Do We Need Bigger Models for Science? Task-Aware Retrieval with Small Language Models

Florian Kelber[1], **Matthias Jobst**[1], **Yuni Susanti**[2], **Michael Färber**[1]
[1]ScaDS.AI, TU Dresden, Germany
[2]FIZ Karlsruhe, Germany
florian.kelber@student.tu-dresden.de, matthias.jobst@student.tu-dresden.de,
yuni.susanti@fiz-karlsruhe.de, michael.faerber@tu-dresden.de

Abstract

Scientific knowledge discovery increasingly relies on large language models, yet many existing scholarly assistants depend on proprietary systems with tens or hundreds of billions of parameters. Such reliance limits reproducibility and accessibility for the research community. In this work, we ask a simple question: *do we need bigger models for scientific applications?* Specifically, we investigate to what extent carefully designed retrieval pipelines can compensate for reduced model scale in scientific applications. We design a lightweight retrieval-augmented framework that performs *task-aware* routing to select specialized retrieval strategies based on the input query. The system further integrates evidence from full-text scientific papers and structured scholarly metadata, and employs compact instruction-tuned language models to generate responses with citations. We evaluate the framework across several scholarly tasks, focusing on scholarly question answering (QA), including single- and multi-document scenarios, as well as biomedical QA under domain shift and scientific text compression. Our findings demonstrate that retrieval and model scale are complementary rather than interchangeable. While retrieval design can partially compensate for smaller models, model capacity remains important for complex reasoning tasks. This work highlights retrieval and task-aware design as key factors for building practical and reproducible scholarly assistants.

Keywords: scientific question answering, retrieval-augmented generation, small language models

1. Introduction

The volume of scientific publications continues to grow rapidly, making it increasingly difficult for researchers to discover and synthesize relevant knowledge. Recent advances in large language models (LLMs) have shown strong potential for supporting scientific tasks such as question answering, summarization, and literature exploration. However, many scholarly assistants rely on proprietary models with tens or hundreds of billions of parameters, creating substantial barriers in terms of computational cost, accessibility, and reproducibility.

Beyond issues of scale, applying general-purpose LLMs to scientific literature presents additional challenges. Scientific documents are highly technical and domain-specific, and models may lack sufficient adaptation to scholarly language and reasoning patterns. As a result, existing systems can produce hallucinated or weakly supported claims when answering scientific questions in research papers (Wadden et al., 2025; Cai et al., 2025; Li et al., 2025; Shen et al., 2024). These limitations raise concerns about reliability and transparency, which are central to scientific applications.

Early efforts to build science-focused language models highlight both the promise and the limitations of current approaches. Systems such as Galactica attempted to encode large amounts of scientific knowledge directly into model parameters, but encountered challenges related to factual accuracy and verification (Taylor et al., 2022; Marcus, 2022). Other approaches incorporate retrieval mechanisms or domain-specific datasets, yet many still depend on large proprietary backbones or remain limited to specific domains. For example, biomedical QA systems built around datasets such as PubMedQA (Jin et al., 2019) demonstrate the benefits of targeted retrieval, but do not necessarily generalize across the broader scientific ecosystem.

Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm for improving reliability and transparency by grounding model outputs in external evidence (Lewis et al., 2020). In scientific domain, prior work has explored knowledge-graph-based retrieval, domain-specific training, and hybrid approaches that combine multiple sources of evidence (Jaradeh et al., 2020; Asai et al., 2024). While these approaches improve grounding, hallucinations remain a persistent challenge (Mishra et al., 2024; Mallen et al., 2023), and many systems continue to rely on large backbone models.

In this work, we examine a simple question: *do we need bigger models for scientific applications?* Specifically, we investigate the extent to which improvements in retrieval design can support the use of smaller models in these settings. Rather than asking whether retrieval can replace model scaling entirely, we focus on the conditions under which retrieval strategies can support smaller models effectively, and the trade-offs that arise in doing so. In particular, we analyze how retrieval, domain cover-

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 108–118
12 May 2026.

age, and task structure interact with model capacity in scholarly question answering and related tasks.

