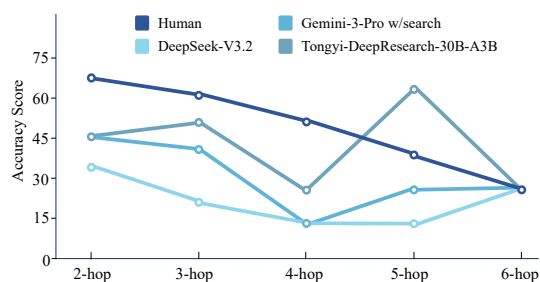


Figure 4: Performance on multi-hop T1 tasks with varying hop lengths.

cross-source integration abilities. Across all evaluated systems, performance on domain-level synthesis remains consistently low, indicating demands well beyond entity- or document-level reasoning. Although Table 3 reports moderate item-level scores—for example, OpenAI Deep Research correctly fills 30.14% of individual cells—this partial correctness does not scale to coherent structured outputs. Models often recover isolated facts

Table 4: T4 formatting accuracy (Format Acc) and primary key recall (PK Recall) across evaluated systems. All values are reported as percentages.

Model	Format_Acc	Recall
<i>LLMs</i>		
DeepSeek-V3.2	94.44	20.15
Qwen2.5-72B-Inst	100.00	16.86
Qwen3-30B-Think	88.89	3.44
Qwen3-235B-Think	94.44	14.95
GPT-5.1	100.00	18.23
Gemini-2.5-Pro	94.44	19.94
Gemini-3-Pro	100.00	19.05
<i>LLMs with Search Tools</i>		
DeepSeek-V3.2 w/ search	88.89	17.01
GPT-5.1 w/ search	100.00	22.02
Gemini-3-Pro w/ search	100.00	21.42
<i>Deep Research Agents</i>		
Tongyi-DeepRes-30B	77.78	19.9
Gemini Deep Research	55.56	29.84
OpenAI Deep Research	88.89	30.04

from multiple sources, yet fail to assemble them into complete and internally consistent scientific representations, revealing a gap between local correctness and global structure.

5 Analysis

5.1 Search Effort and Task Difficulty

To understand the performance differences noted in Section 4, we analyze the relationship between agents’ information-seeking behavior and task accuracy by examining the correlation between average search calls and performance across systems and task types.

Figure 3 displays the average search calls per question for each model on tasks T1–T4. A clear trend emerges: systems with more frequent searches achieve significantly higher accuracy on SCIEXPLORE. For instance, models with minimal search usage, like GPT-5.1-search, average fewer than two search calls per question and show low success rates. In contrast, Deep Research agents, which utilize a larger search budget, consistently outperform others across all tasks. This analysis highlights that search intensity is crucial for scientific activities. Effective scientific assistance requires agents to actively explore and verify information through repeated searches, rather than relying on shallow retrieval.

The number of retrieval hops is not correlated with task difficulty. Table 6 shows a counterintuitive finding: increasing hop length in scientific database navigation does not consistently degrade

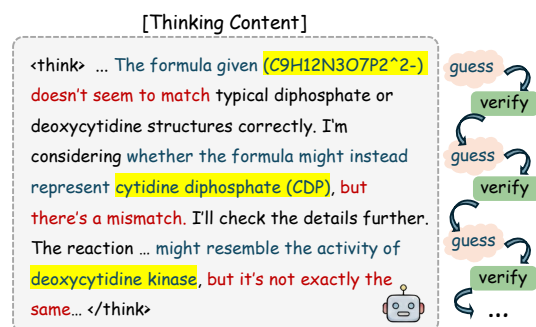


Figure 5: An example of guess-and-verify thinking process. “Guess” and “verify” are highlighted in blue and red, respectively.

performance across models. In some cases, performance remains steady or even improves with higher hop counts. This indicates that the human-defined concept of “search depth”—which implies increasing difficulty with more hops—does not reliably translate to the challenges faced by LLM-based agents. Thus, hop length is an imperfect measure of the cognitive or operational complexities involved in task execution.

This discrepancy arises from differences in problem-solving approaches between humans and models. Human experts typically refine their searches using structured filters and explicit entities, while LLM agents often employ a guess-and-verify strategy as shown in Fig. 5: they hypothesize a target entity based on parametric knowledge and then use retrieval queries mainly to confirm or refute that hypothesis.

5.2 Error Analysis

To deeply analyze the root causes of the poor performance of current systems in SCIEXPLORE, we conducted an in-depth qualitative and quantitative analysis of typical failure cases in the experiments. The analysis reveals that existing models, when dealing with real-world scientific tasks, are primarily constrained by the following four key defects. Representative empirical examples for each failure mode are illustrated in Appendix A.7 (Fig. 8–11).

Premature Abandonment in the Face of Ambiguity. In scientific exploration, perseverance in the face of ambiguous or incomplete cues is crucial for researchers. However, we observe that agents exhibit “search inertia” when confronted with complex problems, often deeming tasks as “unsolvable” and prematurely terminating searches after few attempts. As shown in Fig. 8, agents fre-

quently interpret the lack of an immediate semantic match as proof of non-existence, failing to broaden the query space or reformulate hypotheses. This premature abandonment fundamentally contradicts the exploratory nature of scientific inquiry.

Hallucinated Responses after Search Stagnation. A more concerning issue than explicit refusal is the emergence of hallucinated answers following prolonged search stagnation. Our analysis reveals that when multi-step reasoning fails to yield conclusive evidence, some models begin to relax their internal verification mechanisms. In later reasoning stages, they may invent or overgeneralize details to make partial retrievals appear relevant to the original query. Fig. 9 illustrates this, showing unverified document attributes that are fabricated to meet strict user constraints. In scientific research, where rigor is paramount, such behavior severely undermines trustworthiness.

Information Loss in Long Contexts. Even when agents locate the correct literature, extracting precise answers from lengthy scientific documents remains a significant challenge. Scientific texts are often dense, with critical evidence buried under extensive background or methodology. While models can retrieve the correct web pages or PDFs, their limitations in long-context attention hinder them from isolating fine-grained details. As shown in Fig. 10, models tend to rely on high-level summaries, such as abstracts, while overlooking contradictory or corrective information deeper in the text. This lack of "needle-in-a-haystack" capability ultimately fractures the reasoning chain during final extraction.

