SciRAG: Adaptive, Citation-Aware, and Outline-Guided Retrieval and Synthesis for Scientific Literature

Anonymous ACL submission

Abstract

The accelerating growth of scientific publications has intensified the need for scalable, trustworthy systems to synthesize knowledge across diverse literature. While recent retrieval-augmented generation (RAG) methods have improved access to scientific information, they often overlook citation graph structure, adapt poorly to complex queries, and yield fragmented, hard-to-verify syntheses. We introduce SciRAG, an open-source framework for scientific literature exploration that addresses these gaps through three key innovations: (1) adaptive retrieval that flexibly alternates between sequential and parallel evidence gathering; (2) citation-aware symbolic reasoning that leverages citation graphs to organize and filter supporting documents; and (3) outline-guided synthesis that plans, critiques, and refines answers to ensure coherence and transparent attribution. Extensive experiments across multiple benchmarks such as QASA and ScholarQA demonstrate that SciRAG outperforms prior systems in factual accuracy and synthesis quality, establishing a new foundation for reliable, large-scale scientific knowledge aggregation.

1 Introduction

With over four million journal articles published in 2024, scholarly output continues its decade-long 8% annual growth (Crossref, 2024). The rise of preprint servers and open-access repositories has expanded scientific discourse, fostering cross-disciplinary discovery (Bornmann and Mutz, 2015), but also burdening researchers with reconciling fragmented findings, outpacing manual surveys and bibliometric tools (Beltagy et al., 2019; Asai et al., 2024; Singh et al., 2024).

Retrieval-Augmented Generation (RAG) has advanced rapidly since its introduction as a framework for knowledge-intensive NLP (Lewis et al., 2020). Recent systems couple LLMs with external search, significantly improving performance on knowledge-intensive benchmarks (Asai et al., 2024; Zheng et al., 2024; Skarlinski et al., 2024).

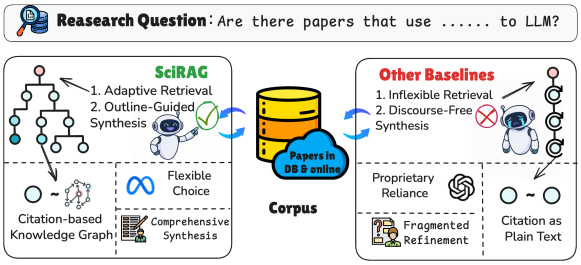


Figure 1: An overview of SciRAG framework.

When applied to scientific literature, however, current RAG systems still exhibit four key limitations: (1) **Superficial exploitation of citations**: references are treated as plain unstructured text rather than as structured relational entities, or, at best, single-hop backlinks, leaving the richer forward–backward citation graph unused (Zhang et al., 2024; Agarwal et al., 2024; Bornmann and Daniel, 2008). (2) **Inflexibility of retrieval**: queries are typically issued in a fixed, one-pass manner without adapting depth or coordinating orthogonal subtopics (Asai et al., 2024; Zheng et al., 2024). This limitation is especially pronounced for scientific literature, where complex multi-aspect questions (*e.g.,* combining theory, methodology, and application) demand dynamic and context-aware retrieval. (3) **Discourse-free synthesis**: models concatenate passage snippets without a global rhetorical plan, yielding answers that drift, overlook caveats, or conflate conflicting evidence (Skarlinski et al., 2024). (4) **Proprietary and costly frameworks**: many frameworks are proprietary and expensive, withholding models, indices, and workflows, which impedes reproducibility and further research.

To address these gaps, we propose **SciRAG**[1], a retrieval and synthesis framework for scientific lit-

[1]Our code is available at https://anonymous.4open.science/r/SciRAG-5C43, and will be released publicly upon publication.

erature. SciRAG pioneers an adaptive architecture that dynamically integrates citation-driven reasoning over the literature graph, symbolic logic, and structured knowledge aggregation. To address the unique challenges of scientific literature QA, such as complex query structure, implicit cross-paper reasoning, and fragmented evidence, SciRAG is built around three tightly integrated components:

1. **Adaptive Retrieval.** A query-aware controller dynamically switches between sequential exploration, which is ideal for answering complex and in-depth questions, and parallel retrieval, which independently handles multiple sub-questions. This design enables flexible and comprehensive evidence gathering.

2. **Citation-Aware Symbolic Reasoning.** Explicitly traverses forward and backward citation paths in literature graphs to uncover conceptual relationships among studies, structuring retrieved evidence into interpretable contribution chains that guide both candidate reranking and logically grounded answer generation.

3. **Outline-Guided Synthesis.** SciRAG proposes an outline-guided synthesis module that first generates a coarse answer outline to structure retrieval, then iteratively identifies factual gaps, retrieves missing evidence, and refines the draft into a coherent and well-supported final answer.

By integrating adaptive retrieval with citation centric symbolic reasoning, SciRAG establishes a new paradigm for trustworthy and scalable scientific knowledge synthesis. Extensive experiments on diverse open retrieval benchmarks, such as ScholarQA and PubMedQA, demonstrate that SciRAG consistently outperforms strong baselines including OpenScholar (Asai et al., 2024) and PaperQA2 (Skarlinski et al., 2024), achieving higher factual accuracy and overall relevance. Moreover, our analysis demonstrates the contributions of each component through ablation studies, verifies factual grounding via hallucination checks and case studies, and evaluates scalability under varying retrieval depths, confirming the framework's robustness and interpretability.

2 Related Work

LLMs for Scientific Research. LLMs are beginning to automate a broad spectrum of scientific workflows, from idea generation and hypothesis formation (Baek et al., 2025; Yang et al., 2024) to code-level experiment design (Huang et al., 2023; Tian et al., 2024) and literature-centric question answering (Asai et al., 2024; Zheng et al., 2024; Skarlinski et al., 2024). Correspondingly, community benchmarks have evolved from single-paper fact checking (Wadden et al., 2020) and abstractive QA (Lee et al., 2023) to multi-disciplinary, multi-paper synthesis suites such as ScholarQABench (Asai et al., 2024; Singh et al., 2025). These datasets highlight an emerging consensus: credible scientific assistance demands verifiable citations and wide coverage across disparate sub-fields. Such requirements are difficult for purely parametric LMs to satisfy, but we believe they can be effectively addressed by the integrated retrieval and reasoning design of SciRAG.

Graph-Enhanced RAG Systems. Several graph- and structure-aware RAG systems have been proposed to enhance open-domain QA by incorporating citation-graph propagation, section segmentation, or multi-hop retrieval. These systems include LitFM (Zhang et al., 2024), LitLLM (Agarwal et al., 2024), EfficientRAG (Zhuang et al., 2024), CoRAG (Wang et al., 2025), DeepRAG (Guan et al., 2025), and CG-RAG (Hu et al., 2025). While these systems show promise on small-scale, static datasets for tasks such as title generation or local citation prediction, they often lack discourse-level reasoning and struggle with query-adaptive exploration over large open corpora. Relying on shallow, sequential chains, they fail to handle multi-hop knowledge trails, resulting in poor scalability and limited recall in open-domain synthesis across millions of documents, a core requirement in scientific literature synthesis.

