# Ariadne: A Multi-Agent Collaborative System for Interactive Literature Analysis and Research Support

**Anonymous ACL submission**

## Abstract

The rapid expansion of scholarly publications has resulted in severe information overload, posing significant challenges for researchers in retrieving, evaluating, and synthesizing scientific knowledge. While large language models (LLMs) have shown potential in assisting scientific workflows, existing approaches often suffer from hallucinations and lack support for iterative, exploratory research. We introduce **Ariadne**, a multi-agent collaborative system designed for interactive literature analysis. Ariadne dynamically adapts to evolving research intents in the course of user interaction, employs flexible retrieval strategies, and performs hierarchical evidence synthesis to more effectively address complex scientific queries. Experiments on single-turn scientific QA benchmarks, including SciFact and SCHOLARQA-MULTI, demonstrate state-of-the-art performance. Moreover, human evaluations human evaluations in real-world research scenarios indicate that Ariadne delivers superior performance compared to existing baselines.

## 1 Introduction

In recent years, the volume of scholarly publications has grown rapidly across diverse disciplines. This information explosion has made research more complex, requiring researchers to continuously synthesize findings from fragmented sources (Shao et al., 2024; Wang et al., 2024). Traditional tools for literature analysis, such as keyword search and citation networks, are no longer sufficient for timely and comprehensive scientific inquiry. To address these limitations, large language models (LLMs) (Park et al., 2023; Anthropic, 2024; Team et al., 2023) have been integrated into research workflows, leveraging their advanced language capabilities to support scientific work, which have significantly advanced literature analysis and research workflows (Jiang et al., 2025).

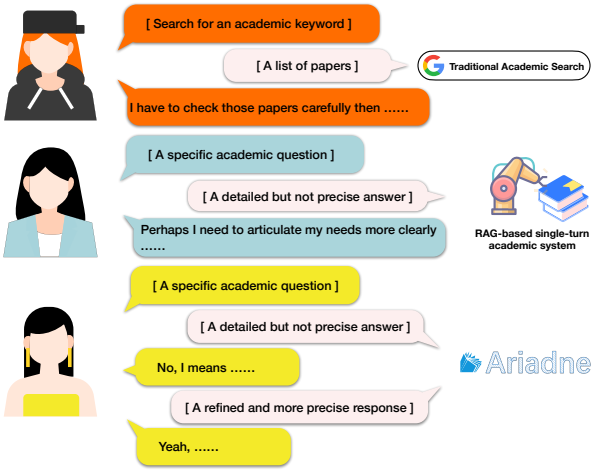


Figure 1: **The Evolution of Literature Analysis.** The progression from manual search and single-turn QA systems to our multi-agent assistant (Ariadne), which more effectively supports scientific research.

Current LLM-based literature analysis approaches fall into two main paradigms. **The first paradigm** stores knowledge directly within model parameters, as seen in systems like Med-PaLM (Singhal et al., 2023) and SciGLM (Zhang et al., 2024). While these models can generate fluent responses, they often suffer from outdated knowledge due to their fixed parameters (Gekhman et al., 2024) and are prone to hallucinations and false citations, making them less suitable for rapidly evolving research fields.

**The second paradigm** is based on Retrieval-Augmented Generation (RAG) frameworks, which leverage external document retrieval for more up-to-date information (Agarwal et al., 2024). For example, AutoSurvey (Wang et al., 2024) and SurveyX (Liang et al., 2025) generate research surveys by iteratively expanding outlines, while PaperQA (Skarlinski et al., 2024) and OPENSCHOLAR (Asai et al., 2024) utilize vector-based retrieval to provide more precise citation grounding in response to user queries. Compared to the first paradigm,

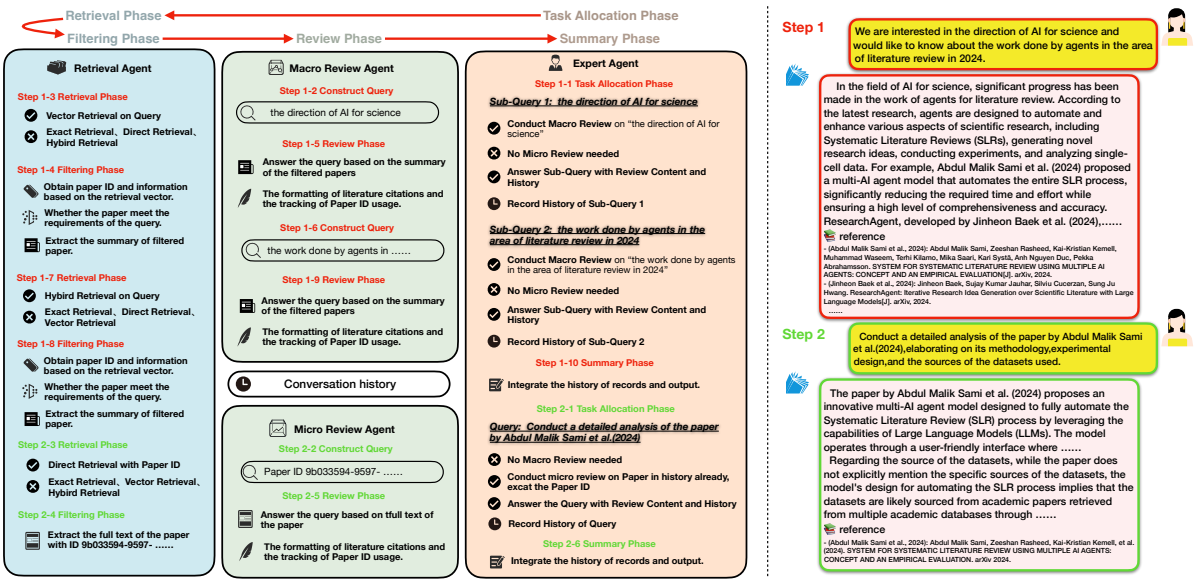


Figure 2: **A practical case of the Ariadne system.** This diagram illustrates the collaborative workflow of the three agents: The Expert Agent coordinates the system, while the Review Agents handle literature review based on the query. The Retrieval Agent ensures accurate literature retrieval.

these methods can switch between base models without retraining, and external retrieval reduces hallucination. However, most systems are **limited to single-turn interactions**, lacking iterative guidance for in-depth exploration. Their **reliance on vector-based retrieval** also limits effectiveness for complex, structured queries, such as filtering by author or publication period.

To address the limitations of existing literature analysis methods in the second paradigm, a multi-agent collaborative literature research system, **Ariadne**, is proposed with the following key contributions:

- **Multi-agent collaboration**: Ariadne employs a hierarchical multi-agent architecture, where Expert, Review, and Retrieval Agents coordinate to decompose tasks, retrieve evidence, and synthesize answers.
- **Hierarchical evidence flow**: The system supports both Macro and Micro review modes, allowing for high-level overviews and detailed analyses, respectively, to ensure comprehensive and traceable answers.
- **Adaptive retrieval and filtering**: Ariadne dynamically selects among vector, exact, and hybrid retrieval strategies, and leverages LLM-based filtering to ensure high relevance and precision.

These innovations enable **Ariadne** to overcome existing limitations, serving as an adaptive research partner that enhances research efficiency and quality.

## 2 Methodology

**Ariadne** is a multi-agent system designed to assist research work through interactive dialogue. The overall architecture and workflow are illustrated in Figure 2.

### 2.1 Expert Agent

The Expert Agent acts as the central controller, responsible for user interaction, task decomposition, and answer integration.

**Task Allocation Phase** As shown in Step 1-1 Task Allocation Phase, the Expert Agent first decomposes the user's query $q$ into a sequence of sub-questions:

$$Q = \text{Decompose}(q, H) \tag{1}$$

where $H$ represents the dialogue history.

For each $q_i \in Q$, the Expert Agent determines the appropriate review mode $m_i$ (Macro or Micro) for $q_i$:

$$m_i = \text{SelectReviewMode}(q_i) \tag{2}$$

Here, Macro mode applies to broad questions, and Micro mode to narrow, detail-oriented ones.

Then, the task is delegated to the Review Agent for further analysis:

$$r_i = \text{ReviewAgent}(q_i, m_i) \tag{3}$$

After obtaining $r_i$, the Expert Agent synthesizes the answer $a_i$ for each $q_i$ based on $r_i$ and the dialogue context $H$:

$$a_i = \text{Synthesize}(r_i, H) \tag{4}$$

**Summary Phase** As shown in Step 1-10 Summary Phase, once responses $a_i$ have been generated for each sub-question $q_i$, the Expert Agent synthesizes the final response by integrating these $a_i$. It ensures consistency with the research context, formats citations appropriately, and produces the final output $A$ for the user:

$$A = \text{Integrate}(\{a_1, ..., a_n\}, H) \tag{5}$$

This design encapsulates the entire workflow within the Expert Agent, abstracting the details of evidence retrieval and review.

## 2.2 Review Agent

The Review Agent serves as the bridge between the Expert Agent and the Retrieval Agent, responsible for synthesizing evidence for each sub-question.