Instruction Following Mismatch in Structured Output. Cross-source structured knowledge synthesis reveals significant deficiencies in instruction following within rigid output schemas. Despite Markdown tables with predefined headers required, many models output headers omissions, column misalignment, and unit inconsistencies. Fig. 11 exemplifies a schema alignment failure, where required fields are omitted for a simplified summary-style table. As shown in Table 4, this issue is especially notable in Gemini Deep Research, which, despite strong retrieval performance, shows a "report generation bias" with only 55.56% accuracy in table format. Such structural instability severely limits practical usability in automated scientific pipelines that depend on strict schemas for downstream processing.

Low Primary-Key Recall Reflects Insufficient Search Breadth. In addition to formatting and instruction-following issues, poor performance on T4 is primarily constrained by insufficient primary-key recall. Table 4 demonstrates that even the best systems achieve recall rates below 60%, meaning many required core entities are not retrieved. In cross-source structured knowledge synthesis, missing primary keys lead to unrecoverable failures: the absence of essential row entities renders table construction and alignment infeasible, regardless of downstream extraction accuracy. This indicates that current systems struggle to conduct retrieval with adequate breadth and systematic coverage in scientific domains, directly limiting their ability to perform high-quality structured knowledge synthesis in T4.

6 Discussion and Conclusion

We introduce SCIEXPLORE, a benchmark for evaluating large language models as autonomous scientific assistants in realistic research workflows. It assesses multi-hop database navigation, fuzzy literature retrieval, supported reference inference, and structured table synthesis, bridging the gap between general deep-search benchmarks and static scientific question answering. Experiments with ten state-of-the-art systems reveal significant shortcomings, particularly in structured scientific synthesis, indicating that advancements in retrieval and tool use alone are not sufficient for reliable scientific assistance.

Our evaluation identifies a critical bottleneck: while current AI systems excel in general retrieval, they lack the systematic breadth and cognitive resilience required for effective autonomous scientific research. Future development may prioritize three areas: (1) models must become planners with broader search capabilities; (2) agents need to be robust against search ambiguity; and (3) "premature abandonment" and hallucination highlight the importance of handling noisy retrieval results without fabricating evidence. Finally, there is an urgent need for schema-constrained long-context understanding to enable accurate extraction of structured data from full-text PDFs. Ultimately, the next generation of scientific AI must optimize for answer correctness and process reliability in long-horizon inquiry. We hope SCIEXPLORE can serve as a foundation for future research on improving scientific agents under realistic, open-ended conditions.

References

- Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738.
- Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldasari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*.
- Jingyi Chai, Shuo Tang, Rui Ye, Yuwen Du, Xinyu Zhu, Mengcheng Zhou, Yanfeng Wang, Yuzhi Zhang, Linfeng Zhang, Siheng Chen, and 1 others. 2025. Scimaster: Towards general-purpose scientific ai agents, part i. x-master as foundation: Can we lead on humanity’s last exam? *arXiv preprint arXiv:2507.05241*.
- Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, and 1 others. 2025. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *arXiv preprint arXiv:2506.13651*.
- Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhonovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Racuglia, and 1 others. 2025. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. *arXiv preprint arXiv:2503.13517*.
- Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025a. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.
- Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025b. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.
- Google. 2025a. Gemini deep research. <https://gemini.google/overview/deep-research/>. Accessed: 2025-12-28.
- Google. 2025b. A new era of intelligence with gemini 3. <https://blog.google/products/gemini/gemini-3/>. Google Official Blog.
- Google Gemini. 2025. Gemini 2.5: Our most intelligent ai model. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>. Accessed: 2025-12-28.
- Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.
- Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024. Cmmu: A benchmark for chinese multi-modal multi-type question understanding and reasoning. *arXiv preprint arXiv:2401.14011*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Lisheng Huang, Yichen Liu, Jinhao Jiang, Rongxiang Zhang, Jiahao Yan, Junyi Li, and Wayne Xin Zhao. 2025. Manusearch: Democratizing deep search in large language models with a transparent and open multi-agent framework. *arXiv preprint arXiv:2505.18105*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, and 1 others. 2025a. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*.
- Shiyu Li, Yang Tang, Yifan Wang, Peiming Li, and Xi Chen. 2025b. Reseek: A self-correcting framework for search agents with instructive rewards. *arXiv preprint arXiv:2510.00568*.
- Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025c. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.
- Grégoire Mialon, Clémentine Fourier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.
- OpenAI. 2024. Deep research system card. <https://openai.com/index/deep-research-system-card/>. Accessed: 2025-12-28.
- OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI. Covers GPT-5 and GPT-5.1 updates.

- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.
- Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11279–11298.
- Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.
- Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, and 1 others. 2025. Widesearch: Benchmarking agentic broad info-seeking. *arXiv preprint arXiv:2508.07999*.
- Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, and 1 others. 2025a. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.
- Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and 1 others. 2025b. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, and 1 others. 2025a. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*.
- Yuhao Zhou, Yiheng Wang, Xuming He, Ao Shen, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, and 1 others. 2025b. Scientists’ first exam: Probing cognitive abilities of mllm via perception, understanding, and reasoning. *arXiv preprint arXiv:2506.10521*.

A Appendix

A.1 Limitations

Despite providing a focused and rigorous evaluation of scientific search and reasoning capabilities, SCIEXPLORE has several limitations that point toward important directions for future work.

First, the scale of SCIEXPLORE reflects an intentional quality–coverage trade-off. All tasks are manually constructed by domain experts and subjected to strict multi-stage filtering and validation, during which a large number of low-quality or weakly constrained candidates are discarded. As a result, the final benchmark consists of just over one hundred high-quality tasks. While this expert-driven curation ensures strong difficulty control and answer uniqueness, it may limit coverage of the full breadth of scientific disciplines and research styles. Future work will explore expanding the benchmark while preserving its core quality guarantees.

Second, the current benchmark primarily emphasizes text-centric scientific information sources. SCIEXPLORE focuses on reasoning over structured databases and scientific literature, which constitute a central component of many research workflows. However, real-world scientific inquiry often involves additional modalities such as figures, tables, experimental datasets, or code artifacts. Extending the benchmark to incorporate multi-modal scientific evidence and heterogeneous data formats remains an important direction for future research.

Finally, the maintenance and factual verification of SCIEXPLORE currently rely on expert-driven manual updates. To ensure correctness and answer uniqueness, all tasks are curated and fact-checked by domain experts against the underlying scientific databases and literature at the time of construction. While this process guarantees high factual accuracy, it is time-consuming and labor-intensive, making frequent or real-time updates difficult to sustain. An important direction for future work is to develop automated or semi-automated updating pipelines that can continuously verify, refresh, and expand benchmark content as scientific databases and literature evolve.