RAG Systems for Scientific Literature Tasks. To address citation fidelity, recent RAG systems combine LLMs with external corpora. OpenScholar (Asai et al., 2024) implements explicit citation verification and self-feedback; OpenResearcher (Zheng et al., 2024) merges dense-sparse retrieval with adaptive query rewriting; and PaperQA2 (Skarlinski et al., 2024) frames the task as a search-refine loop. However, these systems rely on sequential retrieval processes, which can overlook indirect evidence or lead to excessively large context windows when a query spans multiple theoretical threads. SciRAG, on the other hand, orchestrates parallel, citation-graph-aware retrieval, balancing high recall with precision and scaling efficiently to millions of papers. This approach enables SciRAG to

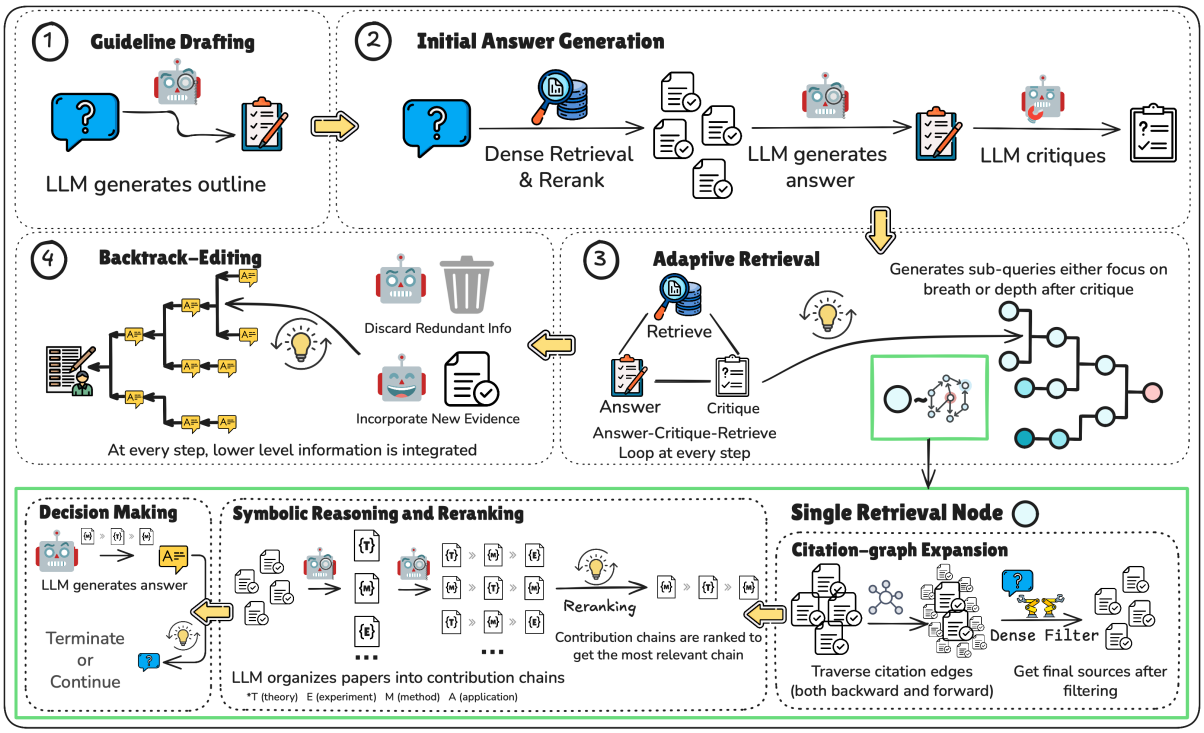


Figure 2: An illustration of the SciRAG pipeline. The process begins with guideline drafting and initial answer generation. Each retrieval node searches documents, decides whether to expand along the citation graph, builds contribution chains, and applies reasoning-based reranking to judge from current information whether to continue or stop. Adaptive retrieval integrates multiple nodes to balance sequential exploration for depth and parallel exploration for breadth. Finally, backtrack-editing consolidates all evidence and produces a coherent, well-documented answer.

address complex queries and multi-faceted tasks more effectively than its sequential counterparts.

3 SciRAG Framework

Scientific literature retrieval and synthesis present distinct challenges beyond those in general-domain QA. Queries are often multi-faceted, requiring the integration of theoretical context, methodological details, and application-specific findings. Relevant evidence is scattered across papers connected by implicit conceptual links and complex citation structures, making isolated retrieval or naive summarization insufficient.

To this end, SciRAG introduces a tightly integrated pipeline comprising three complementary components: (1) **Outline-guided synthesis**, which structures answer generation through planning and iterative refinement; (2) **Citation-aware symbolic reasoning**, which constructs and prunes contribution chains through forward and backward citation expansion and symbolic reasoning; and (3) **Adaptive retrieval**, which dynamically alternates between sequential and parallel search based on query structure. Together, these modules address key obstacles in scientific QA, including retrieval inflexibility, reasoning opacity, and synthesis incoherence, while enabling transparent and verifiable responses grounded in literature. An overview of the full system is shown in Figure 2, and all the prompt templates of our proposed system can be found in Appendix E.

3.1 Outline-Guided Answer Aggregation with Reflective Refinement

To overcome the fragmentation, inconsistency, and shallow structure of conventional retrieval-augmented methods, SciRAG employs an adaptive outline-guided aggregation procedure organized around a "plan–critic–solve" cycle. From the user's query, it first derives a detailed outline that serves as a scaffold to keep retrieval and synthesis aligned with the intended scope and depth. Guided by this outline, SciRAG then enters a reflective refinement phase that critiques preliminary answers, diagnoses logical gaps, and triggers targeted retrieval to collect the additional evidence needed. Corrections are deferred to a backtracking edit phase, preserving coherence and reinforcing factual support.

Finally, SciRAG performs a bottom-up edit-

ing process: beginning from the deepest retrieval branches, each parent layer sequentially integrates the synthesized outputs of its children to iteratively revise and strengthen the draft. At each stage, newly retrieved evidence is incorporated, redundant or conflicting citations are pruned, and precise provenance links are maintained, producing coherent and well-grounded final answers.

3.2 Citation-Graph Expansion with Symbolic Reasoning

Typical RAG pipelines rely primarily on embedding similarity and thus overlook the *structural* and *logical* relations central to scientific discovery. To address this limitation, **SciRAG** enhances retrieval through two coordinated stages: (1) *citation-graph expansion*, which broadens the search space under LLM guidance; and (2) *symbolic reasoning*, which constructs, filters and ranks evidence by conceptual role. By combining these two stages, SciRAG produces a transparent, role-aware evidence set that supports robust multi-paper synthesis, going far beyond what similarity-only or shallow multi-hop systems can provide. For clarity, we provide a simplified snapshot in Appendix D, illustrating how contribution chains are constructed and how reranking and filtering are applied.

Citation-graph expansion. Starting from an initial embedding match set P_0, an LLM first judges whether P_0 suffices to answer query q. If not, SciRAG traverses both *backward* and *forward* edges (≤ 1 hop) to assemble an enriched pool P. This step surfaces foundational works, key replications, and derivative applications that pure similarity search often misses, yielding a broader and more historically grounded candidate set for synthesis.

Symbolic reasoning and reranking. Each paper $p \in P$, represented by its abstract, is segmented and tagged into conceptual roles: T (theory), E (experiment), M (method), A (application), etc., ensuring fine-grained understanding of its content. The LLM then analyzes these tagged segments to uncover conceptual links (*e.g.,*, "[1]T→[2]E" when paper 1's theory supports paper 2's experiment), while pruning contradictory, tangential, or weakly supported branches. Pruning decisions reflect content relevance, query consistency, and logical coherence within the citation context. From the pruned graph, the model extracts *contribution chains*, coherent multi-paper inference paths that collectively address the query. Rather than relying on ad hoc similarity or centrality scores, the LLM performs in-context reasoning to compare chains, evaluating their logical coherence, evidential completeness, and query relevance, then ranks the originating papers accordingly. Papers anchored in the strongest chains are promoted to the answer generator, while those with fragmented or unsupported reasoning are discarded, with justifications recorded for transparency.