To this end, we design a lightweight retrieval-augmented framework that incorporates *task-aware routing* prior to retrieval. Incoming queries are first classified into scholarly task categories, allowing the system to select specialized retrieval strategies tailored to different information needs. The framework integrates evidence from both unstructured scientific documents and structured scholarly metadata from knowledge graph within a unified pipeline. Specifically, we combine a compact 3B-parameter instruction-tuned language model with retrieval over a corpus of 165K research papers (unarXive (Saier et al., 2023)) and a large scholarly knowledge graph (SemOpenAlex (Färber et al., 2023)). This hybrid design enables support for multiple tasks, including scholarly question answering, summarization, and factual metadata queries.

We evaluate the framework across several scholarly tasks, focusing on scholarly QA (ScholarQABench-Multi (Asai et al., 2024))–including both single- and multi-document scenarios–as well as biomedical QA under domain shift (PubMedQA (Jin et al., 2019)) and scientific text compression (SciTLDR (Cachola et al., 2020)). Our results show that small open-weight models can approach the performance of larger systems when paired with appropriate retrieval strategies. However, we also found that system performance remains sensitive to retrieval quality, prompt length, and domain mismatch, highlighting limitations in robustness and generalization. Overall, our findings suggest that improved retrieval design can partially compensate for smaller models. Nevertheless, model capacity remains important for complex reasoning tasks, which means that retrieval and model scale are complementary rather than interchangeable.

Our contributions are summarized as follows:

- We design a task-aware routing strategy that selects retrieval methods based on the information needs of scholarly queries.
- We introduce a hybrid retrieval framework integrating scientific text collections with structured scholarly knowledge graphs.
- We empirically analyze the extent to which compact open-weight models, combined with targeted retrieval, can approach the performance of larger systems, highlighting key trade-offs in robustness and generalization.
- We release resources and source code to support reproducibility and further research.[1]

[1] https://github.com/faerber-lab/lightweight-scholarly-qa

2. Related Work

Retrieval-Augmented Systems for Scholarly QA Retrieval-Augmented Generation (RAG) has become a dominant paradigm for improving factual grounding in language models (Lewis et al., 2020). In the scholarly domain, several systems combine large language models with retrieval from scientific corpora. For example, OpenScholar retrieves and reranks paper sections from Semantic Scholar, incorporating feedback loops and citation verification mechanisms to improve reliability (Asai et al., 2024). Large-scale resources such as unarXive (Saier et al., 2023) and SciQA (Auer et al., 2023) provide curated datasets for scientific question answering and retrieval-based experimentation.

While these approaches demonstrate the effectiveness of retrieval in scientific settings, many rely on heavyweight reranking pipelines or large backbone models. In contrast, our work investigates whether a lightweight architecture can achieve competitive performance without complex reranking or large-scale parametric knowledge. Rather than scaling model size, we focus on structuring the retrieval process through task-aware routing.

Knowledge Graph-Based Scholarly Information Access Knowledge graphs provide a complementary perspective on scholarly knowledge by representing publications, authors, and venues as structured entities and relations. Several large-scale scholarly KGs have been introduced, including SemOpenAlex (Färber et al., 2023), Semantic Scholar (Kinney et al., 2023), ORKG (Auer et al., 2020), and MAKG (Färber and Ao, 2022). These resources enable structured querying through languages such as SPARQL and support applications including metadata exploration and scientific discovery. Prior work has investigated machine-learning-driven interfaces that map natural language queries to KG queries. Our approach builds on this line of work by integrating KG retrieval within a broader retrieval-augmented generation framework. Rather than focusing exclusively on KG-based QA, we combine structured metadata with text-based evidence to support a wider range of scholarly tasks.