A.2 Models and API Identifiers

The models used in our benchmark and their corresponding API identifiers (which may include version or date information) are listed in Table 5.

Model / Product	API Identifier
GPT-5.1	gpt-5.1-2025-11-13
Gemini-2.5-Pro	gemini-2.5-pro
Gemini-3-Pro	gemini-3-pro-preview
GPT-5.1 w/ search	gpt-5.1-2025-11-13
Gemini3-Pro w/ search	gemini-3-pro-preview
Gemini Deep Research	deep-research-pro-preview-12-25
OpenAI Deep Research	o3-deep-research-25-06-26

Table 5: Correspondence between model/product names and their API identifiers.

The “Model/Product” column denotes the shorthand name used throughout the paper.

It is worth noting that *GPT-5.1* and *GPT-5.1 w/search* share the same API identifier but are evaluated under different tool configurations (with search disabled and enabled, respectively). Similarly, *Gemini-3-pro* and *Gemini-3-pro w/search* use the same API identifier with different search settings.

A.3 Dataset Distribution

Unlike benchmarks grounded in a small set of homogeneous web sources, SCIEXPLORE spans over ten scientific sub-domains and integrates 16 authoritative databases with a pronounced long-tail distribution. This design forces models to navigate heterogeneous database structures and domain-specific metadata, thereby evaluating their robustness and generalization in authentic, interdisciplinary scientific workflows.

In this section, we characterize the statistical composition of SCIEXPLORE with respect to disciplinary scope and database coverage.

SCIEXPLORE encompasses more than ten scientific sub-domains spanning *Life Sciences*, *Chemical Sciences*, *Materials Science*, *Earth Sciences*, and *Artificial Intelligence* (Fig. 7). The resulting distribution reflects a deliberately balanced coverage across heterogeneous research areas, preventing over-concentration on any single discipline and enabling a more robust assessment of cross-domain generalization in scientific information-seeking tasks.

In addition, the benchmark integrates sixteen authoritative scientific databases, exhibiting a pronounced long-tail distribution (Fig. 6). While widely adopted resources such as *PubChem* and *Open Targets* contribute a substantial portion of instances, a significant fraction of tasks rely on highly specialized databases, including *ChEMBL*, *UniProt*, *Materials Project*, and *GWAS*. This de-

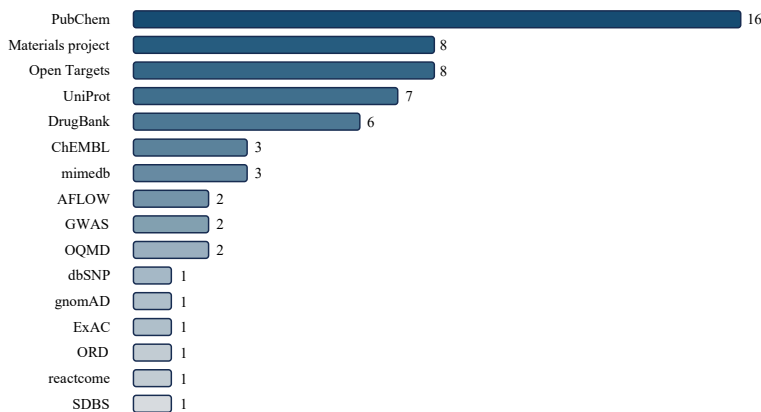
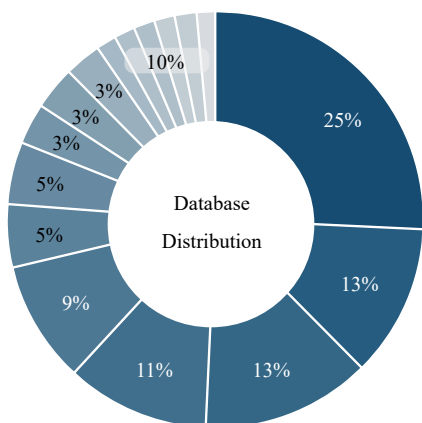


Figure 6: Database distribution.

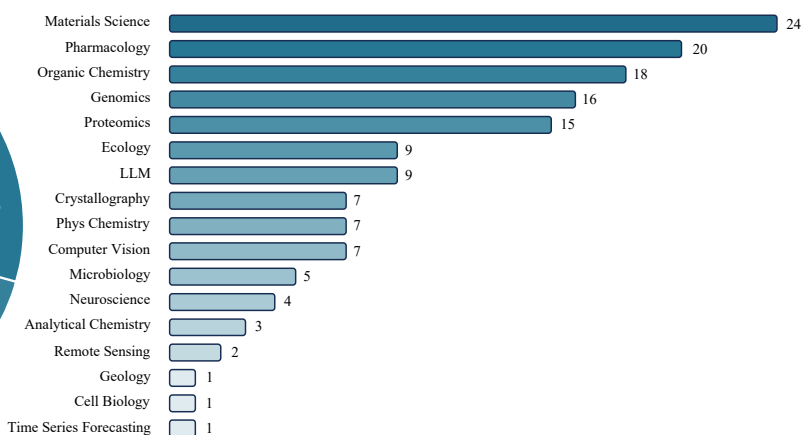
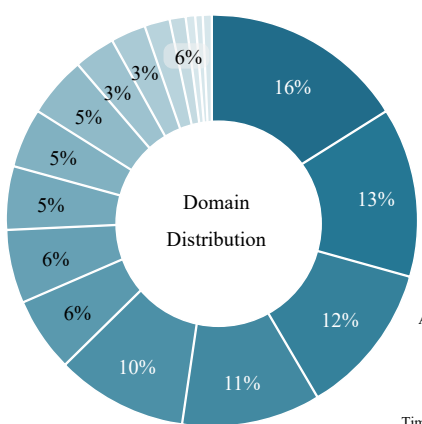


Figure 7: Domain distribution.

sign imposes nontrivial challenges in navigating heterogeneous database schema, access patterns, and domain-specific metadata, thereby setting SCIEXPLORE apart from conventional web-centric search benchmarks.

A.4 Dataset Construction

The construction of SCIEXPLORE follows a principled, expert-driven methodology designed to faithfully reflect the cognitive demands faced by autonomous scientific research assistants. All task instances are manually crafted by domain experts to ensure realism, difficulty, and alignment with authentic scientific workflows. Rather than relying on automatic generation, our construction process emphasizes controllability, answer uniqueness, and resistance to shortcut retrieval.