**Review Phase** For each $q_i$ and its assigned $m_i$, the Review Agent first generates a adapted retrieval query $q_i^{\text{retr}}$ based on $q_i$:

$$q_i^{\text{retr}} = \text{GenerateRetrievalQuery}(q_i) \tag{6}$$

It then invokes the Retrieval Agent with $q_i^{\text{retr}}$ and the specified mode to obtain candidate evidence $C_i$:

$$C_i = \text{RetrievalAgent}(q_i^{\text{retr}}, m_i) \tag{7}$$

Finally, the Review Agent synthesizes the review result $r_i$ as follows:

$$r_i = \begin{cases} \text{MacroReview}(C_i), & \text{if } m_i = \text{Macro} \\ \text{MicroReview}(C_i), & \text{if } m_i = \text{Micro} \end{cases} \tag{8}$$

The MacroReview function focuses on synthesizing high-level overviews from a large volume of literature, identifying underlying patterns, trends, and consensus. When the number of papers is too large, it processes them in batches and then summarizes the results.

The MicroReview function performs in-depth analysis of specific factual content, extracting concrete paper fragments or full texts to uncover more fine-grained information.

## 2.3 Retrieval Agent

The Retrieval Agent is the backbone of **Ariadne**, responsible for preprocessing and retrieving academic papers using multiple strategies tailored to different query types.

**Preprocessing Phase** As illustrated in Figure 3, the preprocessing phase involves extracting textual content from academic papers. This is handled through two complementary approaches: (1) using an LLM to extract paper-level overviews, capturing key aspects such as main ideas, background, methodology, and findings; (2) slicing the full text into sentence-level embeddings for retrieval. Both are organized by paper ID to ensure that related information from the same paper remains interconnected.

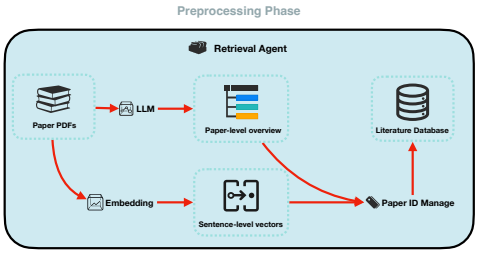


Figure 3: **Preprocessing phase of the Retrieval Agent.** LLMs generate paper-level overviews, while sentence-level embeddings are obtained for retrieval. All data is indexed and stored in the paper database.

**Retrieval Phase** As shown in Step 1-3 Retrieval Phase, for each adapted retrieval query $q_i^{\text{retr}}$ generated by the Review Agent, the Retrieval Agent retrieves a set of candidate contents $C_i$ from the literature database $\mathcal{D}$:

$$C_i = \text{Retrieve}(q_i^{\text{retr}}, \mathcal{D}) \tag{9}$$

where $\mathcal{D}$ represents the preprocessed database of academic papers, including both paper-level overviews and sentence-level content.

To address different query types, the Retrieval Agent supports multiple retrieval strategies:

- **Vector Retrieval**: Content is retrieved based on semantic similarity with $q_i^{\text{retr}}$:

$$C_i^{\text{vec}} = \text{TopK}_{c \in \mathcal{D}}\big(\text{sim}(E(q_i^{\text{retr}}), E(c))\big) \tag{10}$$

  where $E(\cdot)$ denotes the embedding model, sim is a similarity metric (typically cosine similarity), and TopK selects the $K$ most similar items.

- **Exact Retrieval**: For $q_i^{\text{retr}}$ containing precise identifiers (e.g., author names, paper titles, publication dates), the agent directly matches these in the database:

$$C_i^{\text{exact}} = \{c \in \mathcal{D} \mid \text{Match}(q_i^{\text{retr}}, c)\} \quad (11)$$

  The Match function performs fuzzy matching on bibliographic metadata and exact matching on quotes or paper identifiers.

- **Hybrid Retrieval**: For mixed $q_i^{\text{retr}}$, exact matching is first performed to filter relevant paper IDs, followed by vector-based retrieval within this subset:

$$C_i^{\text{hybrid}} = \text{TopK}_{c \in C_i^{\text{exact}}}\left(\text{sim}(E(q_i^{\text{retr}}), E(c))\right) \quad (12)$$

- **Direct Retrieval**: If $q_i^{\text{retr}}$ explicitly includes a paper ID, the agent directly retrieves the corresponding paper (not shown as a formula for brevity).

**Filtering Phase** As shown in Step 1-4 Filtering Phase, after retrieval, the candidate set $C_i$ is further filtered by an LLM to ensure relevance and coherence with respect to $q_i$:

$$F_i = \text{Filter}_{\text{LLM}}(C_i, q_i) \quad (13)$$

This phase significantly improves the relevance and coherence of the final output. The LLM evaluates each candidate's relevance to the original question $q_i$ and removes irrelevant or tangential content.

## 3 Experiment

To evaluate system performance, a series of experiments were designed to cover both single-turn and multi-turn scenarios. Public single-turn QA benchmarks were first used for objective evaluation, followed by human-in-the-loop multi-turn interactions for subjective assessment. Additional analyses, such as ablation studies and phase-wise citation tracking, further reveal the strengths and limitations of the system.

### 3.1 Single-turn QA Evaluation

The first stage evaluates factual accuracy and citation faithfulness in a single-turn QA setting.

**Evaluation Tasks.** Two benchmark tasks are included. The SciFact (Wadden et al., 2020) task involves claim verification with sentence-level evidence identification. The SCHOLARQA-MULTI (Asai et al., 2024) requires generating citation-grounded answers to academic questions.

**Compared Methods.** Five representative methods were tested:ChatGPT-4o (OpenAI, 2024), OPENSCHOLAR (Asai et al., 2024), PaperQA v2 (Skarlinski et al., 2024), Naive RAG (Lewis et al., 2020), and Ariadne. The latter three methods all utilize the "text-embedding-3-small"[1] model for embedding.

**Parameter Settings.** PaperQA v2 was configured with its default parameters, while Naive RAG utilized the same vector store as Ariadne, with top-$k$ set to 5 for SciFact and 10 for SCHOLARQA-MULTI.

**Answer Evaluation Metrics.** For SciFact, we report the average and standard deviation of precision, recall, and F1 for claim verification. For ScholarQA-multi, we use Organization, Coverage, and Relevance metrics (see Appendix B).

**Citation Quality Metrics.** Given a model citation set $C$ and ground truth $G$, compute Precision, Recall, and F1 as:

$$\text{Precision} = \frac{|C \cap G|}{|C|} \quad (14)$$

$$\text{Recall} = \frac{|C \cap G|}{|G|} \quad (15)$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (16)$$

These metrics reflect the system's citation accuracy.

**Result on SciFact.** Table 1 shows Ariadne outperforms existing methods in both judgment and citation accuracy, with stable performance.

**Result on SCHOLARQA-MULTI.** Table 2 shows Ariadne achieves the best overall performance on SCHOLARQA-MULTI, leading in content quality and citation metrics, but at higher cost per question. ChatGPT-4o is lowest cost but less accurate, while PaperQA v2 excels in citation precision.

### 3.2 Tracing Citation Quality

We analyzed citation quality across the Retrieval, Filtering, Review, and Summarization phases, tracking precision, recall, and F1 at each step. Results are averaged over three runs.

[1] https://platform.openai.com/docs/models/text-embedding-3-small

| Metrics | Acc. | Citation | | | Cost ($/question) |
|---|---|---|---|---|---|
| | | Precision | Recall | F1 | |
| **ChatGPT-4o** | 77.9 ± 1.7 | 2.7 ± 0.7 | 6.2 ± 1.6 | 3.7 ± 0.9 | 0.0010 ± 0.0002 |
| **+Naive RAG** | 89.5 ± 0.3 | 38.3 ± 15.1 | 93.3 ± 0.8 | 49.3 ± 12.8 | 0.009 ± 0.001 |
| **+OpenScholar** | 81.3* | - | - | 56.5* | 0.05* |
| **+Paper-QA V2** | 88.1 ± 2.2 | 87.3 ± 8.6 | 92.2 ± 3.9 | 87.9 ± 7.5 | 0.051 ± 0.006 |
| **+Ariadne** | **90.9 ± 0.5** | **90.9 ± 3.6** | **93.8 ± 1.4** | **91.1 ± 2.2** | 0.136 ± 0.026 |

Table 1: **SciFact Benchmark Results.** This table presents the performance of different systems on the SciFact dataset, including overall accuracy (Acc.), citation metrics (Precision, Recall, F1), and cost per query (Cost ($/question)). * means data directly sourced from (Asai et al., 2024), provided as reference. The results are reported as the mean ± standard deviation over three runs.

| Metrics | Generation | | | Citation | | | Cost ($/question) |
|---|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 | |
| **ChatGPT-4o** | 3.54 ± 0.04 | 3.19 ± 0.08 | 3.37 ± 0.03 | 6.0 ± 0.5 | 4.4 ± 0.4 | 4.8 ± 0.1 | 0.0034 ± 0.0006 |
| **+Naive RAG** | 3.71 ± 0.07 | 3.51 ± 0.23 | 3.56 ± 0.04 | 35.8 ± 4.6 | 35.6 ± 3.3 | 33.2 ± 0.3 | 0.011 ± 0.002 |
| **+OpenScholar** | - | - | - | - | - | 37.5* | 0.05* |
| **+Paper-QA V2** | 3.39 ± 0.03 | 3.07 ± 0.01 | 3.37 ± 0.04 | **58.5 ± 0.6** | 31.0 ± 0.8 | 38.5 ± 0.7 | 0.061 ± 0.025 |
| **+Ariadne** | **3.86 ± 0.01** | **3.75 ± 0.07** | **3.55 ± 0.03** | 58.3 ± 2.2 | **46.6 ± 1.0** | **49.2 ± 0.3** | 0.228 ± 0.144 |

Table 2: **SCHOLARQA-MULTI Benchmark Results.** This table summarizes the performance of various systems on the ScholarQA-multi dataset, reporting scores for generation quality (Organization, Coverage, Relevance), citation accuracy (Precision, Recall, F1), and cost per query (Cost ($/question)). * means data directly sourced from (Asai et al., 2024), provided as reference. The results are reported as the mean ± standard deviation over three runs.