Plain Language Summarization of Scientific Literature Another important line of research focuses on improving the accessibility of scientific documents through plain language summarization (PLS). These systems aim to condense complex texts into clear summaries that can be understood by both expert and non-expert audiences (August et al., 2024). Techniques often involve interpreting domain-specific terminology, adding explanatory context, and removing redundant or overly technical details (Guo et al., 2024). Maintaining fac-

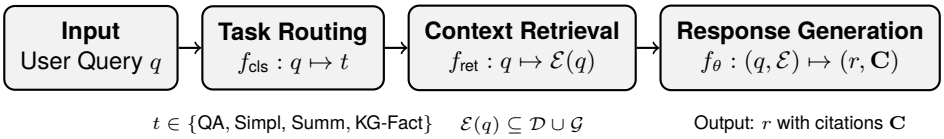

Figure 1: Task-aware retrieval pipeline. Task routing determines which processing strategy and data source are used before response generation. The input query q is classified into a task t, used to retrieve relevant context $\mathcal{E}(q)$ from data sources $\mathcal{D}$ and $\mathcal{G}$, and passed to a lightweight LLM to generate the response r with citations $\mathbf{C}$.

tual accuracy while improving readability remains challenging, particularly for specialized domains such as biomedical literature (Joseph et al., 2024). Evaluation is also non-trivial: traditional metrics such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2020) capture lexical similarity but often fail to fully measure informativeness or faithfulness (Guo et al., 2024; Luo et al., 2022; Ondov et al., 2022). To evaluate summarization capabilities within our framework, we use the SciTLDR dataset (Cachola et al., 2020), which focuses on concise summaries of computer science publications and has been widely used in prior work (Takeshita et al., 2024). Earlier experiments using environments such as CATTS extended these tasks with title generation and denoising objectives (Cachola et al., 2020).

Despite advances in retrieval and summarization, many scholarly assistants continue to rely on large proprietary models. This dependence limits reproducibility and raises the barrier to entry for academic institutions without extensive computational resources. Recent studies have highlighted both the benefits and limitations of adapting general-purpose LLMs to scientific domains (Wadden et al., 2025; Cai et al., 2025; Li et al., 2025; Shen et al., 2024), as well as the persistence of hallucinations in scholarly settings (Mishra et al., 2024; Mallen et al., 2023).

Positioning of this Work. Overall, prior work demonstrates the value of retrieval-augmented generation, domain-specific datasets, and knowledge graphs for scholarly question answering. However, fewer works investigate how these components interact within a lightweight architecture. While prior work often emphasizes scaling model size, fewer studies systematically analyze the trade-offs between retrieval design and model capacity. Our work addresses these gaps in two key ways. First, we explicitly examine whether improved retrieval design can compensate for a reduced model scale. Second, we integrate task-aware routing and hybrid retrieval (text + structured scholarly metadata) within a lightweight pipeline. We next describe the proposed framework in detail.

3. Task-Aware Hybrid Retrieval for Scholarly Applications

Scholarly assistants must handle heterogeneous information needs, ranging from factual metadata queries to multi-document reasoning over scientific literature. Treating these requests uniformly often leads to inefficient retrieval and unnecessary load on large language models.

We design a task-aware retrieval framework that routes user queries to specialized retrieval strategies before generation. The framework combines (i) task classification and routing with a lightweight classifier, (ii) hybrid context evidence retrieval from both scientific text collections and knowledge graphs, and (iii) response generation using a small language model. Figure 1 illustrates the overall architecture of the proposed framework.

Design Rationale Our design is guided by two observations. First, many scholarly queries can be solved through targeted retrieval rather than increased model scale. For example, a question "Which papers propose methods for protein structure prediction?" can be answered by retrieving relevant abstracts or datasets instead of relying on a massive language model to memorize all literature. Second, queries about scholarly metadata are better answered through knowledge graphs or structured database than through text generation alone. For instance, a question "Who are the co-authors of Zang et al., 2023?" is more accurately answered by querying structured bibliographic databases or knowledge graphs. These observations motivate the combination of task-aware routing and hybrid retrieval within a lightweight generation framework.

Formally, given a query $q \in \mathcal{Q}$, the system retrieves evidence

$$\mathcal{E}(q) = \{e_1, ..., e_M\}$$

from a textual corpus $\mathcal{D}$ and/or scholarly knowledge graph $\mathcal{G}$. A lightweight language model then produces an answer

$$r = f_\theta(q, \mathcal{E}(q)).$$

Each answer is accompanied by a citation set

$$\mathbf{C} = \{c_1, ..., c_K\}, \quad c_i \in \mathcal{D} \cup \mathcal{G},$$

allowing the system to expose the evidence underlying generated statements.