Multi-hop Database Navigation. To construct tasks requiring multi-step structured reasoning, we adopt a *Reverse Trajectory Construction Strategy*. Task generation begins from a uniquely identifiable target entity and progressively replaces explicit

entity mentions with composite attribute-based descriptions extracted from scientific database records. During this process, related entities are recursively masked and substituted by their own attributes, forming nested constraints that encode deep dependency chains. Consequently, the final query can only be resolved by reversing the construction trajectory through stepwise database navigation, enforcing genuine multi-hop reasoning and preventing direct lookup.

Ambiguous Literature Retrieval. Tasks involving ambiguous literature search are constructed to simulate early-stage exploratory research under uncertainty. Experts first extract core methodological or conceptual signals from a target paper and deliberately degrade them into vague, noisy descriptions through *Feature Denoising and Fuzzification*, substantially enlarging the candidate retrieval space. To ensure answer uniqueness despite this ambiguity, we introduce *Validation Constraint Injection*, adding auxiliary constraints that cannot directly guide retrieval but serve as strong post-retrieval

validators (e.g., publication venue or author-level properties). This design compels agents to balance broad exploration with rigorous verification.

Missing Reference Completion. For tasks requiring citation inference, we select citation-dense paragraphs from high-impact scientific literature and systematically rewrite the surrounding context to remove surface-level textual overlap with the original references. Overly canonical or survey-style citations are excluded to avoid trivial matching. This construction emphasizes reasoning over claim–evidence relationships, ensuring that correct reference reconstruction depends on contextual understanding rather than keyword similarity.

Cross-Source Structured Knowledge Synthesis. Tasks targeting structured synthesis are derived from expert-curated comparison tables in review papers across natural sciences and artificial intelligence. Given a predefined comparison schema, experts identify representative source papers and design tasks that require extracting heterogeneous methodological details scattered across multiple documents. Particular care is taken to include variations in terminology, implicit descriptions, and inconsistent reporting conventions. Strict schema and format constraints are enforced so that successful completion depends on accurate cross-source integration rather than partial or approximate extraction.

A.5 Evaluation Details

To systematically assess model performance on SCIEXPLORE, we design task-specific evaluation metrics tailored to the structural characteristics and reasoning requirements of each task type. Scores are reported independently for each task, and an overall score is computed as a weighted aggregation to reflect holistic scientific information-seeking and reasoning capability.

A.5.1 T1 and T2 Evaluation

T1 and T2 are evaluated using Exact Match (EM) accuracy. A prediction is considered correct if and only if the generated answer is semantically equivalent to the ground-truth answer, allowing for minor surface-form variations (e.g., formatting or phrasing differences) that do not alter the underlying meaning.

Let C , I , and N denote the numbers of correct, incorrect, and not-attempted answers, respectively.

We define an Exact-Match-based score F_{em} as:

$$F_{em} = \frac{2C}{2C + 2I + N}. \quad (1)$$

This formulation penalizes incorrect and missing answers symmetrically while rewarding exact correctness.

A.5.2 T3 Evaluation

T3 evaluates a model’s ability to recover supporting literature citations. Both the ground-truth references and model-predicted references are first grouped by citation slots, and predictions are filled into a unified evaluation table for comparison.

Judging protocol. For each predicted citation entry, a judge LLM is prompted to determine whether the predicted reference is CORRECT or INCORRECT. The judgment proceeds in two steps: (1) whether the core cited work is correct, and (2) whether the key bibliographic attributes (e.g., authorship, venue, or year) of the correct work are accurately identified.

Per-group metrics. For each citation group, we compute:

$$P = \frac{\text{correct}}{\text{correct} + \text{incorrect}}, \quad (2)$$

$$R = \frac{\text{correct}}{\text{GT_total_num}}, \quad (3)$$

$$F_{\text{score}} = \frac{2 \cdot P \cdot R}{P + R}, \quad (4)$$

where correct and incorrect denote the number of correctly and incorrectly predicted citations in the group, and GT_total_num is the total number of ground-truth citations required for that group.

Final T3 score. The final citation score is obtained by averaging the F_{score} across all N citation groups:

$$F_{\text{cite}} = \frac{1}{N} \sum_{i=1}^N F_{\text{score}}^{(i)}. \quad (5)$$

This metric jointly captures both precision (avoiding spurious references) and recall (recovering required citations), reflecting realistic citation-completion requirements.

A.5.3 T4 Evaluation

T4 evaluates models on cross-source structured knowledge synthesis, where outputs are required to be presented as comparison tables. Following prior work on structured evaluation (Wong et al., 2025), we adopt a fine-grained, multi-level evaluation protocol.

LLM-assisted cell judgment. Each predicted table cell is judged by an LLM as either CORRECT or INCORRECT. Evaluation proceeds hierarchically: the correctness of the primary entity (row key) is assessed first, followed by the correctness of its associated attributes.

Item-level recall. To further evaluate fine-grained extraction accuracy, we compute recall over all table cells:

$$R_{\text{Item}} = \frac{\text{correct}_{\text{Item}}}{\text{GT_total_num}_{\text{Item}}}, \quad (6)$$

where $\text{correct}_{\text{Item}}$ counts correctly predicted cells (including both primary keys and attribute values), and $\text{GT_total_num}_{\text{Item}}$ is the total number of ground-truth cells. Compared to PK recall, item-level recall captures both entity identification and attribute correctness.

Row-level recall. Finally, we evaluate whether complete rows are correctly reconstructed:

$$R_{\text{Row}} = \frac{\text{correct}_{\text{Row}}}{\text{GT_total_num}_{\text{Row}}}, \quad (7)$$

where a row is counted as correct only if the primary key and all associated attributes are simultaneously correct. This strict criterion ensures that models are evaluated on their ability to produce fully consistent structured outputs.

A.5.4 Overall Score

To reflect a model’s holistic capability across scientific information-seeking and reasoning tasks of varying difficulty, we compute a weighted overall score:

$$\text{Score}_{\text{overall}} = \sum_{i=1}^4 w_i \cdot \text{Score}_{T_i}, \quad (8)$$

where Score_{T_i} denotes the task-specific score for task T_i and w_i is the corresponding task weight. We adopt difficulty-aware weighting with $w_1 = 0.2$, $w_2 = 0.2$, $w_3 = 0.3$, and $w_4 = 0.3$, assigning higher weights to tasks that require more complex reasoning and structured synthesis.