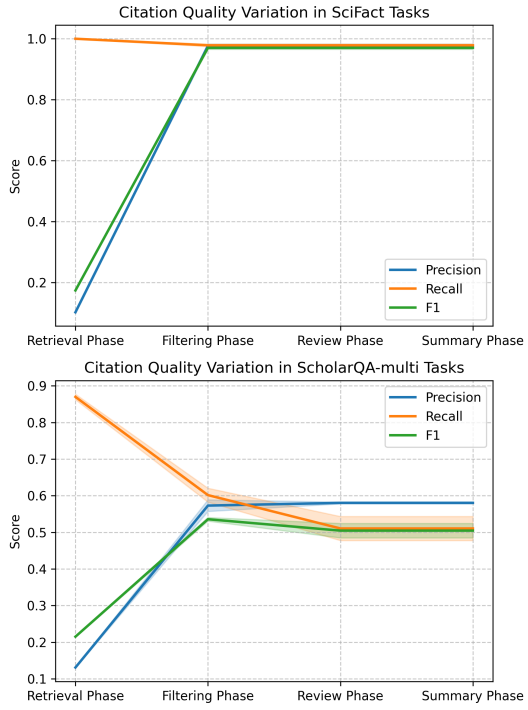


Figure 4: **Citation Quality Variation Across Different Phases.** The top panel presents the precision, recall, and F1 scores across the four phases (Retrieval, Filtering, Review, and Summarization) in the SciFact benchmark, while the bottom panel illustrates the same metrics for the SCHOLARQA-MULTI benchmark.

**Experimental Setup** To evaluate the evolution of citation quality throughout the reasoning process, a phase-wise analysis of Ariadne's citation workflow was conducted, encompassing the **Retrieval Phase**, **Filtering Phase**, **Review Phase**, and **Summarization Phase**. At each phase, the set of retained citations was tracked and their alignment with gold-standard references was assessed using precision, recall, and F1 score. All results are averaged over three independent runs, with both means and standard deviations reported.

**Results and Analysis** Figure 4 presents the phase-wise citation quality metrics, revealing several common trends across two benchmarks. In the **Retrieval Phase**, recall is generally high, indicating that the initial paper retrieval phase effectively captures most relevant citations. However, precision and F1 scores are typically lower in this phase, reflecting the inclusion of a substantial number of non-essential papers, which dilutes the overall quality.

In the **Filtering Phase**, a significant increase in precision and F1 is observed, driven by the removal of irrelevant citations. This phase effectively narrows the citation set to a more precise subset, although this refinement often comes at the cost of reduced recall, as some potentially relevant but

lower-confidence citations are also filtered out.

As the workflow progresses into the **Review Phase** and **Summarization Phase**, the citation set continues to stabilize, with precision and F1 scores typically reaching their peak, reflecting the final consolidation of contextually relevant evidence. This trend suggests that the later phases effectively prioritize citation quality over coverage, aligning the retained citations more closely with the target answers.

**Benchmark-Specific Differences** Despite the overall consistency in trends across both benchmarks, notable differences still emerge. In the **Filtering Phase**, SciFact exhibits a sharper improvement in precision with only a slight drop in recall, whereas SCHOLARQA-MULTI experiences a more significant decline in recall.

As shown in the Appendix A, these differences can be attributed to the nature of the questions in each banchmark: SciFact focuses on factual verification based on a small number of papers, where citations are few and closely aligned with the question. In contrast, SCHOLARQA-MULTI questions draw on a broader set of references, with each document contributing partially to the final answer, making them more prone to being discarded during filtering.

### 3.3 Ablation Study on the Review Agent

An ablation study was conducted to assess the individual contributions of Micro Review and Macro Review in the Review Phase.

**Experimental Settings** On both benchmarks, either Macro or Micro Review was disabled. Each experiment was repeated three times, and the mean ± standard deviation was reported.

**Results and Analysis** As shown in Table 3 and Table 4, using only Micro or Macro Review degrades overall performance, with only marginal, inconsistent gains in some metrics.

Notably, SCHOLARQA-MULTI drops more without Macro Review, while SciFact is more sensitive to removing Micro Review, reflecting dataset-specific preferences.

Using both strategies, Ariadne achieves the best or near-best performance, highlighting their complementarity and the importance of flexible coordination.

### 3.4 Multi-turn Human Evaluation

To complement single-turn benchmarks, multi-turn, subjective human evaluations were conducted to capture richer user experience insights, particularly in scenarios requiring extended dialogue and complex reasoning.

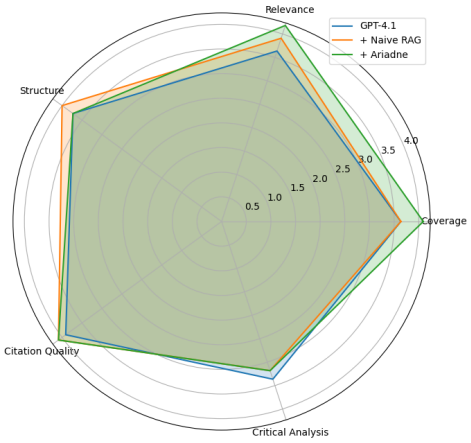


Figure 5: **Human Evaluation Results.** Average scores of Ariadne, Naive RAG, and ChatGPT on five evaluation dimensions. Ariadne achieves the best overall performance, especially in coverage and relevance.

**Experimental Setup** Given the education domain's reliance on paywalled papers, it provides a fitting testbed for evaluating the proposed method. To support this, a curated database of 1,613 top-tier education journal articles was constructed, simulating realistic academic search conditions. Ten graduate students in education were recruited for multi-turn interactions, focusing on real-world applicability and user experience.

For comparison, three methods were included: GPT-4.1, GPT-4.1 + Naive RAG, and GPT-4.1 + Ariadne. The Naive RAG used the same embeddings as the retrieval agent in Ariadne, retrieving the top 10 most similar segments for each user query at the beginning of each conversation. Ariadne also set the retrieval agent to extract 10 passages or segments to ensure consistency in the retrieval step.

Participants engaged with each method for 5-10 rounds, with system identities blinded to reduce bias, and the outputs were rated according to Table 5.

**Results and Analysis** Figure 5 presents the average scores of Ariadne, Naive RAG, and ChatGPT across five evaluation dimensions. Ariadne

| Metrics | Acc. | Citation | | |
|---|---|---|---|---|
| | | Precision | Recall | F1 |
| **Ariadne** | **90.9 ± 0.5** | 90.9 ± 3.6 | **93.8 ± 1.4** | 91.1 ± 2.2 |
| **-Micro** | 87.9 ± 0.6 ↓ | 79.9 ± 1.6 ↓ | 80.6 ± 1.1 ↓ | 79.8 ± 1.4 ↓ |
| **-Macro** | **90.9 ± 0.5** | **91.2 ± 2.7** ↑ | 93.6 ± 1.2 ↓ | **91.3 ± 1.6** ↑ |

Table 3: **SciFact Ablation Study Results.** Performance comparison between Ariadne and its ablated variants on the SciFact dataset, including overall accuracy (Acc.) and citation metrics (Precision, Recall, F1). Results are reported as mean ± standard deviation over three runs. ↑/↓ indicates performance increase/decrease compared to the base Ariadne model. **Bold** numbers denote the best results in each column.

| Metrics | Generation | | | Citation | | |
|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 |
| **Ariadne** | **3.861 ± 0.008** | **3.747 ± 0.070** | 3.546 ± 0.033 | 58.3 ± 2.2 | **46.6 ± 1.0** | 49.2 ± 0.3 |
| **-Micro** | 3.812 ± 0.038 ↓ | 3.623 ± 0.038 ↓ | **3.556 ± 0.013** ↑ | **60.8 ± 0.5** ↑ | 45.9 ± 1.1 ↓ | **50.0 ± 0.8** ↑ |
| **-Macro** | 3.818 ± 0.004 ↓ | 3.642 ± 0.029 ↓ | 3.543 ± 0.024 ↓ | 56.8 ± 1.5 ↓ | 40.9 ± 1.7 ↓ | 45.6 ± 1.5 ↓ |

Table 4: **SCHOLARQA-MULTI Ablation Study Results.** Performance comparison between Ariadne and its ablated variants. Results show mean ± standard deviation across three runs. ↑/↓ indicates performance increase/decrease compared to the base Ariadne model. **Bold** numbers denote the best results in each column.

consistently outperforms the other systems, demonstrating its advantage in understanding user intent, managing context, and providing comprehensive responses. These results highlight the benefits of multi-agent collaboration in supporting complex academic research.