Algorithms for Expansion & Reasoning & Reranking

Require: query q, initial set P_0
if LLMJUDGE(q, P_0) **then**
  $P \leftarrow$ EXPANDGRAPH(P_0)
else
  $P \leftarrow P_0$
end if
for all $p \in P$ **do** ▷ segment & role tags
  seg$(p) \leftarrow$ TAG(p)
end for
$G \leftarrow$ BUILDRELATIONGRAPH(P)
$G \leftarrow$ PRUNECHAINS(G) ▷ remove contradictions
for all $p \in P$ **do**
  RANK$(p) \leftarrow$ LLMREASON(q, p, G)
end for
return TOPK(P, rank)

3.3 Adaptive Retrieval: Sequential and Parallel Mechanisms

Traditional retrieval strategies for scientific literature often struggle to balance coverage with depth, especially for multifaceted queries. SCIRAG addresses this limitation with an adaptive scheme that alternates between sequential and parallel retrieval, guided by the structural complexity and granularity of the user's information need. To further boost precision, it retrieves short text fragments rather than entire documents and maps them back to sources for citation expansion. Passage-level retrieval not only raises topical relevance but also strengthens the foundation for evidence fusion. The entire process is driven by an "answer–critique–retrieval" loop, where each iteration diagnoses information gaps and generates sub-queries. Instead of being explicitly told to choose sequential or parallel search, the LLM automatically issues both deepening and broadening sub-queries, each launching its own retrieval thread and naturally combining depth- and breadth-oriented exploration.

For queries requiring deep, context-dependent exploration, SciRAG builds retrieval rounds step by step: earlier results provide context that guides later searches, preserving logical continuity while progressively enriching the evidence. When a query decomposes into independent sub-questions, dis-

Dataset	*Task*	*Domain*	*Size*	*Metric*
SciFact	Claim Verification	Biomedicine	208	Corr & Cite
PubMedQA	Yes/No Judgement	Biomedicine	843	Corr & Cite
QASA	Q&A	Computer Science	1,375	Corr & Cite
ScholarQA-CS	Q&A	Computer Science	100	Corr & Cite
ScholarQA-BIO	Q&A	Biomedicine	1,451	Cite
ScholarQA-NEURO	Q&A	Neuroscience	1,308	Cite
ScholarQA-MULTI	Q&A	Multi-domain	108	Corr & Cite

Table 1: Overview of benchmarks evaluated in this study. Metrics: Corr=Correctness Score, Cite=Citation F1.

tinct retrieval threads operate concurrently, each targeting a specific sub-query (Appendix C), thus improving efficiency and ensuring balanced coverage of divergent facets. By integrating sequential and parallel exploration with snippet-level retrieval, SciRAG achieves comprehensive, precise, and coherently organized coverage of the scientific literature, outperforming the rigid, one-dimensional pipelines of prior systems.

4 Experiment Setup

4.1 Baseline Systems

To evaluate the effectiveness of SciRAG, we evaluate it against several strong baselines, briefly described below: (1) **SciRAG**: Our proposed advanced framework. When using vector retrieval, We utilize the OpenScholar Datastore (Asai et al., 2024) as our vector retrieval corpus, which contains over 45 million papers and more than 200 million snippets. Throughout the entire pipeline, we use the standard GPT-4o Legacy model for retrieval, reasoning, and answer generation. (2) **OpenScholar** (Asai et al., 2024): A large-scale scientific RAG system with an iterative self-feedback process to improve citation accuracy and content quality. It uses the same Datastore as SciRAG and provides four model versions (GPT-4o, Llama3.1-70B, OS-70B, OS-GPT4o), where OS-70B and OS-GPT4o are fine-tuned on scientific corpora. (3)**PaperQA2** (Skarlinski et al., 2024): A retrieval-augmented framework designed for literature synthesis. Implemented with its official open-source code and crawler, though our replication is limited by lack of access to private or license-protected papers. (4) **GPT-4o with Online Search** (OpenAI et al., 2024): GPT-4o augmented with real-time web search to enhance response accuracy. (5) **Perplexity Pro**[2] : A commercial RAG system combining LLMs with real-time web search to deliver conversational responses with citations.

[2] https://www.perplexity.ai/

4.2 Evaluation Benchmarks

For benchmark settings, we follow OpenScholar (Asai et al., 2024) and adopt the same four tasks: SciFact, PubMedQA, QASA, and ScholarQA, while explicitly adapting them to open-retrieval settings as detailed below. Together, these tasks cover diverse scientific domains and query types, enabling robust evaluation of answer synthesis.

SciFact (Wadden et al., 2020) is a biomedical claim verification benchmark, originally single-document. We adapt it to open retrieval by requiring evidence from a large corpus to verify claims as *supported* or *contradicted*, testing both fact-checking and retrieval.

PubMedQA (Jin et al., 2019) contains expert-written yes/no biomedical questions, originally paired with abstracts. We convert it into open retrieval, where models must locate literature to determine the binary answer.

QASA (Lee et al., 2023) contains reasoning-heavy questions from single AI/ML papers. We adapt it to open retrieval, where models must locate source documents before answering.

ScholarQA (Asai et al., 2024) evaluates multi-document synthesis for literature review questions across four domains: computer science (CS), biomedicine (BIO), neuroscience (NEURO), and a mixed set (MULTI). CS includes expert references and rubrics; BIO and NEURO provide curated questions demanding deep synthesis; MULTI offers long-form answers with citations for detailed assessment of coverage and citation quality.

Detailed dataset statistics are shown in Table 1.

4.3 Evaluation Protocols

Automated Evaluation. We adopt two core metrics. **Correctness Score** evaluates factual consistency and relevance. For SciFact and PubMedQA, we use Exact Match against expert-labeled binary

Method / Dataset	SciFact		PubMed		QASA		CS		MULTI		BIO	NEURO	Cost
	Corr	Cite	Corr	Cite	Corr	Cite	Corr	Cite	Corr	Cite	Cite	Cite	USD/query
Finetuned Baselines													
OpenScholar-OS-70B	82.1	47.5	79.6	74.0	**23.4**	64.2	52.5	45.9	4.03	54.7	55.9	63.1	0.01
OpenScholar-OS-GPT4o	81.3	56.5	74.8	77.1	18.7	60.4	57.7	39.5	4.51	37.5	51.5	43.5	0.12
Untuned / Legacy Baselines													
GPT-4o	77.8	0.0	65.8	0.0	21.2	0.0	45.0	0.1	4.01	0.7	0.2	0.1	0.06
OpenScholar-GPT4o	79.3	47.9	75.1	73.7	18.3	53.6	52.4	31.1	4.03	31.5	36.3	21.9	0.12
OpenScholar-Llama3.1-70B	78.2	42.5	77.4	71.1	22.7	63.6	48.5	24.5	4.24	41.4	53.8	58.1	0.00
PaperQA2†	–	–	–	–	–	–	45.6	48.0	3.82	47.2	56.7	56.0	0.3–2.3
Perplexity Pro†	–	–	–	–	–	–	40.0	–	4.15	–	–	–	0.002
Ours (Untuned)													
SciRAG(Llama3.1-70B)	81.5	44.1	78.2	71.7	21.5	63.8	60.2	28.4	4.51	37.1	43.1	45.9	0.00
SciRAG(GPT-4O)	**84.1**	52.9	**84.1**	74.8	19.2	54.2	**69.0**	34.0	**4.79**	37.8	44.8	36.2	0.16

Table 2: Performance comparison across multiple scientific QA datasets. Note that the evaluation scale for ScholarQA-MULTI ranges from 0 to 5, whereas the other benchmarks adopt a 0–100 scale.†: PaperQA2 is designed for multi-document synthesis and relies on a local PDF-based corpus, which is not publicly available. Perplexity Pro may incorporate non-scholarly sources and does not expose citation snippets, preventing full evaluation. As a result, we evaluate both baselines only on a subset of benchmarks.

answers. QASA is scored by ROUGE-L overlap with references. ScholarQA-CS uses expert rubrics specifying must-have and nice-to-have elements, while ScholarQA-MULTI is assessed via *Prometheus-8x7b-v2.0* (Kim et al., 2024), which rates relevance, coverage, and organization. **Citation F1** captures citation accuracy as the harmonic mean of precision (citation supports claim) and recall (all citation-worthy claims are cited). It is used for all datasets and serves as the sole metric for ScholarQA-BIO and NEURO. These metrics support a realistic and comprehensive evaluation of SciRAG in open-retrieval scientific QA.