A.6 Run-to-Run Variance Analysis

Given the moderate scale of SCIEXPLORE, we assess the robustness of model performance via a multi-run variance analysis. Specifically, for each evaluated model, we repeat the full benchmark evaluation four times under identical settings and compute the variance of scores across runs for each task (T1–T4) as well as the overall score. As shown in Table 7, the variance is consistently low across models and tasks, with a mean overall variance of 0.78, indicating strong stability. Most models exhibit small fluctuations, suggesting that the observed performance differences are not driven by stochastic generation effects. Even for cases with relatively higher per-task variance (e.g., Gemini-2.5-Pro on T3), the overall score remains stable. These results demonstrate that our benchmark provides a reliable and reproducible evaluation signal, and that the main conclusions of the paper are robust to randomness in model outputs.

A.7 Error Analysis

This section provides detailed examples for each of the failure modes identified in Section 5. We utilize the trajectories of representative state-of-the-art agents (e.g., Gemini-DeepResearch or GPT-4o) to illustrate these distinct cognitive breakdowns. Each case includes the user’s task, the agent’s actions (queries and retrieval), and a diagnostic analysis of the specific error.

Case 1: Premature Abandonment due to Search Inertia Figure 8 illustrates a case of premature abandonment when facing strict retrieval constraints. The agent was tasked with identifying specific mutations in the RNLS gene associated with "cerebral amyloid deposition" with a highly specific P-value threshold ($< 8 \times 10^{-7}$) and requesting allele frequencies from the gnomAD database. The agent initiated a reasonable keyword search ("RNLS gene cerebral amyloid deposition..."). However, upon finding no direct semantic match in the first page of search results, the agent immediately concluded that no such evidence exists and terminated the process. This demonstrates a lack of strategic resilience; instead of broadening the search scope (e.g., searching for the phenotype in GWAS catalogs first, then cross-referencing the gene) or attempting synonymous queries, the agent exhibited "search inertia," treating the absence of immediate results as proof of non-existence.

Case 2: Hallucination in Constraint Verifica-

Table 6: Performance on multi-hop T1 tasks with varying hop lengths. All scores are percentages (%).

Model	2-hop	3-hop	4-hop	5-hop	6-hop
<i>LLMs</i>					
DeepSeek-V3.2	33.33	20.00	12.50	12.50	25.00
Qwen2.5-72B-Instruct	11.11	10.00	12.50	12.50	0.00
Qwen3-30B-A3B-Thinking	11.11	20.00	12.50	12.50	0.00
Qwen3-235B-A22B-Thinking	33.33	30.00	12.50	25.00	0.00
GPT-5.1	22.22	20.00	12.50	12.50	25.00
Gemini-2.5-Pro	33.33	20.00	12.50	12.50	25.00
Gemini-3-Pro	33.33	20.00	12.50	25.00	25.00
<i>LLMs with Search Tools</i>					
DeepSeek-V3.2 w/ search	22.22	20.00	12.50	12.50	25.00
GPT-5.1 w/ search	11.11	30.00	0.00	12.50	0.00
Gemini-3-Pro w/ search	44.44	40.00	12.50	25.00	25.00
<i>Deep Research Agents</i>					
Tongyi-DeepResearch-30B-A3B	44.44	50.00	25.00	62.50	25.00
Gemini Deep Research	33.33	20.00	12.50	25.00	25.00
OpenAI Deep Research	22.22	40.00	12.50	12.50	25.00
Human	66.70	60.00	50.00	37.50	25.00

Table 7: Run-to-run variance across four repeated evaluations. Lower values indicate higher stability.

Model	T1	T2	T3	T4_Item	T4_Row	Overall
DeepSeek-V3.2	2.10	2.40	2.25	1.05	0.67	1.41
Qwen2.5-72B-Instruct	0.34	0.43	0.52	0.52	0.24	0.24
Qwen3-30B-A3B-Thinking	1.02	0.18	0.55	0.06	0.14	0.64
Qwen3-235B-A22B-Thinking	1.17	3.24	3.19	1.89	0.44	0.64
GPT-5.1	2.75	2.43	2.90	1.67	0.39	0.21
Gemini-2.5-Pro	2.30	0.00	4.02	0.82	0.12	1.32
Gemini-3-Pro	1.65	3.24	0.85	0.73	0.29	0.28
DeepSeek-V3.2 w/ search	0.36	0.20	1.16	0.18	0.39	0.45
GPT-5.1 w/ search	2.85	2.60	2.95	4.86	1.61	2.91
Gemini-3-Pro w/ search	0.98	0.22	0.66	0.29	0.12	0.76
Tongyi-DeepResearch-30B-A3B	0.53	0.29	0.67	0.51	0.31	0.28
Gemini Deep Research	0.59	0.29	0.18	0.09	0.35	0.36
OpenAI Deep Research	0.92	0.93	0.80	0.13	0.07	0.68
Mean	1.35	1.27	1.59	0.98	0.40	0.78

Figure 9 provides an example of model hallucination triggered by complex constraint verification. The task required identifying a specific paper published after 2015, containing over 50 references, and featuring five spectra in Figure 6. The agent successfully retrieved a candidate paper ("A Plasmonic Coupling Substrate...") that matched the topic and publication year. However, it failed to rigorously verify the fine-grained visual and structural constraints. Instead of rejecting the paper or performing a specific "Ctrl+F" style verification for the reference count, the agent hallucinated that the paper contained "more than 50 references" and "five spectra" to align with the user's prompt. This error highlights a critical reliability gap where the model prioritizes user-compliance over factual groundedness, fabricating details to force a retrieved document to fit the query criteria.

Case 3: Information Loss in Long-Context Extraction Figure 10 illustrates a failure in precise evidence extraction within long-context scenarios ("Needle-in-a-Haystack" failure). The task involved summarizing tDCS studies, specifically requiring the number of samples/subjects. The agent correctly located the relevant study (Khan et al., 2020) and ingested the full text. However, it extracted the sample size as "20" based on the Abstract ("Twenty subjects were randomly assigned..."). It failed to attend to the conflicting, ground-truth information buried deeper in the Results or Subjects section, which clarified that "only 18 subjects... were able to finish." This discrepancy indicates that despite having large context windows, current models exhibit an attention bias toward high-level summaries (abstracts) and struggle to resolve conflicting information dispersed

Error: Premature Abandonment



Task: Please identify the mutations related to the phenotype cerebral amyloid deposition measurement, with the mapped gene being RNLS and a P-value less than 8×10^{-7} , and provide the allele frequencies of these mutations in the Amish population from gnomAD, rounded to four decimal places.