## 4 Discussion and Future Direction

**Impact of Base Models** In the experiments, representative models such as ChatGPT-4o and ChatGPT-4.1 were employed, both demonstrating strong performance. Additionally, a broader range of models, including ChatGPT-4o mini and DeepSeek V3, was explored. Notably, Ariadne achieved comparable performance when combined with DeepSeek V3. However, a significant performance drop was observed when using ChatGPT-4o mini, which lagged behind other methods. This indicates that Ariadne relies on the capabilities of more powerful base models.

**Cost Considerations** As shown in Table 1 and 2, Ariadne's multi-step collaborative approach incurs higher costs compared to other methods—particularly when using more expensive models—this issue should be viewed in the context of advancing model efficiency. In the long term, as large models continue to improve in performance and decrease in cost, a more favorable balance between performance and cost can be achieved.

**Corpus Size Trade-offs** Some RAG-based methods (Asai et al., 2024; Wang et al., 2024) construct extremely large retrieval databases, demanding heavy storage, computation, and maintenance. However, for most researchers, a small database covering major venues in their field, supplemented by legally obtained papers, is sufficient and more practical. Furthermore, paper retrieval is itself a mature area, and systems like PaSa (He et al., 2025) allow online filtering without tightly coupling retrieval with intelligent systems. Therefore, an overly large database was not constructed to validate the performance of Ariadne in the experiments.

**Future Work** As discussed in sections 3.2 and 3.3, future research should focus on two main directions. The first is to improve the F1 score in the retrieval and filtering stages. The second is to conduct a more in-depth analysis of the mechanisms underlying **Micro Review** and **Macro Review** to propose more effective review strategies.

## 5 Conclusion

This paper introduces **Ariadne**, a multi-agent collaborative system designed to support interactive literature analysis and research. Through the integration of expert guidance, detailed review, and efficient retrieval within a hierarchical multi-agent framework, Ariadne addresses key challenges in academic research support. Experimental results on established benchmarks demonstrate that Ari-

| Dimension | 5 (Excellent) | 4 (Good) | 3 (Average) | 2 (Poor) | 1 (Very Poor) |
|---|---|---|---|---|---|
| **Coverage** | Fully covers the task requirements, all key points included, comprehensive content | Covers most key points, only minor omissions | Covers some key points, but with noticeable gaps | Fragmented coverage, significant omissions | Fails to cover task requirements, lacks essential content |
| **Relevance** | Highly relevant to the topic, focused and on-point | Mostly relevant, with minor off-topic sections | Partially relevant, with noticeable digressions or redundant content | Mostly irrelevant, significant off-topic content | Completely off-topic, chaotic and irrelevant |
| **Structure** | Clear structure, well-organized, logical progression, smooth transitions | Generally well-structured, with occasional inconsistencies | Basic structure present, but lacks clarity and coherence | Poorly structured, disconnected ideas, lacks coherence | No recognizable structure, disorganized and chaotic |
| **Citation Quality** | Accurate, reliable, and clearly sourced citations, sufficient in number, fully supports arguments | Mostly accurate and clearly sourced citations, generally sufficient, but with minor omissions | Mixed quality, some accurate, some unverifiable, or insufficient in number | Mostly inaccurate or poorly sourced citations, clearly insufficient | Mostly inaccurate, fabricated, or misleading citations, almost entirely unsupported |
| **Critical Analysis** | Demonstrates deep analysis and balanced evaluation, identifies complex issues and offers insights | Shows some critical thinking, able to identify issues or weigh pros and cons | Superficial analysis, often relies on surface-level observations | Lacks independent analysis, mostly repetition or summary | No critical thinking, blindly accepts or oversimplifies information |

Table 5: **Rubric for subjective evaluation of response quality across five dimensions.**

adne achieves superior performance in both answer quality and citation accuracy compared to existing methods, highlighting its effectiveness in enhancing academic research workflows. These findings underscore the potential of multi-agent collaboration for advancing intelligent literature analysis and supporting future developments in automated scientific inquiry.

## 6 Related Work

**LLM for literature analysis** With the rapid advancement of natural language processing (NLP) technologies, particularly LLMs, significant progress has been made in automating various stages of scientific research workflows (Jiang et al., 2025). LLMs have proven effective in document processing tasks, including information retrieval, citation text generation, and paper review. For instance, PaperRobot (Wang et al., 2019) supports incremental draft generation, while (Xing et al., 2020) focuses on accurate citation text generation using pointer-generator networks. In addition, (Zimmermann et al., 2024) demonstrates the potential of LLMs in automating paper review writing. Beyond writing assistance, LLMs have been employed in the peer review process to generate explainable reviews based on synthesized knowledge from large volumes of scientific paper (Wang et al., 2020; Yu et al., 2024), and to identify errors for quality validation (Liu and Shah, 2023). Moreover, LLMs have been utilized for automated hypothesis generation by extracting key insights from extensive bodies of paper (Yang et al., 2024; Zeng et al., 2024).

**RAG-Based Methods for literature analysis** Another line of work employs RAG methods, such as AutoSurvey (Wang et al., 2024) and SurveyX (Liang et al., 2025), which plan an outline and retrieve literature to generate surveys on given topics. While these approaches offer broad overviews of research fields, they often produce lengthy outputs, require considerable time, and struggle to address specific queries or support interactive exploration. In contrast, PaperQA (Skarlinski et al., 2024) and OpenScholar (Asai et al., 2024) focus on retrieving literature and answering user queries. However, existing systems are typically limited to single-turn interactions, lack task decomposition, and heavily rely on vector retrieval, making them inadequate for exploratory research that demands iterative refinement and the handling of complex tasks.

## Limitations

**Dataset Limitations** The study's evaluation was constrained by the limited availability of publicly annotated datasets, preventing broader task coverage. Nevertheless, multiple validation rounds were conducted to ensure experimental stability.

**Human Evaluation** Human evaluation was confined to a single academic discipline due to time and resource constraints. However, the focus remained on participants' subjective experiences and objective assessments, as the SCHOLARQA-MULTI already established the methods' effectiveness across diverse disciplinary contexts as shown in Appendix C.

## Ethics Statement

This study involved human participants for evaluation. To ensure privacy, all personal information in the collected responses and related materials was anonymized. Data used for methodological analysis was included only with the explicit consent of the participants.

We used ChatGPT to assist with language polishing during the preparation of this manuscript. However, all conceptual development, analysis, and argumentation were carried out by the human authors.

## References


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

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf. Accessed: 2025-04-26.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Daniel S. Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *CoRR*, abs/2411.14199.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7765–7784. Association for Computational Linguistics.

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

Xue Jiang, Weiren Wang, Shaohan Tian, Hao Wang, Turab Lookman, and Yanjing Su. 2025. Applications of natural language processing and large language models in materials discovery. *npj Computational Materials*, 11(1):79.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. 2025. Surveyx: Academic survey automation via large language models. *CoRR*, abs/2502.14776.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *CoRR*, abs/2306.00622.

OpenAI. 2024. Chatgpt-4o (may 13 version). https://chat.openai.com. Large language model by OpenAI.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6252–6278. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela M. Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *CoRR*, abs/2409.13740.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019), Volume 1: Long Papers*, pages 1980–1991, Florence, Italy. Association for Computational Linguistics.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020, Dublin, Ireland, December 15-18, 2020*, pages 384–397. Association for Computational Linguistics.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Autosurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6181–6190, Online. Association for Computational Linguistics.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13545–13565. Association for Computational Linguistics.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024. Automated peer reviewing in paper SEA: standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10164–10184. Association for Computational Linguistics.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. In *Artificial Intelligence for Research and Democracy - First International Workshop, AI4Research 2024, and 4th International Workshop, DemocrAI 2024, Held in Conjunction with IJCAI 2024, Jeju, South Korea, August 5, 2024, Proceedings*, volume 14917 of *Lecture Notes in Computer Science*, pages 20–38. Springer.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Sciinstruct: a self-reflective instruction annotated dataset for training scientific language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Robert Zimmermann, Marina Staab, Mehran Nasseri, and Patrick Brandtner. 2024. Leveraging large language models for literature review tasks - a case study using chatgpt. In Teresa Guarda, Filipe Portela, and Jose Maria Diaz-Nafria, editors, *Advanced Research in Technologies, Information, Innovation and Sustainability*, pages 313–323. Springer Nature Switzerland, Cham.

## A Details of the Dataset

Figure 6 presents the distribution of cited references per question in SCHOLARQA-MULTI and SciFact. In SCHOLARQA-MULTI, most questions cite 3 to 7 references, with some citing up to 10. In contrast, SciFact questions predominantly cite a single reference. This highlights the multi-document nature of SCHOLARQA-MULTI compared to SciFact.

## B Criteria for SCHOLARQA-MULTI

Table 6 lists the assessment criteria adapted from Asai et al. (2024).

## C Subject Area Distribution and Per-Discipline Results

Figure 7 shows the distribution of questions across scientific domains in SCHOLARQA-MULTI.

Tables 7–12 report per-discipline results for six metrics: Organization, Coverage, Relevance, Precision, Recall, and F1. Results are shown for Raw, Naive-RAG, Paper-QA, and Ariadne. For each metric, the highest mean is bolded. These tables support detailed comparison across domains.

## D Implementation Details

### D.1 Workflow Pseudocode

The multi-agent workflow is summarized in Algorithm 1. Prompt templates are listed in Section D.2.