Human Evaluation. We conducted a human evaluation with three expert annotators, each holding at least a Master's degree in CS. The evaluation compared answers generated by SCIRAG, OpenScholar, and PaperQA2 for 30 randomly sampled queries from the ScholarQA-CS dataset. Each annotator evaluated the three answers for each query, scoring them on four aspects: *Relevance*, *Coverage*, *Organization*, and *Overall Usefulness*, using a 1-5 scale. The final score for each aspect was calculated as the average of the three annotators' scores. The evaluation criteria are further detailed in Appendix B.

5 Experiment

We perform extensive experiments to evaluate the effectiveness, reliability, and reasoning capabilities of SciRAG across multiple scientific QA datasets.

5.1 Main Results

As shown in Table 2, across all evaluated benchmarks SciRAG consistently delivers top-tier answer quality, ranking first in correctness score on 4 out of 5 datasets. This demonstrates the effectiveness of our symbolic reasoning and outline-guided synthesis in producing coherent, accurate, and logically structured responses. Notably, SciRAG outperforms strong baselines such as OpenScholar-OS-GPT4o and PaperQA2, with clear gains on SciFact (+2.8), PubMedQA (+9.3), ScholarQA-CS (+11.3), and ScholarQA-MULTI (+0.28) compared to OpenScholar-OS-GPT4o, highlighting its strength in tackling complex, multi-faceted queries across diverse scientific domains.

However, SciRAG does not achieve the highest citation F1 on several benchmarks. This gap stems primarily from two factors. First, our system uses the general GPT-4o model, which lacks the specialized fine-tuning for citation accuracy that OS-GPT4o benefits from. Second, SciRAG sometimes generates summary-style or inferential statements that, while logically sound, do not have explicit one-to-one citation links, leading to lower scores under strict F1 evaluation criteria. Nevertheless, SciRAG maintains competitive citation precision scores, supported by robust citation graph expansion, symbolic reranking, and citation verification modules. Further manual analysis reveals that many seemingly uncited statements are in fact implicitly supported through multi-hop citation chains, a nuance missed by surface-level metrics.

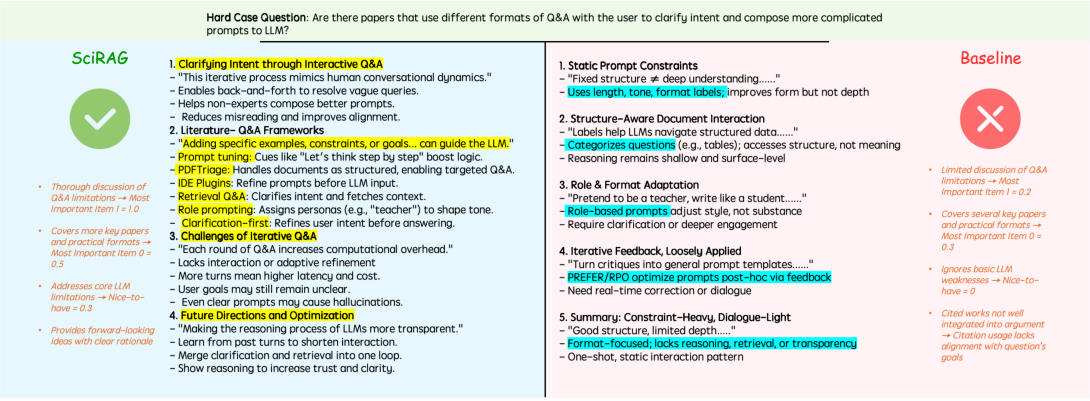


Figure 3: An example comparing SciRAG with a representative baseline.

Metric	SciRAG	OpenScholar	PaperQA2
Org.	**3.83**	3.75 (-0.08)	3.69 (-0.14)
Cov.	**4.00**	3.51 (-0.49)	2.62 (-1.38)
Rel.	3.49	**3.63** (+0.14)	3.38 (-0.11)
Usef.	**3.67**	3.30 (-0.37)	2.72 (-0.95)
Avg.	**3.75**	3.55 (-0.20)	3.10 (-0.65)

Table 3: Human evaluation on ScholarQA-CS. Numbers in parentheses show the difference vs. SciRAG.

5.2 Human Evaluation

We also conducted a qualitative evaluation of the generated outputs using three expert annotators The protocol is detailed in Section 4.3. The evaluation result is shown in Figure 3. Among the four aspects, SciRAG leads in *Organization*, *Coverage*, *Usefulness*, demonstrating SciRAG could generate a well-organized, comprehensive and useful result to scientific literature queries. SciRAG scores lower than OpenScholar in *Relevance*, primarily because its answers often include additional background or contextual information. While this content is designed to support comprehensive understanding, evaluators may sometimes perceive such introductory material as unrelated to the main question. To further assess reliability, we examined inter-annotator agreement. The average pairwise correlation between annotator scores was consistently high (around 0.87), confirming that the evaluations are stable and not driven by individual annotator variance.

5.3 Hallucination Analysis

To assess the reliability of uncited statements, we conducted a sentence-level evaluation on 100 randomly sampled generation cases, using the same granularity as citation F1. In each case, all sentences without explicit citations were extracted and reviewed by three independent LLM judges (GPT-4o, DeepSeek-R1, and Gemini 2.5 Pro). The judges determined whether each sentence was supported by the retrieved context, either directly or through multi-hop reasoning. The proportion of sentences judged unsupported was 6.0% for GPT-4o, 5.3% for DeepSeek-R1, and 6.6% for Gemini 2.5 Pro. These highly consistent results across models indicate that the vast majority of SciRAG's responses remain well-grounded in evidence.

5.4 Case Studies

To further evaluate the performance of SciRAG against the baseline, we conduct a representative hard case study, with the LLM-based evaluation results visualized in Figure 3. Compared to the baseline, which offers only a limited discussion of Q&A limitations, partial coverage of related works, and poorly aligned citations, SciRAG provides a more thorough treatment of Q&A challenges, integrates a wider range of key papers and practical formats, and explicitly addresses LLM weaknesses while offering forward-looking insights. This contrast shows that SciRAG produces answers that are both more comprehensive and analytically grounded. For completeness, we also include in Table 5 in Appendix the human evaluation of this case, where expert annotators likewise preferred SciRAG for its broader coverage and deeper reasoning.