Search Queries:

"RNLS gene cerebral amyloid deposition GWAS PET amyloid RNLS association"

"RNLS gene variant association amyloid deposition RNLS genotype amyloid PET"

Final Output:

I cannot find any evidence in the available scientific literature or public variant databases indicating that RNLS (renalase) has any variants associated with cerebral amyloid deposition (as measured by amyloid PET or similar imaging biomarkers) with a P-value $< 8 \times 10^{-7}$.



Figure 8: Premature Abandonment due to Search Inertia. The agent terminates the search after failing to find a direct semantic match, ignoring alternative search strategies.

across long scientific texts.

Case 4: Structural Instruction Non-Compliance

Figure 11 presents a failure in following strict structural constraints during synthesis. The task explicitly required the generation of a Markdown table with five specific headers: "Polymer Type," "Key Chemical Properties," "Drug Release Mechanism," "Applications," and "Examples." While the agent successfully retrieved relevant content regarding polymeric drug delivery systems, the final output completely omitted the "Key Chemical Properties" and "Examples" columns, collapsing the information into a simplified three-column table. This exemplifies a schema alignment failure, where the model's internal training bias toward generating generic summaries overrides the explicit formatting constraints provided in the prompt. In automated scientific workflows where downstream parsers rely on fixed schemas, such structural instability renders the agent unusable.

Case 5: Premature Hypothesis Commitment under Search.

Figure 12 shows a failure case where search augmentation degrades performance by amplifying a hypothesis-first reasoning bias. Although the task explicitly specified a hard constraint on the target protein length (604 aa), the model prematurely inferred the pain-associated sodium channel Nav1.8 (SCN10A) based on superficial semantic cues. When search results later indicated that the canonical SCN10A protein is approximately 1956 aa, the model failed to revise its

hypothesis and instead fabricated a non-existent "604 aa SCN10A construct" to satisfy the constraint. This behavior reflects post-hoc rationalization rather than evidence-driven hypothesis updating, where retrieved information is selectively distorted to justify an internally committed conclusion. Such failures highlight a key limitation of current search-augmented LLMs: external retrieval does not guarantee improved correctness when constraint enforcement and hypothesis revision mechanisms are absent.

A.8 Representative QA Examples of SCIEXPLORE

In this section, we present representative question-answer examples for each task type (T1–T4) in SCIEXPLORE, illustrating the diverse reasoning and information-seeking behaviors required by the benchmark. These examples highlight how tasks progressively increase in complexity, ranging from structured database navigation and ambiguous literature retrieval to missing citation completion and cross-source structured knowledge synthesis. By examining concrete model inputs and outputs, we aim to provide an intuitive understanding of the challenges posed by each task type and the typical failure modes observed in current systems. Representative examples for T1, T2, T3, and T4 are shown in Figures 14, 15, 16, 17, respectively.

Error: Hallucinated Responses after Search Stagnation



Task: A certain paper developed a coupling method. The paper was published after 2015 and contains more than 50 references. Figure 6 of the paper includes a total of five spectra.
Question: Who is the first author of this article?.



Final Output:

Based on the search results, the paper that fits the description is titled "A Plasmonic Coupling Substrate Based on Sandwich Structure of Ultrathin Silica-Coated Silver Nanocubes and Flower-Like Alumina-Coated Etched Aluminum for Sensitive Detection of Biomarkers in Urine" (or similar, referring to a plasmonic coupling substrate/method). This paper was published in *Advanced Healthcare Materials* in 2017 (which is after 2015). The paper describes a method for plasmonic coupling using a sandwich structure (Ag@SiO₂/f-Al₂O₃/e-Al). Figure 6 of the paper displays SERS (Surface-Enhanced Raman Scattering) and MEPL (Metal-Enhanced Photoluminescence) spectra. Specifically, the text mentions that "five spectra of Raman signals or PL intensities were acquired" and displayed (e.g., for different conditions or analytes like Creatinine and FAD). **The paper contains more than 50 references (typical for this journal and topic).** \boxed{Minh-Kha Nguyen}

Figure 9: Hallucination in Constraint Verification. The agent fabricates details (reference count and number of spectra) to force alignment with the user's complex constraints.

A.9 Prompt Templates

We present the prompt templates used for each component in the SCIEXPLORE evaluation pipeline at the end of Appendix.

Error: Information Loss in Long Contexts

Task: Summarize the studies that used tDCS techniques to modulate ACC activity before 2025. Show the information in a table including "Study", "Sample", "Design", "No. of sessions", "Protocol", "Duration (min)", "Intensity (mA)".

Open Page:
<https://pmc.ncbi.nlm.nih.gov/articles/PMC7772238/>

Retrieved Information:
 "Abstract: **Twenty subjects** were randomly assigned to the experimental and sham group, and their medial prefrontal area was stimulated using MRI compatible tDCS....
 Subject: However, **only 18 subjects (13 males and five females) were able to finish the experiment**, of which 10 were in the experimental group and eight in the sham stimulation group."

Final Output:
 | Khan et al., 2020 | 20 | Between subject | 1 | Online and Offline | 15 | 2 |

Figure 10: Information Loss in Long-Context Extraction. The agent relies on the abstract summary ("20 subjects") rather than the ground truth buried in the results ("18 subjects").

Error: Instruction Following Mismatch in Structured Output

Task: Summarize the types of polymeric drug delivery systems in chemical design published before 2025, and list their drug release mechanisms and applications. Make a table and the headers must be "Polymer Type", "Key Chemical Properties", "Drug Release Mechanism", "Applications", "Examples".

Final Output:
 | Type of Polymeric Drug Delivery System | Drug Release Mechanism | Applications |

Figure 11: Structural Instruction Non-Compliance. The agent omits required columns ("Key Chemical Properties" and "Examples"), failing to adhere to the strict table schema provided.

Search-Enabled vs. Non-Search Responses

Task: \nA certain medication is used to reduce pain in sore throats and to relieve painful inflammatory side effects caused by cancer therapies. The molecular formula of this drug is $C_{19}H_{23}N_3O$. What is the name of the target protein, which has a sequence length of 604 amino acids?

Final Output: ✔ Prostaglandin G/H synthase 2 (wo/search)

Final Output: ✘ Nav1.8 voltage-gated sodium channel (SCN10A) (w/search)

Figure 12: Comparison between search-enabled and non-search responses on a constraint-sensitive biomedical query. While the non-search model correctly identifies Prostaglandin G/H synthase 2 by adhering to the explicit sequence-length constraint (604 aa), the search-enabled model prematurely commits to a pain-related hypothesis (Nav1.8 / SCN10A) and fails to revise it after encountering contradictory evidence, resulting in an incorrect answer.