3.1. Task Routing

User queries in scholarly assistants are highly heterogeneous. A request for a paper summary requires different evidence than a factual query asking about an author's affiliation. Applying a single retrieval strategy to all queries therefore introduces unnecessary noise and computational cost. We introduce a lightweight task routing module that predicts the information need prior to retrieval. Given a query q, a classifier assigns a task label

$$f_{\text{cls}} : q \mapsto t \in \mathcal{T}$$

where

$$\mathcal{T} = \{\texttt{General QA}, \texttt{Simplification}, \texttt{Summarization}, \texttt{KG-Fact}\}$$

The four task categories reflect distinct information needs that require different retrieval strategies. General QA involves multi-document reasoning over unstructured text, while summarization and simplification require transformation of specific documents. In contrast, KG-Fact queries target structured metadata that can be answered more reliably via knowledge graphs than text retrieval. This taxonomy is not intended to be exhaustive but rather to capture common patterns in scholarly information-seeking. Future work could explore learned or hierarchical task taxonomies.

We implement the classifier using a fine-tuned small language model (*Llama 3.2 3B Instruct*). If classifier confidence falls below a predefined threshold, the system defaults to the general QA retrieval pathway. The predicted task t determines the retrieval strategy used in the next stage, as summarized in Table 1. This routing mechanism reduces unnecessary retrieval noise and allows the system to allocate resources more efficiently.

3.2. Context Evidence Retrieval

Traditional RAG systems typically rely on unstructured text corpora. However, scholarly information exists both in textual form and in structured metadata sources such as citation graphs and author databases. Unlike conventional RAG pipelines relying solely on unstructured text corpora, our framework integrates both unstructured scientific documents and structured scholarly knowledge graphs. This hybrid design enables the system to answer both narrative questions and precise metadata queries within a unified architecture.

Once a query has been classified into a specific task, the system retrieves task-appropriate *evidence* from curated text corpora or structured knowledge graphs. The context retrieval step thus serves as the bridge between user intent and generation, ensuring that responses are grounded in identifiable sources. Depending on the task category, one of the following retrieval strategies is invoked, as summarized in Table 1.

QA: General Scholarly Question Answering. For general QA tasks, text passages are retrieved as evidence from the open subset of the *unarXive* corpus (Saier et al., 2023), comprising 165K publications. Documents are preprocessed into *markdown* containing metadata and plain text (tables and figures removed). The text is segmented into sections or subsections, further subdivided into chunks of at least 800 characters. Each chunk is embedded using Sentence Transformers (*all-MiniLM-L6-v2*) into 384-dimensional vectors. Both vector embeddings and metadata (e.g., *title, authors, venue*) are stored in a FAISS vector index (Douze et al., 2024; Chase, 2022). At query time, the user query q is embedded in the same space, and the top-k most similar chunks based on vector similarity are retrieved:

$$\mathcal{E}_{\text{QA}}(q) = \text{Top-}k\big(\text{FAISS}(\text{Embed}(q))\big).$$

Retrieved chunks are assigned reference numbers and appended to the user query, following the multi-paper QA prompt design from the OpenScholar (Asai et al., 2024). To enrich bibliographic accuracy, the pipeline also issues secondary lookups to the Semantic Scholar API (Kinney et al., 2023) using paper titles from the retrieved set. References are tracked for the generation of a corresponding bibliographic list.

Simplification and Summarization For simplification and summarization tasks, the system attempts to detect whether the user refers to a specific scientific paper. A Named Entity Recognizer (NER) component is used to identify candidate paper titles, and if a specific title is detected, a fuzzy match is performed against the corpus (*unarXive* corpus (Saier et al., 2023)) to retrieve the full text of the corresponding paper. If no reliable title match is found, the system assumes that the query refers to arbitrary input text rather than a specific paper and performs summarization or simplification directly on the provided input. This fallback ensures that the system remains functional even when grounding is not possible.

The NER component is implemented using a spaCy *EntityRecognizer*[2] trained on synthetic data generated from unarXive titles and manually designed prompt templates. This step enables the system to ground summaries in the full paper when possible, while still supporting general text queries.