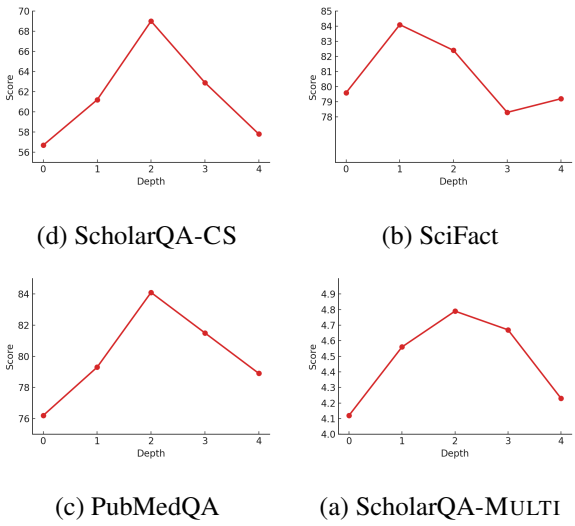
5.5 Ablation Studies on Reasoning Modules

We disentangle the contributions of SciRAG's two core reasoning modules—*symbolic reranking* and the *outline-based meta-planner*—through controlled ablations on three representative benchmarks (Table 4). Removing either module leads to performance degradation, but in different ways.

Symbolic reranking governs evidence preci-

Model	SciFact	ScQA-CS	ScQA-Multi
Full SciRAG	**84.1**	**69.0**	**4.79**
w. Dense Rerank	73.4 (-10.7)	58.6 (-10.4)	4.45 (-0.34)
Sequential-only	75.7 (-8.4)	59.1 (-9.9)	4.41 (-0.38)
Parallel-only	77.3 (-6.8)	58.2 (-10.8)	4.56 (-0.23)
w/o Planner	76.1 (-8.0)	52.1 (-16.9)	4.07 (-0.72)

Table 4: Impact of removing each module of SciRAG.



(d) ScholarQA-CS (b) SciFact

(c) PubMedQA (a) ScholarQA-Multi

Figure 4: Effect of retrieval depth on answer quality across four benchmarks.

sion: replacing it with a dense reranker results in sharp drops of 10.7 on SciFact, 10.4 on ScholarQA-CS, and 0.34 on ScholarQA-Multi. Dense models rely only on embedding similarity, often retrieving papers with overlapping terms but divergent focus. In contrast, our symbolic reranker filters evidence by conceptual role and logical consistency within citation chains, while also providing reasoning traces and rationales that enhance transparency.

The effects of retrieval strategy are similarly clear. **Sequential-only** and **Parallel-only** variants show that depth preserves logical continuity and breadth expands coverage, yet both fall short of the adaptive integration that combines their strengths. **The meta-planner**, meanwhile, drives global answer structure. Without it, the model loses rhetorical guidance, producing fragmented or incoherent responses. This limitation is most pronounced in complex synthesis tasks, where performance drops by 16.9 points on ScholarQA-CS and 0.72 on ScholarQA-Multi, underscoring the planner's role in coordinating multi-branch integration.

Taken together, symbolic reranking and the meta-planner form the backbone of SciRAG's reasoning pipeline: the former sharpens evidence selection, while the latter orchestrates synthesis. Their complementary roles are essential for producing precise, well-structured, and verifiable answers.

5.6 Scaling Behavior: Retrieval Tree Depth

We further evaluate the robustness of SciRAG under increasing retrieval tree depths. Figure 4 illustrates answer quality trends across ScholarQA-CS, SciFact, PubMedQA, and ScholarQA-Multi benchmarks as retrieval depth varies from 0 to 4 levels. A common trend emerges across all benchmarks: answer quality improves as retrieval depth increases from 0 to 2, before deteriorating at greater depths. Performance consistently peaks at depth 2 (or 1), suggesting an optimal balance between sufficient context and avoiding noise. For instance, PubMedQA improves from 76.2 (depth 0) to 84.1 (depth 2), indicating successful integration of relevant evidence. However, further expansion (depth 3–4) reduces quality, likely due to citation noise and tangential content overwhelming the model's reasoning capacity. This analysis confirms our design choice of using moderate depths, maximizing supportive evidence while minimizing irrelevant information. Consequently, these findings not only provide empirical validation for SciRAG's adaptive retrieval mechanism but also inform practical guidelines for optimal retrieval-depth selection in future research.

6 Conclusion

We presented **SciRAG**, a novel retrieval-augmented generation framework designed specifically for scientific literature exploration. By integrating adaptive retrieval, citation-aware symbolic reasoning, and outline-guided synthesis, SciRAG addresses key limitations of existing approaches, including inflexible retrieval, superficial citation usage, and fragmented answer construction. Comprehensive experiments on open-retrieval benchmarks such as SciFact, PubMedQA, and ScholarQA demonstrate that SciRAG achieves state-of-the-art performance in factual accuracy, coherence, and overall usefulness. It consistently generates coherent, well-organized responses grounded in verifiable evidence, as further validated by expert human evaluations.Looking forward, SciRAG provides a scalable and transparent foundation for trustworthy scientific question answering, with strong potential to assist researchers in navigating the ever-growing volume and complexity of scholarly literature.

Limitations

Despite its strong performance, **SciRAG** has certain limitations. First, it relies on general-purpose language models such as GPT-4o and Llama-3.1, which are not fine-tuned for scientific citation accuracy and may miss precise attribution. Second, the symbolic reasoning and outline-based synthesis introduce non-trivial computational overhead, which may affect real-time applicability in large-scale deployments. Also, the human evaluation was conducted with a limited number of expert annotators, which may not capture disciplinary variance. Future work could explore domain-specific model tuning and lightweight alternatives to improve both precision and efficiency.

References

Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80.

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222.

Crossref. 2024. Crossref metadata statistics. https://www.crossref.org/. Accessed: 2025-09-29.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieval step by step for large language models. *arXiv preprint arXiv:2502.01142*.

Yuntong Hu, Zhihan Lei, Zhongjie Dai, Allen Zhang, Abhinav Angirekula, Zheng Zhang, and Liang Zhao. 2025. Cg-rag: Research question answering by citation graph retrieval-augmented llms.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Mlagentbench: Evaluating language agents on machine learning experimentation. In *International Conference on Machine Learning*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William 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.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Hong-In Lee, and Moontae Lee. 2023. QASA: Advanced question answering on scientific articles. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19036–19052. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian

Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D. Hwang, Jason Dunkleberger, Matt Latzke, Smita Rao, Jaron Lochner, Rob Evans, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. 2025. Ai2 scholar qa: Organized literature synthesis with attribution.

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.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. 2024. Scicode: A research coding benchmark curated by scientists.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025. Chain-of-retrieval augmented generation.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery.

Jiasheng Zhang, Jialin Chen, Ali Maatouk, Ngoc Bui, Qianqian Xie, Leandros Tassiulas, Jie Shao, Hua Xu, and Rex Ying. 2024. Litfm: A retrieval augmented structure-aware foundation model for citation graphs. *arXiv preprint arXiv:2409.12177*.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, et al. 2024. Openresearcher: Unleashing ai for accelerated scientific research. *arXiv preprint arXiv:2408.06941*.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. 2024. Efficientrag: Efficient retriever for multi-hop question answering. *arXiv preprint arXiv:2408.04259*.