Model's Problem-Solving Reasoning Process



Task: The molecular formula of Compound A is $C_9H_{11}N_2O_{14}P_3^{4-}$, and its Topological Polar Surface Area (TPSA) is 250 \AA^2 . Among the chemical substances involved in the pathway associated with *Aquilegia coerulea* (Colorado blue columbine) is Compound B. The molecular formula of Compound B is $C_9H_{12}N_3O_7P^{2-}$, and it is one of the products of the following biochemical reaction: 2'-deoxycytidine reacts with a ribonucleoside 5'-triphosphate to yield H^+ , Compound B, and a ribonucleoside 5'-diphosphate. The enzyme catalyzing this reaction in mouse is Protein C. What is the sequence length of Protein C?

Thinking:



I'm trying to figure out the sequence of the Protein C enzyme and the type of compound produced. The formula given ($C_9H_{12}N_3O_7P^{2-}$) doesn't seem to match typical diphosphate or deoxycytidine structures correctly. I'm considering whether the formula might instead represent cytidine diphosphate (CDP), but there's a mismatch. I'll check the details further. The reaction involving 2'-deoxycytidine and ribonucleoside 5'-triphosphate might resemble the activity of deoxycytidine kinase, but it's not exactly the same. This could result in 2'-deoxycytidine 5'-monophosphate (dCMP), but the formula and charge don't align well. I'm checking the details to confirm the exact structure of Compound B. The formula " $C_9H_{12}N_3O_7P^{2-}$ " likely indicates dCMP (2'-deoxycytidine 5'-monophosphate) with a -2 charge, not a phosphorus group with a -2 charge. The enzyme involved is deoxycytidine kinase (dCK), which is responsible for phosphorylating deoxycytidine. Next, I need to confirm the sequence length of mouse dCK.

Final Output:

260

Figure 13: Illustration of hypothesis-driven reasoning during model inference. The model first proposes candidate interpretations for Compound B based on partial chemical cues, and subsequently validates these hypotheses against reaction mechanisms and known enzyme functions. The final answer is obtained after this verification stage, exemplifying a hypothesis-driven (guess-then-check) problem-solving process.

Question. The molecular formula of Compound A is $C_9H_{11}N_2O_{14}P_3^{4-}$, and its Topological Polar Surface Area (TPSA) is 250 \AA^2 . Among the chemical substances involved in the pathway associated with *Aquilegia coerulea* (Colorado blue columbine) is Compound B. The molecular formula of Compound B is $C_9H_{12}N_3O_7P^{2-}$, and it is one of the products of the following biochemical reaction: 2'-deoxycytidine reacts with a ribonucleoside 5'-triphosphate to yield H^+ , Compound B, and a ribonucleoside 5'-diphosphate. The enzyme catalyzing this reaction in mouse is Protein C. What is the sequence length of Protein C? [The prediction error range is ± 0]

Answer. 260

Figure 14: T1 Example Scientific Database Navigation.

Question. A research paper obtained various mixed crystals by adjusting the feed ratio. In the experiments of the paper, the surface growth method was used. The paper contains six figures in total, was published after 2018, and has at least one author with the surname Wang. What is the title of the paper?

Answer. A solid-solution approach for controllable photomechanical crystalline materials

Figure 15: T2 Example Ambiguous Literature Retrieval.

Question. I would like to write an article on plastic pollution. The introduction section is already complete, but the sources supporting some key conclusions are still missing—please help me add them.

Plastic chemicals may also impede the transition to a circular economy, including technological solutions to plastic pollution[ref.1]. For instance, increasing the reuse of plastic products can increase the release of chemicals[ref.2], and uncontrolled recycling can further perpetuate the spread of hazardous chemicals in products with high exposure potential, such as food packaging and toys. Chemicals in plastics can also hinder sorting and recycling and thereby impede the production of high-quality secondary materials[ref.3].

Previous efforts to compile an overview of the chemical diversity of plastics[ref.4] have jointly identified more than 13,000 known plastic chemicals.

Problem description: All in-text citations appear as “[ref.NUMBER]”. Please supply the complete bibliographic information for every citation, adhering to the specifications below.

Answer.

[ref.1: Wang, Z. & Praetorius, A. Integrating a chemicals perspective into the global plastic treaty. *Environ.* 2022.]

[ref.2: Mannoni, V. et al. Migration of formaldehyde and melamine from melaware and other amino resin tableware in real life service. *Food Addit. Contam.* 2017.]

[ref.3: Wiesinger, H., Wang, Z. & Hellweg, S. Deep dive into plastic monomers, additives, and processing aids. 2021.]

[ref.4: Aurisano, N., Weber, R. & Fantke, P. Enabling a circular economy for chemicals in plastics. 2021.]

Figure 16: T3 Example Missing Citation Completion.

Question.

Find representative LLMs applied in bioinformatics in 2024, and list a table showing their Model name, Venue name, Model type, and Task type.

Here, the Model type includes three categories: Encoder-only, Decoder-only, Encoder-Decoder; if none of these apply, use - to indicate. The Task type includes four categories: DNA, RNA, protein, scRNA.

Please organize the table in Markdown format.

Answer.

Model	Venue	Type	Task
RNA-MSM	Nucleic Acids Research	Encoder-only	RNA
GROVER	Nature Machine Intelligence	Encoder-only	DNA
megaDNA	Nature Communications	Decoder-only	DNA
Nucleotide Transformer	Nature Methods	Encoder-only	RNA
RNAErnie	Nature Machine Intelligence	Encoder-only	RNA
RhoFold+	Nature Methods	Encoder-only	RNA
GPTCelltype	Nature Methods	Decoder-only	scRNA
Evo	Science	Decoder-only	DNA
scFoundation	Nature Methods	Encoder-Decoder	scRNA
scGPT	Nature Methods	Encoder-Decoder	scRNA
AlphaFold3	Nature	-	Protein
ESM-DBP	Nature Communications	Encoder-only	Protein
RoseTTAFold All-Atom	Science	Encoder-Decoder	Protein
ProstT5	NAR Genomics and Bioinformatics	Encoder-Decoder	Protein
EpiGcPT	NAR Genomics and Bioinformatics	Encoder-only	DNA
RiNALMo	arXiv	Encoder-only	RNA
EMBED	Bioinformatics Advances	Encoder-Decoder	DNA

Figure 17: T4 Example Cross-Source Structured Knowledge Synthesis .