### D.2 Prompt Design

Ariadne adopts a modular prompt design, where each agent (Expert, Retrieval, MacroReview, and MicroReview) is equipped with specialized prompt templates that reflect its distinct function within the multi-agent system, as listed from Prompt 1-12.

We use publicly available code and data under the MIT License, with proper attribution to the original sources.

### D.3 Interface Design

Our system interface is implemented based on Gradio[2], a user-friendly and interactive web UI with Apache 2.0 License. Users can input academic questions, view multi-turn dialogue history, export chat records, and directly access cited references. The interface is designed for clarity and ease of use, supporting efficient literature analysis and interactive research. As shown in Figure 8, the main interface allows users to input questions and receive structured answers. Figure 9 demonstrates the multi-turn dialogue and literature citation display.

[2]https://github.com/gradio-app/gradio?tab=Apache-2.0-1-ov-file

## E Details of Human Evaluation

**Participant Demographics.** Participant demographics are summarized in Table 13. The participants were graduate students and visiting researchers with diverse levels of experience in educational technology research.

**Test Procedure.** Each participant interacted with three intelligent agent systems, each designed for different scenarios in educational technology research. Participants were instructed to ask questions based on their own research interests, such as exploring a research direction or querying a specific academic paper. For each system, participants conducted 5–10 rounds of dialogue.

To ensure consistent understanding of the task, participants were shown an instruction modal before beginning the evaluation (Figure 10). The modal clearly explained the purpose and structure of the test in both English and Chinese, including the number of interactions, anonymization policy, and the post-evaluation rating procedure.

After completing all dialogue sessions, participants were provided with the full conversation history for each system. They then rated the quality of the responses using the evaluation rubric described in Appendix B.

All dialogue records were anonymized during analysis and presentation to ensure that no personally identifiable information was disclosed.

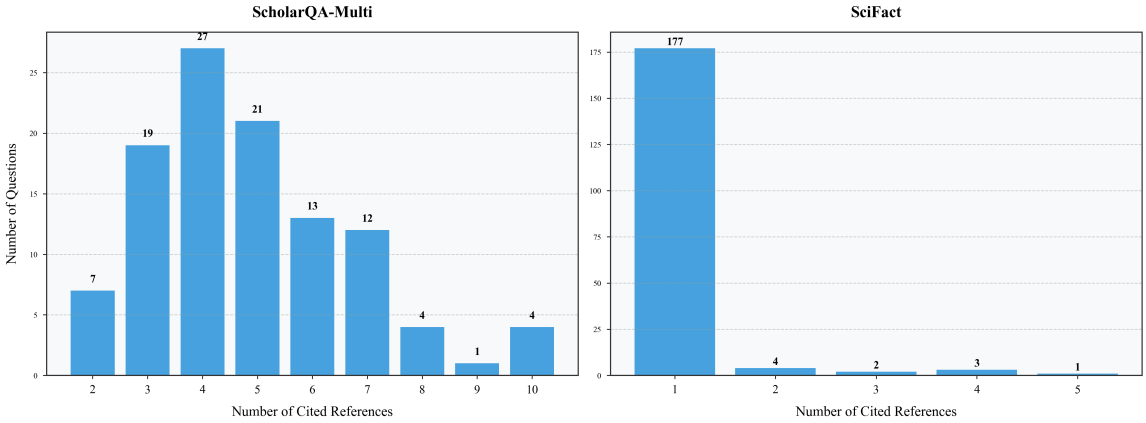


Figure 6: **Distribution of the number of cited references per question in SCHOLARQA-MULTI and SciFact.**

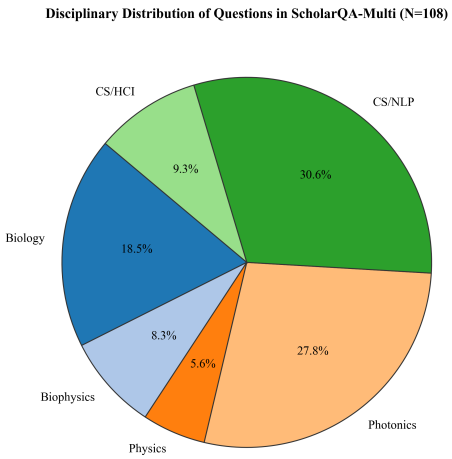


Figure 7: **Disciplinary distribution of questions in SCHOLARQA-MULTI.**

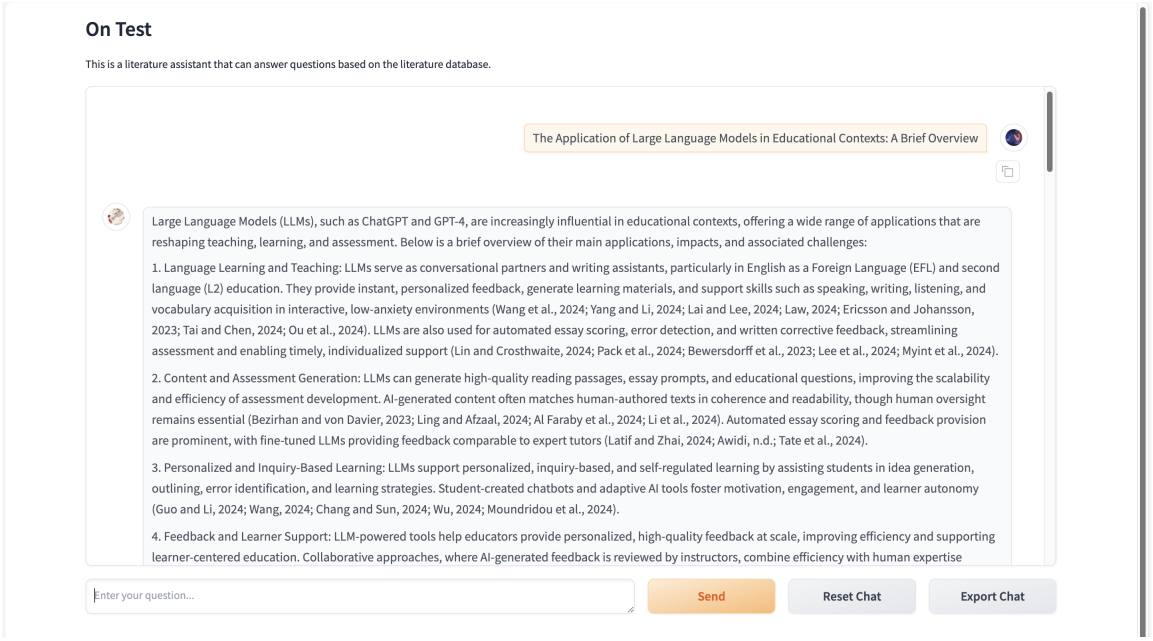


Figure 8: Main interface: users input academic questions and receive structured answers.

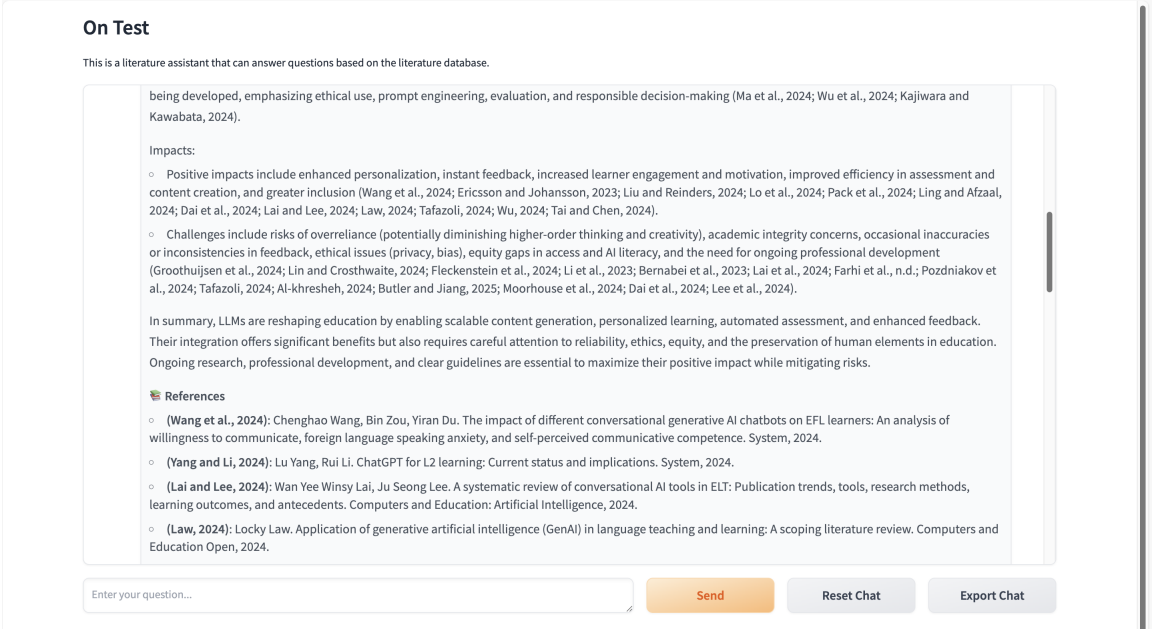


Figure 9: Interface example: multi-turn dialogue and literature citation display.