A Case Study

Hard Case Question: Are there papers that use different formats of Q&A with the user to clarify intent and compose more complicated prompts to LLM?
SciRAG: 1. Significance of Using Q&A Formats for Intent Clarification and Enhanced Prompt Composition in LLMs The use of Q&A formats to clarify user intent and enhance prompt composition is gaining prominence as an effective strategy to improve the quality [3][6] and relevance of interactions ... 2. Survey of Literature on Q&A-Based Frameworks for Refining User Intent and Generating Complex Prompts Recent studies have investigated various Q&A-based frameworks and methodologies for improving user interaction with LLMs[4]. These frameworks emphasize iterative clarification, intent disambiguation, and the construction of structured, context-rich prompts. One notable approach is ..... 3. Challenges and Limitations of Q&A Formats for Intent Clarification in LLMs Despite their advantages, Q&A-based frameworks for intent clarification face several challenges and limitations. One significant issue is scalability, particularly when .....[5] 4.Conclusion and Future Directions In summary, Q&A-based frameworks offer a promising approach to improving the quality and relevance of interactions with LLMs by refining user intent and ...[10] **Fine-grained Assessments:** Org 5.0 Rel 5.0 Con 5.0
OS-GPT4o Yes, several papers explore the use of different formats for question-and-answer interactions to clarify user intent and construct more sophisticated prompts . . . [2] Structured and Granular Prompt Engineering Structured and granular prompt engineering involves designing prompts that align with specific structural constraints, such as numerical, descriptive, and formatting requirements . . . [0] Dynamic Interaction with Structured Documents This section discusses user interactions with structured documents, focusing on numerical, descriptive, and formatting structural constraints. . . . [2] Adaptive and Role-Specific Prompts Taxonomy Numerical Constraining text to a set or a bounded number of words, sentences, or paragraphs is valuable in all aspects of writing . . . [2] Iterative Feedback and Post-Processing Iterative feedback and post-processing are valuable in improving the quality of responses generated by language models. Specifically, incorporating user feedback into . . . [1] Summary and Implications Collectively, these studies highlight the importance of using structured, adaptive, and iterative Q&A formats to clarify user intent and enhance prompt engineering for LLMs. Techniques such as feedback loops . . . [3] **Fine-grained Assessments:** Org 4.0 Rel 5.0 Con 4.0
Preference: SciRAG **Explanation:** Compared to the baseline, SciRAG offers more comprehensive topic coverage and a clearer structural flow. It systematically introduces key concepts, explores multiple relevant methods, and highlights potential limitations, enabling a deeper understanding of the question. In contrast, the baseline answer is fragmented, with less clear organization and limited depth across sections, making it harder to grasp the full scope of the discussion.

Table 5: Evaluation Case Study

B Human Evaluation Criteria

Inspired by OPENSCHOLAR (Asai et al., 2024), we define four fine-grained aspects to evaluate the generation. We ask the evaluator to give a score from 1 to 5 with regarding to the four aspects, *Relevance,*

Coverage, Organization and Overall Usefulness. The definition and instruction of scoring is shown in table 6

Aspect	Definition	Instructions
Organization	Evaluate if the output is well-organized and logically structured.	Score 1 means the response is disorganized, with no clear structure. Score 5 means the response is exceptionally well-organized, with a flawless logical structure
Coverage	Evaluate if the output provides sufficient coverage and amount of information.	Score 1 means the answer misses most of the key areas with few resources. Score 5 means the answer covers a diverse range of papers and viewpoints offering a thorough overview of the area.
Relevance	Evaluate if the response stay on topic and maintain a clear focus to provide a useful response to the question.	Score 1 means the content significantly deviates from the original question. Score 5 means the response remains tightly centered on the subject matter with enough depth for every piece of information.
Overall Usefulness	Evaluate if the output contains useful information to fulfill the information needs.	Score 1 means the response does not answer the question or provides rather confusing information. Score 5 means the response provides a comprehensive overview of the area, and sufficiently answers the question without additional references.

Table 6: Evaluation Criteria Descriptions

C Case Examples

Below is an example of our query depth and parallel dynamic retrieval. In this example, the sub-queries 1 and 3 were further refined through deeper retrieval after initial searching to gather more detailed information. On the other hand, sub-queries 2 and 4 employed a parallel retrieval strategy, where multiple sub-queries were independently expanded at the same time, thereby broadening the scope of retrieval and ensuring comprehensive coverage of multidimensional information.

Root Question

Are there papers that use different formats of Q&A with the user to clarify intent and compose more complicated prompts to LLM?

1. What are the different clarification question formats used to clarify user intent in Q&A systems with LLMs?

1.1

What are the different clarification question formats used to clarify user intent in Q&A systems with LLMs?

2. What performance metrics are used to assess the impact of clarification and multi-step questioning techniques on LLM accuracy?

2.1

What are the most widely used benchmarks to assess LLM performance in tasks involving clarification and multi-step questioning?

2.2

How is user engagement specifically measured when applying clarification or multi-step questioning techniques in LLM-based systems?

3. How does multi-step question generation work in LLMs to improve the precision of prompts?

3.1

What techniques are used for the adaptive generation of multi-step questions in LLMs to ensure prompt precision?

4. What are some real-world applications where multi-step Q&A strategies have been successfully applied to LLM systems?

4.1

What measurable impact has multi-step Q&A had on customer satisfaction in service-based industries?

4.2

How have multi-step Q&A strategies been specifically applied in healthcare for improving diagnostic accuracy or patient interactions?

D Simplified Snapshot of Contribution Chains and Symbolic Reranking

To complement the method description in §3.2 , we provide a compact snapshot that makes the mechanics of contribution-chain construction and symbolic reranking/filtering more transparent.


```
{ "query": "How do different formats of multi-turn Q&A (e.g., structured,
    iterative, and conversational) specifically address user intent
    clarification and prompt accuracy in complex scenarios like legal, medical
    , or technical domains?",

"path": "Are there papers that use different formats of Q&A with the user to
    clarify intent and compose more complicated prompts to LLM? -> How do
    different formats of multi-turn Q&A (e.g., structured, iterative, and
    conversational) specifically address user intent clarification and prompt
    accuracy in complex scenarios like legal, medical, or technical domains?",

"papers_count": 10,

"step1_analysis": { "output": "{"papers": [{"paper_index": 1,"segments": [{"
    label": "I","description": "Explores clarifying questions as a key
```

property in conversational search and mixed-initiative IR, highlighting the role of multi-turn Q&A.","relevance": "High"},{"label": "E","description": "Compares GPT-based simulators with human query reformulation patterns and discusses evaluation challenges for multi-turn clarification.","relevance": "Medium"}]},.....]}" },

"step2_relationships": { "output": "{"relationships": [{"symbol": "[1]T -> [Q]","rationale": "Paper 1's exploration of clarifying questions in conversational search informs the understanding of multi-turn Q&A formats for user intent clarification."},{"symbol": "[3]T -> [5]M","rationale": "Paper 3's discussion of multi-turn LLM performance challenges addresses issues relevant to improving accuracy in complex scenarios from the problems coming from Paper 5's methodology."},.....]}" },

"step3_analysis": { Crafting analysis plan > We need to follow the instructions step by step. I should avoid web-based responses, focusing on a detailed yet concise breakdown. I'll identify key papers, highlight contradictions and irrelevancies, and stick to the exact data provided. I'll include symbolic relationships and ensure responses address the query directly, using numbered lists for clarity. No extra claims should be made .> Breaking down paper contributions > I went through the symbolic relationships and the paper breakdowns and produced a focused, step-by-step analysis of how the set of papers coheres (or not) with the retrieval query: How do different formats of multi-turn Q&A (structured, iterative, conversational) specifically address user intent clarification and prompt accuracy in complex scenarios (legal, medical, technical)? # 1) Core papers (most central) and why they matter These papers most directly speak to mechanisms, failure modes, and empirical tests of multi-turn Q&A styles that aim to clarify intent and improve prompt accuracy. 1. Paper 1 (T + E) - T (Theory): Directly frames clarifying questions as core to conversational search and mixed-initiative interaction. This is central because clarifying Qs are the foundational behavior across structured, iterative, and conversational formats. - E (Evaluation): .....