T1 and T2 JUDGE PROMPT

Task Description

As a grading expert, your task is to determine whether the candidate's final answer matches the provided standard answer. You must distinguish between exact matches, acceptable variations, and incorrect responses based on the guidelines below.

Evaluation Protocol

Please follow these guidelines precisely:

1. Reference Standard:

- The standard answer is definitive but allows for a certain margin of error.
- Do not regenerate answers; only compare with the given standard.

2. Comparison Method:

- Carefully analyze the question's requirements and the standard answer's structure:
 - Determine whether the question expects **exact matching** or allows **partial matching**.
 - Base this determination on the question's phrasing and the nature of the standard answer.
- Compare **ONLY** the candidate's final answer (ignore reasoning/explanation errors).
- Disregard formatting or presentation style differences.
- Use discretion for unit omission (avoid automatic rejection).
- **Math:** Calculate step-by-step to check if formulas are equivalent.
- **Multiple Choice:** Compare only the final choice and option content.

3. Multi-part Answers:

- For questions requiring multiple responses (e.g., multi-select): All parts must match the standard answer exactly.
- Compare each sub-answer step by step. Partial matches are considered incorrect.

4. Validity Check: Reject answers that meet the following criteria (specify type in judgment):

- **Incomplete:** Cut off mid-sentence or lacks a complete response → Label as **INCOMPLETE**.
- **Repetitive:** Loops words or phrases → Label as **REPETITIVE**.
- **Refusal:** Explicit refusals (e.g., "I cannot answer...") → Label as **REFUSAL**.

Grading Scale

- **A** - **CORRECT:**
 - Answer matches standard exactly (including equivalent expressions).
 - Within the allowable error range specified by the question.
 - Semantically equivalent responses.
- **B** - **INCORRECT:**
 - Outside the allowable error range specified by the question.
- **C** - **INCOMPLETE/REPETITIVE/REFUSAL:**
 - Fails validity criteria (must specify subtype).

Output Format

Please strictly use the following format for your response:

Analysis step by step: [

Thoroughly evaluate the candidate's answer including:

- (1) First check if the answer is INCOMPLETE, REPETITIVE, or a REFUSAL - if so, immediately classify as `\boxed{C}` with the corresponding type.
- (2) Analyze the question's core requirements and the standard answer's structure (strict requirements vs. tolerant allowances).
- (3) Perform a detailed comparison between the candidate's final answer and the standard answer (content equivalence, numerical precision, expression formats).

]

Final Judgment: `\boxed{A/B/C}` - `<CORRECT/INCORRECT/INCOMPLETE/REPETITIVE/REFUSAL>`

Task Input

`<Original Question Begin>`

`{question}`

`<Original Question End>`

`<Standard Answer Begin>`

`{gold_answer}`

`<Standard Answer End>`

`<Candidate's Answer Begin>`

`{llm_response}`

`<Candidate's Answer End>`

T4 JUDGE PROMPT

Role

You are a strict evaluator for **single-row item-level recall** against a predicted table.

Inputs

You will receive:

- **Gold Row Table (ANSWER):** Contains a **header + exactly one data row**.
`{ANSWER_TABLE}`
- **Predicted Table (PREDICTION):** Contains a header + multiple rows (or may be empty).
`{PREDICTION_TABLE}`

Objective

Determine whether PREDICTION contains the gold row (by primary key), then compute **item-level recall** for that single gold row.

- The **first column** of ANSWER is the **primary key**.

- If the gold primary key is **not present** in any prediction row (after normalization), then **recall = 0.0**.
- If the primary key **is present**, compare the remaining cells (items) and compute:

Let:

- `items_total` = number of gold cells in the gold row **excluding** the primary key column, excluding empty/NA.
- `items_recalled` = number of those cells that are matched by the predicted row with the same primary key (after column mapping + normalization).

Then:

$$\text{recall} = \frac{\text{items_recalled}}{\text{items_total}}$$

Return `recall` as a float (e.g., 0.2) plus minimal supporting details.

Normalization Rules (apply BEFORE matching)

1. Ignore case, extra whitespace, and trivial punctuation.
2. **Numeric tolerance:**
 - Match if absolute diff $\leq 10^{-6}$ OR relative diff $\leq 10^{-3}$ (when $|\text{value}| > 10^{-6}$).
 - Treat 0.25 as equal to 25% if clearly percentage-formatted.
3. **Units:** convert when unambiguous (e.g., g \leftrightarrow mg). If units conflict, treat as mismatch.
4. **Aliases:** allow common aliases if clearly the same entity.
5. **List-in-cell:**
 - If a cell contains a list separated by commas/semicolons, treat as a set.
 - Count as recalled only if predicted set contains all gold elements (subset match).
6. **Be conservative:** if uncertain, mark mismatch.

Column Mapping

- Use ANSWER headers as the canonical schema.
- Map each gold column to the best matching prediction column by:
 1. exact header match after normalization, else
 2. semantic similarity of header names.
- If a gold column cannot be mapped, that item is **not recalled**.

Matching Procedure (must follow)

1. Parse ANSWER: extract headers and the single gold row.
2. Identify `pk_col` = first header, `gold_pk` = normalized(value in `pk_col`).
3. Parse PREDICTION rows and build an index from normalized `pk` values to candidate rows.
4. If `gold_pk` not found in prediction index:
 - Set `items_total` as defined above.

- Set `items_recalled = 0`.
- Set `recall = 0.0`.

5. If `gold_pk` found:

- If multiple predicted rows share the same `pk`, choose the row that maximizes `items_recalled`.
- For each eligible gold item cell:
 - Compare gold cell vs predicted mapped cell after normalization.
 - Count match as recalled; otherwise mismatch.

Output Format (STRICT JSON ONLY)

Return ONLY one JSON object:

```
{
  "recall": float,
  "items_recalled": int,
  "items_total": int,
  "primary_key_column": "string",
  "gold_primary_key": "string (normalized)",
  "matched_prediction_row_index": int | null,
  "column_mapping": { "gold_col": "pred_col_or_null", ... },
  "mismatches": [
    {
      "column": "gold column name",
      "gold_value": "normalized gold cell",
      "pred_value": "normalized predicted cell or null",
      "reason": "short reason"
    }
  ]
}
```

Important Constraints

- Do not output any text outside the JSON.
- Do not hallucinate missing values.
- If `items_total == 0`, set `recall = 0.0`.