| Aspect | 5 | 4 | 3 | 2 | 1 |
|---|---|---|---|---|---|
| **Organization** | Exceptionally well-organized, flawless logical structure, points grouped perfectly, seamless flow, clear discourse markers or section headers, no contradictions or unnecessary repetition. | Well-organized, clear and logical structure, points effectively grouped, smooth flow, clear opening or section headers, minor lapses in coherence, minimal repetition or contradictions. | Generally well-organized, clear structure mostly maintained, points grouped logically, minor lapses in flow or coherence, mostly clear, occasional repetition or slight contradictions. | Some organization, inconsistent structure, occasional lapses in coherence, minor contradictions or repetitive statements disrupt clarity. | Disorganized, no clear structure, points scattered, difficult to follow, lacks coherence, contradictions or irrelevant repetitions throughout. |
| **Coverage** | Comprehensive and exceptional coverage, diverse range of papers and viewpoints, thorough overview, additional important discussion points, all necessary and sufficient information, no irrelevant details. | Good coverage, variety of representative papers and sources, broad overview, may miss a few minor areas, mostly sufficient information, avoids excessive irrelevant details, minor points could benefit from deeper exploration. | Acceptable coverage, several representative works, satisfactory overview, addresses core aspects, may miss some details, reasonable amount of relevant information, may lack some helpful details. | Partial coverage, covers some key aspects, misses significant lines of research, focuses too narrowly, lacks well-rounded view, limited information, leaves out important details. | Severely lacking coverage, focuses on a single line of work, misses holistic view, greatly limited depth, lacks essential details to understand the topic. |
| **Relevance** | Exceptionally focused and entirely on topic, tightly centered on the subject, enough depth and coverage, every piece of information contributes directly to understanding. | Mostly on-topic, clear focus, minor digressions or slightly irrelevant details, infrequent deviations, does not significantly undermine clarity or usefulness. | Somewhat on topic, several digressions or irrelevant information, frequent deviations, distract from the main question or redundant information. | Frequently off-topic, limited focus, addresses the question to some extent but often strays, several irrelevant or tangential points, difficult to maintain focus. | Off-topic, content significantly deviates from the question, difficult to discern relevance, distracts the user. |

Table 6: Assessment criteria for SCHOLARQA-MULTI across three core aspects.

| Metrics | Generation | | | Citation | | |
|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 |
| **Raw** | 3.78 ± 0.05 | 3.42 ± 0.09 | 3.73 ± 0.02 | 0.8 ± 1.2 | 0.6 ± 0.8 | 0.7 ± 0.9 |
| **Naive-RAG** | **3.85 ± 0.04** | 3.70 ± 0.18 | **3.92 ± 0.02** | 31.4 ± 2.6 | 38.3 ± 7.6 | 33.3 ± 4.3 |
| **Paper-QA** | 3.37 ± 0.06 | 3.00 ± 0.04 | 3.65 ± 0.07 | **70.0 ± 1.8** | 40.3 ± 2.6 | 48.7 ± 2.0 |
| **Ariadne** | **3.85 ± 0.04** | **3.78 ± 0.05** | 3.78 ± 0.08 | 63.9 ± 1.7 | **60.6 ± 0.6** | **60.0 ± 1.1** |

Table 7: Detailed results of different methods on bio.

| Metrics | Generation | | | Citation | | |
|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 |
| **Raw** | 3.44 ± 0.09 | 3.15 ± 0.05 | 3.63 ± 0.05 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |
| **Naive-RAG** | 3.74 ± 0.14 | 3.56 ± 0.24 | **3.85 ± 0.14** | 51.2 ± 8.1 | 35.5 ± 2.7 | 39.7 ± 2.2 |
| **Paper-QA** | 3.41 ± 0.05 | 3.19 ± 0.05 | 3.52 ± 0.05 | **100.0 ± 0.0** | 38.1 ± 2.3 | 53.3 ± 2.4 |
| **Ariadne** | **3.96 ± 0.05** | **3.74 ± 0.05** | 3.70 ± 0.14 | 96.1 ± 2.4 | **60.0 ± 3.0** | **72.4 ± 2.5** |

Table 8: Detailed results of different methods on biophysics.

| Metrics | Generation | | | Citation | | |
|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 |
| **Raw** | 3.83 ± 0.00 | 3.44 ± 0.21 | 3.72 ± 0.08 | 17.9 ± 2.4 | 17.4 ± 2.1 | 16.7 ± 1.9 |
| **Naive-RAG** | 3.78 ± 0.08 | 3.61 ± 0.31 | 3.67 ± 0.14 | 23.4 ± 12.6 | 31.2 ± 2.6 | 24.3 ± 7.1 |
| **Paper-QA** | 3.44 ± 0.08 | 3.06 ± 0.08 | 3.72 ± 0.08 | **56.5 ± 8.6** | 34.6 ± 1.2 | 40.4 ± 2.5 |
| **Ariadne** | **3.94 ± 0.08** | **3.83 ± 0.14** | **3.94 ± 0.08** | 53.6 ± 6.3 | **50.2 ± 1.2** | **49.9 ± 3.2** |

Table 9: Detailed results of different methods on physics.

| Metrics | Generation | | | Citation | | |
|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 |
| **Raw** | 3.56 ± 0.11 | 3.18 ± 0.08 | 3.66 ± 0.08 | 9.6 ± 1.3 | 6.1 ± 0.6 | 7.1 ± 0.1 |
| **Naive-RAG** | 3.78 ± 0.11 | 3.68 ± 0.29 | 3.86 ± 0.03 | 36.9 ± 5.4 | 36.8 ± 3.5 | 34.9 ± 0.6 |
| **Paper-QA** | 3.41 ± 0.06 | 3.18 ± 0.07 | 3.71 ± 0.04 | 50.2 ± 2.7 | 29.2 ± 1.5 | 35.2 ± 1.7 |
| **Ariadne** | **3.94 ± 0.03** | **3.88 ± 0.08** | **3.87 ± 0.03** | **53.8 ± 2.6** | **41.0 ± 3.1** | **44.0 ± 2.2** |

Table 10: Detailed results of different methods on photonics.

| Metrics | Generation | | | Citation | | |
|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 |
| **Raw** | 3.47 ± 0.04 | 3.06 ± 0.11 | 3.46 ± 0.04 | 7.3 ± 0.7 | 5.3 ± 0.8 | 6.0 ± 0.7 |
| **Naive-RAG** | 3.74 ± 0.05 | 3.42 ± 0.19 | 3.75 ± 0.08 | 46.4 ± 7.3 | 44.4 ± 2.8 | 41.5 ± 1.7 |
| **Paper-QA** | 3.49 ± 0.04 | 3.11 ± 0.08 | 3.52 ± 0.09 | 65.8 ± 0.4 | 33.8 ± 1.7 | 42.6 ± 1.2 |
| **Ariadne** | **3.87 ± 0.08** | **3.76 ± 0.13** | **3.77 ± 0.08** | **67.3 ± 6.8** | **52.9 ± 3.0** | **55.8 ± 3.4** |

Table 11: Detailed results of different methods on cs_nlp.

| Metrics | Generation | | | Citation | | |
|---|---|---|---|---|---|---|
| | Organization | Coverage | Relevance | Precision | Recall | F1 |
| **Raw** | 3.17 ± 0.12 | 3.07 ± 0.09 | 1.00 ± 0.00 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |
| **Naive-RAG** | 3.07 ± 0.21 | 2.83 ± 0.17 | 1.00 ± 0.00 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |
| **Paper-QA** | 3.00 ± 0.22 | 2.60 ± 0.08 | 1.00 ± 0.00 | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |
| **Ariadne** | **3.47 ± 0.26** | **3.20 ± 0.00** | **1.00 ± 0.00** | 0.0 ± 0.0 | 0.0 ± 0.0 | 0.0 ± 0.0 |

Table 12: Detailed results of different methods on cs_hci.

**Algorithm 1** Ariadne: Multi-Agent Collaboration Main Workflow

**Require:** User query $q$, conversation history $H$

**Ensure:** Structured academic response (JSON format)

**Task Allocation Phase:**
Expert Agent uses `ExpertPrompts.query_decomposition_prompt` to decompose $q$ into sub-questions $Q$; For each $q_i$, calls `ExpertPrompts.route_query_prompt` for task allocation, determining whether macro/micro review is needed and which paper IDs can be reused.

**for** each sub-question $q_i$ in $Q$ **do**

  **Review Phase (First Part):**
  Review Agent generates an adapted retrieval query $q_i^{\text{retr}}$ based on $q_i$ using `ReviewPrompts.generate_retrieval_query_prompt`;

  **Retrieval phase:**
  Retrieval Agent uses $q_i^{\text{retr}}$ and the specified mode $m_i$ to select retrieval strategy (exact/vector/hybrid/full) and retrieve relevant papers or content.

  **Filtering Phase:**
  For retrieved content, calls `LiteraturePrompts.filter_relevance_prompt` to evaluate relevance and filter highly relevant content; If there are multiple moderately relevant items, calls `LiteraturePrompts.select_best_moderate_prompt` for supplementary selection; For long text, calls `LiteraturePrompts.compress_content_prompt` to compress into concise paragraphs.