[2] https://spacy.io/api/entityrecognizer

Task Category	Retrieval Strategy	Output Type
General QA	Passage retrieval from unarXive; metadata enrichment via Semantic Scholar API	Factual answer with inline citations
Simplification, Summarization	NER-based paper title detection; fuzzy match against unarXive full-text corpus	Simplified or summarized text with optional paper grounding
KG-Fact	SPARQL queries over SemOpenAlex; 18 predefined templates (author/work)	Structured metadata (e.g., h-index, DOI, ORCID, affiliations)

Table 1: Overview of task categories, retrieval strategies, and output in the classification module.

KG-Fact: Structured Scholarly Metadata Retrieval. Certain scholarly queries request structured metadata such as citation metrics, identifiers, or affiliations. These questions are better addressed through knowledge graphs rather than through text generation alone. For such queries, the system maps user questions to one of 18 predefined SPARQL templates, focusing on author and scientific work-related queries. A rule-based precheck detects explicit mentions of scholarly identifiers such as *h-index*, *i10-index*, *ORCID*, or *DOI*. If such identifiers are not present, an NER component extracts candidate author or paper entities, and a classifier assigns the query to the most appropriate SPARQL template. The resulting query is executed against the SemOpenAlex knowledge graph via its SPARQL endpoint, returning structured metadata $\mathcal{E}_{\text{KG}}(q)$ in a verifiable format.

Our template-driven approach to knowledge graphs provides a reliable and interpretable interface for knowledge graph access. The current implementation uses a fixed set of templates, but these can be readily extended as new query types are added.

Final Prompt Composition

Across all pathways, context retrieval produces an evidence set $\mathcal{E}(q)$ tailored to the classified task t. Formally:

$$\mathcal{E}(q,t) = \begin{cases} \mathcal{E}_{\text{QA}}(q) & t = \textit{QA}, \\ \mathcal{E}_{\text{Simpl}}(q) & t = \textit{Simpl}, \\ \mathcal{E}_{\text{Summ}}(q) & t = \textit{Summ}, \\ \mathcal{E}_{\text{KG}}(q) & t = \textit{KG-Fact}. \end{cases}$$

After retrieval, the system constructs a task-specific prompt $\mathcal{P}$ by concatenating the original user query q with retrieved evidence $\mathcal{E}(q,t)$. The retrieved evidence may consist of text passages or structured knowledge-graph outputs depending on the routing decision.

$$\mathcal{P}(q,\mathcal{E},t) = \text{Template}(t) \parallel \mathcal{E} \parallel q,$$

where $\parallel$ denotes string concatenation and $\text{Template}(t)$ is an instruction prefix corresponding to the predicted task (e.g., *"Summarize the following paper:"*, *"Answer the following question..."*). This explicit prompt structure encourages the generator to rely on retrieved evidence rather than parametric memory.

3.3. Response Generation

Instead of relying on large proprietary models, we employ a compact open model (Llama 3.2 3B Instruct) as the response generator. This design choice reflects our central hypothesis: strong retrieval can compensate for smaller model capacity. The model is instruction-tuned on the OpenScholar (Asai et al., 2024) dataset to support scholarly tasks including multi-paper QA, summarization, and editing.

Fine-tuning is performed using the *SFTTrainer* framework (von Werra et al., 2020) with BF16 precision. Multiple variants are trained with either Low-Rank Adapters (LoRA) (Hu et al., 2022) of rank $r = 32$ or full fine-tuning. Training configurations include variable numbers of epochs $\{1, 2, 5\}$ and context lengths $\{8192, 10000, 16384\}$. All models use the *AdamW* optimizer (Loshchilov and Hutter, 2019) with a *cosine-annealing* learning rate schedule, initial learning rate $5 \cdot 10^{-6}$, 200 warmup steps, and an effective batch size of 64. We provide the detailed implementation settings in our repository.

During inference, the system receives the composed prompt $\mathcal{P}(q,\mathcal{E},t)$ constructed in the previous stage, the model generates a response r and a set of citations $\mathbf{C}$:

$$r, \mathbf{C} = g_\phi(\mathcal{P}(q,\mathcal{E},t))$$
$$g_\phi : \mathcal{P} \mapsto (\text{Textual Answer}, \text{Citations})$$

where g_ϕ denotes the fine-tuned LLM parameterized by ϕ. The generated answer includes references corresponding to the retrieved evidence passages. Model serving is performed using *vLLM* (Kwon et al., 2023).