Simulator realism / external validity tension - Papers involved: Paper 1 (E) vs Paper 5 (M/E). - Tension: Paper 1 s evaluation of GPT-based simulators diverging from human query reformulations raises questions about how well simulated user studies (including USi in Paper 5) generalize to real users. Paper 5 claims accurate simulation via crowdsourced validation, but Paper 1 flags evaluation challenges. - Implication: Any conclusions about format effectiveness based mainly on simulated interactions need corroboration with real users, or at least strong validation against human reformulations. 2. Performance drop (Paper 3) vs targeted improvements (Papers 6 & 8) - .....

Mapping the papers to the three format families (quick cheat-map) - Structured Q&A (explicit templates, decision policies, feature-driven): Paper 2 (E), Paper 6 (T/E), Paper 3 (T). - Iterative Q&A (stepwise clarification, short loops): Paper 1 (T), Paper 5 (M/E), Paper 3 (E). - Conversational / Proactive Q&A: .....

Most central: Papers 1, 3, 5, 6, 8 they collectively explain why multi-turn formats are necessary, where they fail, how to evaluate them, and concrete mitigation strategies (including a domain example in finance). - Useful context: Papers 2, 4, 7, 9 add signals ..... } ,
"step3_final_selection": { "final_selection": [{"paper_index": 1,"rank": 1,"justification": "Frames clarifying questions as central to conversational and iterative Q&A, with evaluation evidence on simulators vs human reformulations; directly relevant to intent clarification across formats ."}, {"paper_index": 3,"rank": 2,"justification": "Identifies multi-turn performance degradation and its causes, providing a theoretical and empirical foundation for why structured and iterative formats are needed to maintain prompt accuracy."}, ...] } }

E Prompt Templates

In this appendix, we include the key prompt templates used in the SciRAG framework for different stages of the pipeline. These prompts are designed to ensure consistency, interpretability, and modularity across planning, retrieval, reasoning, and synthesis phases.

Outline Generation Prompt

"Given a knowledge-intensive scientific question, please understand and analyze the question carefully. Consider key aspects like the intent behind the question, the motivations, potential flaws, and possible solutions. Based on this analysis, create a simple outline of the answer, specifying the parts and information that should be included in the response and the proportion each part should contribute. The total should add up to 100%. The outline should guide what should be covered in the response, offering clarity on the scope and balance of each section."
"Your answer should be marked as [Response_Start] Answer [Response_End]."
"Here's an example outline:"
"Question: What strategies are used to improve robustness and safety of quadrotor UAVs in extreme weather conditions?"
"Answer: [Response_Start]"
"1. (33%) The answer should begin by explaining the importance of robustness and safety for quadrotor UAVs in extreme weather conditions."
"2. (33%) The answer should discuss strategies and solutions to improve the robustness and safety of quadrotor UAVs in extreme weather conditions."
"3. (33%) The answer should highlight the limitations or challenges associated with designing robust and safe solutions for quadrotor UAVs under extreme weather conditions."
"[Response_End]"
"Now, please create an outline for this question: {question}"

Initial Answer Generation Prompt

"Provide a detailed and informative answer to the following research-related question. Your response should offer a comprehensive overview and be clearly structured in multiple paragraphs."
"Organize your answer according to the key themes or sections identified in the outline below, and please note that the length of each part of the answer should be roughly the same as the percentage in the outline. Ensure each section is well-supported by multiple references, not just a single source."
"Focus on giving a comprehensive overview of the topic, rather than providing a short or surface-level response."
"Ensure the answer is well-structured, coherent and informative so that real-world scientists can gain a clear understanding of the subject. Rather than simply summarizing multiple papers one by one, try to organize your answers based on similarities and differences between papers."
"Make sure to add citations to all citation-worthy statements using the provided references (References). More specifically, add the citation number at the end of each relevant sentence e.g., 'This work shows the effectiveness of problem X [1].' when the passage [1] in References provides full support for the statement."
"Not all references may be relevant, so only cite those that directly support the statement."
"If multiple references support a statement, cite them together (e.g., [1][2]). Yet, for each citation-worthy statement, you only need to add at least one citation, so if multiple evidences support the statement, just add the most relevant citation to the sentence."
"Your answer should be accurate and rigorous, preferably with citations to support each sentence."
"References: {context}"
"Question: {input}"
"Outline: {outline}"
"Your answer should be marked as [Response_Start] Answer [Response_End]."

Gap Identification and Subquery Generation Prompts

Prompt 1: Gap Identification
"Please review the 'Current Answer' based on the 'Outline Guidance' and the 'Original Query'.
Your sole task is to accurately identify and describe what information, as required by the 'Outline Guidance', is missing from the 'Current Answer'. List these gaps clearly and specifically.
When reviewing, ignore any content related to future work, conclusions, or acknowledgments.
If the 'Current Answer' already fulfills all requirements in the 'Outline Guidance', state this explicitly. For example: 'The answer is complete and contains no information gaps.'
Outline Guidance: {guidance}
Current Answer: {answer}
Original Query: {query}"

Prompt 2: Decision and Subquery Generation
"Please analyze the 'Identified Gaps' provided below.
If the 'Identified Gaps' list indicates that the answer is complete (e.g., the list is empty or explicitly states there are no gaps), you must return only '[end]terminate' in lowercase.
Otherwise, create one or more new search queries to gather the information needed to fill these gaps.
Follow these rules for creating queries:
- Merge Similar Queries: If multiple gaps can be addressed by similar queries, merge them.
- Minimize Query Count: Ensure each query explores a different sub-problem and provide as few queries as possible.
- Be Clear and Concise: Each query must be clear, concise, and contain necessary keywords to guide the retrieval process effectively.
- Do Not Reference the Answer: Do not mention or reference specific content from the 'Current Answer' in your new queries.
Please return your queries in the format below:
(1) Your query content.
(2) Additional query content if needed.
...
Identified Gaps: {gap_analysis}
Original Query (for context): {query}"

Additional Answer Generation Prompt

"You are in a retrieval chain that has been expanded to better answer the initial research-related core query."
"The retrieval path is: {path}."
"Currently, you are at the retrieval step for: {query}."
"Provide a detailed and informative answer only to the query at current step. Your response should offer a concrete answer."
"Make sure your answer includes summaries of relevant literature or texts or clear descriptions of their contribution to the query. When you make a claim, it is always best to have excerpts or citations to support them."
"Ensure your answer is well-supported by references. Focus on giving a concrete answer to the query, rather than providing a short or surface-level response."
"Ensure the answer is well-structured, coherent and informative so that real-world scientists can gain a clear understanding of the query. Rather than simply summarizing multiple papers one by one, try to organize your answers based on similarities and differences between papers."
"Make sure to add citations to all citation-worthy statements using the provided references (References). More specifically, add the citation number at the end of each relevant sentence e.g., 'This work shows the effectiveness of problem X [1].' when the passage [1] in References provides full support for the statement."
"Not all references may be relevant. You can read through the rationales and think on your own, and only cite those that directly support the statement."
"If multiple references support a statement, cite them together (e.g., [1][2]). Yet, for each citation-worthy statement, you only need to add at least one citation, so if multiple evidences support the statement, just add the most relevant citation to the sentence."
"Your answer should be accurate and rigorous, preferably with citations to support each sentence."
"References: {context}"
"Your answer should be marked as [Response_Start] Answer [Response_End]."