  **if** task is Macro Review **then**

    Review Agent calls `MacroReviewPrompts.analyze_papers_prompt` to analyze paper overviews and generate $r_i$;

    **if** number of papers is large **then**

      Process in batches, finally use `MacroReviewPrompts.integrate_batch_results_prompt` to integrate all batch results

    **end if**

  **else if** task is Micro Review **then**

    Review Agent calls `MicroReviewPrompts.answer_query_prompt` to analyze detailed content or full text and generate $r_i$

  **end if**

  Expert Agent synthesizes $a_i$ based on $r_i$ and dialogue context $H$ by calling `ExpertPrompts.synthesize_prompt`

**end for**

**Summary Phase:**
Expert Agent calls `ExpertPrompts.integrate_sub_responses_prompt` to integrate all $a_i$ and form the final answer $A$; Then calls `ExpertPrompts.general_response_prompt` for citation formatting and final output.

**return** Final structured response (with citations, JSON format)

Prompt 1: ExpertPrompts.route_query_prompt

```
Extract academic search queries from the user's question to find relevant academic
literature.

User Query: {query}

Conversation History:
{conversation_history}

Task:
Extract academic search query from the User Query. Focus only on the academic
content, ignoring any non-academic requirements. The query should be transformed
into a clear, academic question format that captures the core research interest.

- Choose EITHER macro_review OR micro_review based on the query's nature
- A query should not be classified as both - select the most appropriate category

Return in JSON format:
{
   "query": "",  # Academic search query in question form, transformed to capture
    the core research interest
    "use_macro_review": true/false,  # If query involves research trends or
    categories
   "use_micro_review": true/false,  # If query involves specific facts or methods
   "macro_context_paper_ids": [],  # Relevant paper IDs from history for context,
    if can be used for macro review
   "micro_direct_analysis": true/false,  # If query requires to analyze specific
    papers from history
    "micro_direct_paper_ids": []  # Paper IDs from history to analyze directly,
    which is required by query
}

Return only the JSON format result without any other explanation.
```

Prompt 2: ExpertPrompts.integrate_results_prompt

```
Based on the user's query, conversation history, and search results, generate a
comprehensive response.

User Query:
{user_query}

Conversation History:
{conversation_history}

Macro Literature Analysis Results:
{macro_results}

Micro Literature Analysis Results:
{micro_results}

Please generate a complete and coherent response that:
1. Directly addresses the user's question based on available information
2. References information from previous conversation turns when relevant
3. Appropriately integrates both macro and micro analysis results
4. Organizes information in a clear, logical structure
5. If any analysis contains error messages (e.g. "No relevant papers found" or
"Unable to determine relevant papers"):
   - Still provide any useful information from successful analyses
   - Naturally incorporate questions for additional information that would help
   provide better results

Citation Rules:
1. In-text citation format:
   - Single author: (Smith, 2023)
   - Two authors: (Smith and Brown, 2023)
   - Three or more authors: (Smith et al., 2023)
2. For multiple papers by the same author(s) in the same year, add letters
(a, b, c...)
   Example: (Smith, 2023a), (Smith, 2023b)
3. Full citation format in citations_used:
   - Keep maximum THREE authors in the full citation, followed by "et al."
   - Format: (First author et al., Year):
   First Author, Second Author, Third Author, et al. (Year). Title.
   Example:
   (Schulz et al., 2023): Schulz, A., Stathatos, S., Shriver, C., et al. (2023).
  Utilizing online and open-source machine learning toolkits to leverage the future
   of sustainable engineering.

Return in JSON format:
{
    "answer": "Complete response including both available information and any
   necessary follow-up questions. The response should be well-structured with clear
    introduction, logical flow of ideas, and concise conclusion.",
    "citations_used": {
        "(Citation Key)": ["Full citation details", "Paper ID"],
        // Include all citations used in the answer
    }
}
Return only the JSON format result without any other explanation.
```

Prompt 3: ExpertPrompts.general_response_prompt

```
Based on the conversation history and current query, generate a comprehensive
response.

Current Query: {query}

Conversation History:
{conversation_history}

Please generate a complete and coherent response that:
1. Directly addresses the user's question based on available information
2. References information from previous conversation turns when relevant
3. Organizes information in a clear, logical structure

Citation Rules:
1. In-text citation format:
   - Single author: (Smith, 2023)
   - Two authors: (Smith and Brown, 2023)
   - Three or more authors: (Smith et al., 2023)
2. For multiple papers by the same author(s) in the same year, add letters
(a, b, c...)
   Example: (Smith, 2023a), (Smith, 2023b)
3. Full citation format in citations_used:
   - Keep maximum THREE authors in the full citation, followed by "et al."
   - Format: (First author et al., Year):
   First Author, Second Author, Third Author, et al. (Year). Title.
   Example:
   (Schulz et al., 2023): Schulz, A., Stathatos, S., Shriver, C., et al. (2023).
  Utilizing online and open-source machine learning toolkits to leverage the future
   of sustainable engineering.

Return in JSON format:
{
    "answer": "Complete response including both available information and any
   necessary follow-up questions. The response should be well-structured with clear
    introduction, logical flow of ideas, and concise conclusion.",
    "citations_used": {
        "(Citation Key)": ["Full citation details", "Paper ID"],
        // Include all citations used in the response
    }
}

Return only the JSON format result without any other explanation.
```

Prompt 4: ExpertPrompts.query_decomposition_prompt

```
Analyze the user query. If it contains multiple independent questions, split them
into separate sub-queries. If it's a single question, keep it as is.

User Query: {query}

Conversation History:
{conversation_history}

Rules:
1. Only split when the query contains multiple independent questions
2. Keep the exact original expression of each question, do not modify any wording
3. If it's a single question, return the original question
4. Maximum 3 sub-queries

Examples:
Input: "How has neural architecture search evolved for efficient transformers?"
Output: {
    "sub-query": [
       "How has neural architecture search evolved for efficient transformers?"
    ]
}

Input: "I'm studying Zhang's 2023 paper on transformer efficiency. Could you
explain their approach to reducing computational complexity? Also, how does their
method compare with previous work, and what are the main limitations they found in
experiments?"
Output: {
    "sub-query": [
     "I'm studying Zhang's 2023 paper on transformer efficiency. Could you explain
        their approach to reducing computational complexity?",
     "Also, how does their method compare with previous work, and what are the main
        limitations they found in experiments?"
    ]
}

Return in JSON format:
{
    "sub-query": ["question1", "question2", "question3"]
}

Return only the JSON format result without any other explanation.
```

Prompt 5: ExpertPrompts.integrate_sub_responses_prompt

```
Based on the original query and its Conversation History, generate a comprehensive
answer.

Original Query:
{original_query}

Conversation History:
{conversation_history}

Please generate a complete and coherent response that:
1. Directly addresses the user's question based on available information
2. References information from previous conversation turns when relevant
3. Appropriately integrates both macro and micro analysis results
4. Organizes information in a clear, logical structure
5. If any analysis contains error messages (e.g. "No relevant papers found" or
"Unable to determine relevant papers"):
   - Still provide any useful information from successful analyses
   - Naturally incorporate questions for additional information that would help
   provide better results

Citation Rules:
1. In-text citation format:
   - Single author: (Smith, 2023)
   - Two authors: (Smith and Brown, 2023)
   - Three or more authors: (Smith et al., 2023)
2. For multiple papers by the same author(s) in the same year, add letters
(a, b, c...)
   Example: (Smith, 2023a), (Smith, 2023b)
3. Full citation format in citations_used:
   - Keep maximum THREE authors in the full citation, followed by "et al."
   - Format: (First author et al., Year):
   First Author, Second Author, Third Author, et al. (Year). Title.
   Example:
   (Schulz et al., 2023): Schulz, A., Stathatos, S., Shriver, C., et al. (2023).
  Utilizing online and open-source machine learning toolkits to leverage the future
   of sustainable engineering.

Return in JSON format:
{
   "answer": "Complete response integrating all sub-answers with citations in the
    context",
    "citations_used": {
        "(Citation Key)": ["Full citation details", "Paper ID"],
        // Include all citations used in the answer
    }
}

Return only the JSON format result without any other explanation.
```

Prompt 6: LiteraturePrompts.search_analysis_prompt

```
Analyze the following literature search query to determine the most appropriate
SEARCH MECHANISM that ensures COMPLETE and ACCURATE paper retrieval.

CRITICAL: Your primary goal is to determine HOW to retrieve papers, NOT how to
analyze them.
- Focus on paper retrieval completeness and accuracy
- Choose the search mechanism that ensures no relevant papers are missed
- The actual analysis of paper content will be handled separately by review agents

Query: {query}

STEP 1: DETERMINE SEARCH TYPE
Choose the most appropriate search type for RETRIEVING papers:

- "full": When the query indicates a need for ALL papers in the database
  * Choose this when completeness is required AND there is no clear analysis focus
  * Examples: "all papers", "entire database", "every paper", "show all",
  "analyze all papers"
  * This ensures NO papers are missed
  * If the query has a specific analysis focus (e.g. "analyze all papers about
  deep learning"), use "vector" or "hybrid" instead
  * When in doubt and there's no clear analysis focus, choose "full" to ensure
  completeness

- "exact": When papers can be found using precise matching criteria
  * Use when query contains specific identifiers
  * Examples: author names, years, exact titles
  * Example: "Find papers by John Smith from 2023"

- "vector": When papers need to be found based on topic similarity
  * Use for topic-based searches without exact criteria
  * Example: "Find papers about deep learning applications"

- "hybrid": When both exact matching and topic similarity are needed
  * Combines exact and vector search
  * Example: "Find John Smith's papers about deep learning"

STEP 2: CREATE SEARCH QUERY
(Skip this step if search_type is "full")

For vector search, create a specific query text that DIRECTLY addresses the user's original query:

CRITICAL: The generated query text MUST:
- Be SPECIFICALLY designed to help answer the user's original query
- Use key terms and concepts from the original query
- Maintain the same intent and focus as the original query
- Be detailed enough to capture the semantic meaning of the search intent

Example:
If original query is "How does gamification affect student motivation?":
{
    "vector_query_text": "Research on gamification effects and impact on student motivation and
    engagement in education, including methods, implementations and results"
}

IMPORTANT RULES:
1. Keep query text focused and specific YET DIRECTLY RELATED to the original query
2. Include all relevant aspects of the search intent in a single comprehensive query
3. ALWAYS ensure the generated query text helps find papers that answer the user's specific question

RESPONSE FORMAT:
{
    "search_type": "exact|vector|hybrid|full",
    "exact_criteria": {
      "authors": ["exact author names"],  // Only when query explicitly mentions specific authors
        "year": "specific year",  // Only when query explicitly mentions specific year
        "title": "specific title"  // Only when query explicitly mentions specific title
    },
   "vector_query_text": "comprehensive query text for finding relevant papers" // Single string for
    vector search
}

Return only valid JSON without any additional text in English.
```