Branch-Based Synthesis Prompt

"You are in a retrieval chain that has been expanded to better answer the initial research-related core query."
"The retrieval path is: {path}."
"Currently, you are at the retrieval step for query: {query}."
"Please review the current research-related query and its initial answer and read them carefully."
"The initial answer may have some shortcomings, so we performed additional searches and supplemented information. Now please combine the information from the supplemented query and answer to optimize the original answer, offering a comprehensive overview and clearly structured in multiple paragraphs."
"Also you should try to keep the original answer content's structure unchanged."
"Ensure the answer is well-structured, coherent and informative so that real-world scientists can gain a clear understanding of the subject, rather than providing a short or surface-level response."
"And re-cite the citations in the answer according to the latest reference list below."
"Make sure your answer includes summaries of relevant literature or texts or clear descriptions of their contribution to the query. When you make a claim, it is always best to have excerpts or citations to support them."
"Make sure to add citations to all citation-worthy statements using the provided references (References). More specifically, add the citation number at the end of each relevant sentence e.g., 'This work shows the effectiveness of problem X [1].' when the passage [1] in References provides full support for the statement."
"Not all references may be relevant. You can read through the rationales and think on your own, and only cite those that directly support the statement."
"If multiple references support a statement, cite them together (e.g., [1][2]). Yet, for each citation-worthy statement, you only need to add at least one citation, so if multiple evidences support the statement, just add the most relevant citation to the sentence."
"Your answer should be accurate and rigorous, preferably with citations to support each sentence."
"Here is the initial answer: {answer}"
"Here is the supplemented queries and answers: {supplement}"
"Here is the references: {context}"
"Your answer should be marked as [Response_Start] Answer [Response_End]."

System Instruction: Symbolic Reasoning Setup

"You are in a retrieval chain that has been expanded to better answer the initial core query."
"The retrieval path is: {path}."
"Currently, you are at the retrieval step for: {query}."
"You have a set of partial paper texts (abstracts or snippets)."
"Your goal is to analyze each text's contribution and the relationship between them, build symbolic relationships,"
"and decide which texts are most relevant and contributing to the query and the overall chain."

Symbolic Reasoning — Step 1: Role Tagging and Relevance Assessment

"We have the following candidate texts from different papers (abstracts or snippets): {paper_text}"
"The query is: {query_text}"

Step 1 Task:
1. For each paper, identify its key content segments and label them with:
- T (theoretical part: theorem, definitions, main theoretical results),
- E (experimental part: methodology, experiment details, results),
- A (applications),
- or other labels if needed (e.g., 'M' for methodology if it's not purely experimental).
2. For each segment, provide a brief summary (1–2 sentences)
and assess its relevance to the query as High, Medium, or Low.
Output format (example):

```
{
"papers": [
{
"paper_index": 1,
"segments": [
{ "label": "T", "description": "...", "relevance": "High" },
{ "label": "E", "description": "...", "relevance": "Medium" }
]
},
...
]
}
```

Please keep the output structure strictly without additional comments.

Symbolic Reasoning — Step 2: Link Construction and Reasoning

"Below is the structured breakdown of each paper's segments from Step 1:"
{step1_result_json}
"Using that breakdown, please establish symbolic relationships among the papers and the query: {query}"
" - T (theoretical part: theorem, definitions, main theoretical results),"
" - E (experimental part: methodology, experiment details, results),"
" - A (applications),"
" - or other labels if needed (e.g., 'M' for methodology if it's not purely experimental)."
"For example:"
" - [1]T -> [2]T means paper1's theoretical part informs or extends paper2's theoretical part."
" - [1]E -> [Q] means paper1's experiment part contributes directly to answering the query."
" - [3]A -> [2]T means paper3's application part provides insights for paper2's theory."
"In each relationship, use the format: [paper_index][label] -> [paper_index or Q][label (if paper)]."
"If the second target is the query itself, just use [Q]."
Output format (example):
{
"relationships": [
{ "symbol": "[1]T -> [Q]", "rationale": "Paper1's theoretical result directly addresses the phenomenon in the query." },
{ "symbol": "[2]E -> [3]T", "rationale": "Paper2's experiment suggests data that confirms the theorem in Paper3." },
...
]
}
Please keep the rationale concise, and keep the output structure strictly without additional comments.

Symbolic Reasoning — Step 3a: Coherence and Relevance Analysis

"Given the symbolic relationships from Step 2 and the paper breakdowns from Step 1, your task is to perform a detailed analysis."
"Symbolic Relationships: {step2_relationships_json}"
"Paper Breakdowns: {step1_result_json}"
"Query: '{query}'"
"Retrieval Chain: '{path}'"
Your Task:
Analyze the coherence and relevance of the papers and their relationships in the context of the query. **Do NOT** decide which papers to keep or discard yet. Instead, provide step-by-step reasoning that addresses the following:
- **Identify Core Papers**: Which papers (and their segments like T, E, A) appear to be most central to answering the query? Explain why.
- **Identify Supporting Papers**: Which papers provide useful context or supplementary information but may not be essential?
- **Identify Contradictions or Weak Links**: Are there any relationships in the chain (e.g., T $\rightarrow$ T) that seem weak, irrelevant, or contradictory? For instance, does one paper's experiment invalidate another's theory?
- **Identify Irrelevant Papers**: Are there papers that seem entirely tangential or irrelevant to the specific query? Explain your reasoning.
Your output should be a clear, textual analysis that will be used in the next step to make final decisions.

Symbolic Reasoning — Step 3b: Final Decision and Ranking

"We now have the symbolic relationships from Step 2:"
{step2_relationships_json}
Where we have the symbols:
- T (theoretical part: theorem, definitions, main theoretical results),
- E (experimental part: methodology, experiment details, results),
- A (applications),
- or other labels if needed (e.g., 'M' for methodology if it's not purely experimental).
And the breakdown of each paper from Step 1:
{step1_result_json}
For the query "{query}" within the context of the overall retrieval chain "{path}".
And based on the detailed 'Coherence and Relevance Analysis' provided below, your task is to make the final decisions on paper selection and ranking.
Coherence and Relevance Analysis: {analysis_from_step3a}
Your Task:
- **Finalize Selection**: Decide which papers to keep and which to discard based only on the provided analysis.
- **Rank Kept Papers**: Create a final ranked list (from most to least relevant) of the papers you decide to keep.
- **Justify Decisions**: For each kept paper, provide a concise justification for its rank. For each discarded paper, provide a brief reason for its exclusion. All justifications must be derived from the analysis. (You may consider the segments ([1]T, [1]E, etc.) internally, but your final output should only include paper indexes.)
Output Format:
You must return the final result in a valid JSON structure, exactly as shown in the example below, without any additional text or comments.
{
"final_selection": [
{ "paper_index": 1, "rank": 1, "justification": "..." },
{ "paper_index": 3, "rank": 2, "justification": "..." }
],
"discarded_items": [
{ "paper_index": 2, "reason": "Not relevant to the query" }
]
}

F Efficiency and Scaling Analysis

We also report efficiency alongside scaling results on ScholarQA-CS here. Table 7 summarizes the average runtime and cost per query at different retrieval depths. As expected, deeper retrieval incurs higher time and monetary cost, but the growth remains manageable relative to the performance gains reported in the main text.

Table 7: Average runtime and cost per query at different retrieval depths (ScholarQA-CS).

Depth	Reasoning / Rerank Time	Other Components Time	Total Time	Total Cost (USD/query)
1	1m16s	0m38s	1m54s	0.04
2	1m52s	0m47s	2m39s	0.09
3	3m38s	1m14s	4m52s	0.17
4	5m13s	1m59s	7m12s	0.29