Prompt 7: LiteraturePrompts.filter_relevance_prompt

```
Please evaluate how relevant this content is to the query.

Query: {query}

Content:
{content}

Evaluate the relevance on a scale of 0-5:
5 - Perfectly relevant: The content directly and comprehensively answers the query
4 - Highly relevant: The content directly answers most aspects of the query
3 - Moderately relevant: The content contains helpful information that partially
answers the query
2 - Somewhat relevant: The content has some related information but doesn't directly
answer the query
1 - Marginally relevant: The content has only tangential or contextual relevance
0 - Not relevant: The content does not contain helpful information for the query

Note: If the content is clearly the reference section of a paper, return 0.

Return in JSON format:
{
    "relevance_score": 0,  // Score from 0-5
}
```

Prompt 8: LiteraturePrompts.select_best_moderate_prompt

```
Given a research query and a list of moderately relevant items, select the most
suitable items that best complement the highly relevant results.

Query: {query}

Number of slots to fill: {remaining_slots}

Available items:
{items_text}

Return your selection as a JSON object with this format:
{
    "selected_indices": [0, 2, 5]  // List of selected item indices, maximum
    {remaining_slots} items
}

Note: Only return indices of items that would be truly helpful in answering the
query. You don't need to use all available slots if fewer items would suffice.
```

Prompt 9: LiteraturePrompts.compress_content_prompt

```
Compress the following research content to approximately 150 words while
maintaining the most relevant information to the query.

Query: {query}

Content to compress:
{content}

Requirements:
1. Focus on information related to the query
2. Target length: ~150 words

Return in JSON format:
{
    "compressed_content": "The compressed text here..."
}
```

Prompt 10: MacroReviewPrompts.analyze_papers_prompt

```
Based on the literature provided below, please answer the following research
question:

Question: {query}

Relevant Literature:
{papers_text}

Write a focused, well-structured answer that directly addresses the question.
Synthesize only the relevant insights from multiple papers, compare approaches
when appropriate, and support your points with specific details.
Avoid summarizing papers unless it helps answer the question. Use a scholarly
tone, cite sources as (Author, Year), and list only the citations actually
used (with citation and paper_id) in the citations_used field.
The goal is to provide a clear, helpful answer—not to review the literature.

Return your response as a JSON object with the following structure:
{
    "answer": "Your comprehensive answer here, including citations in the text.
   The response should be well-structured with clear introduction, logical flow of
    ideas, and concise conclusion.",
    "citations_used": {
      "(Smith et al., 2023)": ["Full citation for Smith et al., 2023 in Relevant
       Literature", "paper_id for Smith et al., 2023 in Relevant Literature"],
     "(Johnson, 2020)": ["Full citation for Johnson, 2020 in Relevant Literature",
        "paper_id for Johnson, 2020 in Relevant Literature"]
    }
}

Return only the JSON object without any additional explanations.
```

Prompt 11: MacroReviewPrompts.integrate_batch_results_prompt

```
I will provide you with multiple batches of paper analysis results. Please integrate
these results into a comprehensive summary.

Original Question: {query}

Analysis Results from Multiple Batches:
{batch_results}

Please integrate these analysis results into a complete response. Your integration
should:
1. Avoid redundant information
2. Maintain a coherent narrative that directly addresses the original question
3. Preserve all relevant citations and evidence
4. Combine similar findings while maintaining specificity
5. Ensure the integrated response is comprehensive yet detailed

Return your response in the same JSON format as the input:
{
   "answer": "Your integrated comprehensive answer here, including citations in the
    text",
    "citations_used": {
        "(Author et al., Year)": ["citation", "paper_id"],
        ...
    }
}

Return only the JSON object without any additional explanations.
```

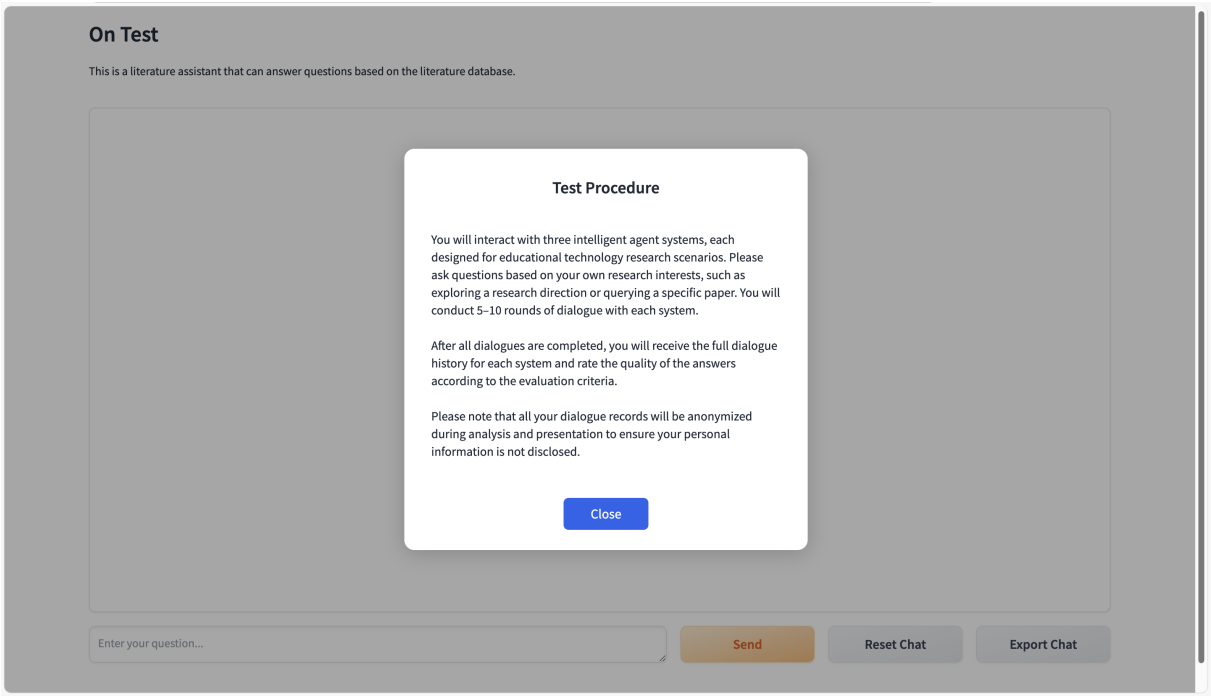


Figure 10: Screenshot of the instruction modal shown at the beginning of the test.

Prompt 12: MicroReviewPrompts.answer_query_prompt

```
Based on the literature provided below, please answer the following research
question:

Question: {query}

Relevant Literature:
{paper_content}

Write a focused, well-structured answer that directly addresses the question.
Synthesize only the relevant insights from multiple papers, compare approaches
when appropriate, and support your points with specific details.
Avoid summarizing papers unless it helps answer the question. Use a scholarly tone,
cite sources as (Author, Year), and list only the citations actually used (with
citation and paper_id) in the citations_used field.
The goal is to provide a clear, helpful answer—not to review the literature.

Return your response as a JSON object with the following structure:
{
   "answer": "Your comprehensive answer here, including citations in the text. The
    response should be well-structured with clear introduction, logical flow of
    ideas, and concise conclusion.",
    "citations_used": {
       "(Smith et al., 2023)": ["citation for Smith et al., 2023", "paper_id for
        Smith et al., 2023"],
       "(Johnson, 2020)": ["citation for Johnson, 2020", "paper_id for Johnson,
        2020"]
    }
}

Note: You can only use the citation from Relevant Literature itself which marked
with citation and paper_id, don't refer to the citation in the "content"!

Return only the JSON object without any additional explanations.
```

| Participant Type | Count |
|---|---|
| Master's (Year 1) | 3 |
| Master's (Year 3) | 1 |
| PhD (Year 1) | 1 |
| PhD (Year 2) | 2 |
| PhD (Year 3) | 2 |
| Visiting Scholar | 1 |

Table 13: Demographics of human evaluation participants.