Model	Org↑	Cov↑	Rel↑	Mean↑	Length	Citation Quality		
						R↑	P↑	F1↑
Llama3-8B †*	–	–	–	3.79	–	–	–	0.0%
OS-8B *	3.92	4.44	4.02	4.12	578.6	–	–	42.8%
Llama 3.2 3B Instruct †	3.96	4.06	4.54	4.19	450	0.00%	0.00%	0.00%
Llama 3.1 8B Instruct †	4.01	4.19	4.76	4.32	475	0.00%	0.00%	0.00%
Llama 3.2 3B Instruct	3.67	3.51	3.99	3.72	363	16.49%	19.05%	17.68%
Llama 3.1 8B Instruct	3.83	3.90	4.38	**4.04**	453	23.50%	**26.45%**	**24.89%**
LoRA FT 1ep 8k	3.74	3.78	4.20	3.91	784	23.41%	21.06%	22.18%
LoRA FT 1ep 16k	3.60	3.87	4.08	3.85	1,017	23.35%	21.63%	22.46%
LoRA FT 2ep 8k	3.65	3.81	4.18	3.88	841	24.74%	23.72%	24.22%
LoRA FT 2ep 10k	3.59	**3.92**	**4.31**	3.94	1,297	**25.64%**	22.77%	24.12%
LoRA FT 2ep 16k	3.55	3.82	4.21	3.86	887	24.07%	25.02%	24.53%
LoRA FT 5ep 16k	3.55	3.80	4.27	3.87	1,024	20.10%	18.75%	19.40%
Full FT 1ep 16k	3.81	3.80	4.20	3.94	532	24.51%	23.33%	23.91%
Full FT 2ep 10k	3.85	3.83	4.19	3.96	468	24.52%	22.69%	23.57%
Full FT 2ep 16k	**3.89**	3.88	4.23	4.00	466	21.20%	19.66%	20.40%
Full FT 5ep 16k	3.82	3.84	4.26	3.98	484	23.81%	23.96%	23.89%

Table 2: Multi-Paper QA Evaluation results (ScholarQABench-Multi dataset). Best value per *model type* is underlined; overall best values (excluding external baselines) are **bolded**; LoRA FT rows denote LoRA fine-tuning applied to Llama 3.1 8B Instruct; * results from (Asai et al., 2024); † without retrieval.

4. Evaluation

We evaluate our framework to assess whether our design combining task-aware retrieval and lightweight model can effectively support scientific applications. Our evaluation focuses on the extent to which small language models, when combined with task-aware retrieval strategies, can achieve competitive performance on scholarly tasks. In addition, we examine the robustness of the system under varying retrieval conditions, including differences in retrieval quality and domain alignment between the corpus and evaluation datasets. We select evaluation datasets to reflect complementary aspects of scholarly tasks:

1. **ScholarQABench-Multi (Asai et al., 2024)** for multi-document QA and reasoning,

2. **PubMedQA (Jin et al., 2019)** for domain transfer and robustness, and

3. **SciTLDR (Cachola et al., 2020)** for extreme summarization.

All experiments follow the pipeline described in Section 3. Incoming queries are first routed to a task category, relevant context is retrieved, and the composed prompt is passed to the language model for response generation. We evaluate the framework on two primary scholarly tasks: **scientific question answering** (§4.1) and **text compression** (§4.2), detailed in the following sections.

4.1. Scientific Question Answering Evaluation

For scientific question answering, we evaluate in two settings: (1) multi-paper QA (§4.1.1), which requires synthesizing information across several documents, and (2) single-paper QA (§4.1.2), focusing on questions grounded in a single paper.

4.1.1. Multi-Paper Question Answering

We evaluate the framework on multi-paper question answering using the ScholarQABench-Multi benchmark (Asai et al., 2024). The dataset contains 108 questions spanning computer science, physics, and biomedical research. Answers are evaluated using LLM-based judges (Prometheus models) across three dimensions: Organization (*Org*), Relevance (*Rel*), and Coverage (*Cov*). Each dimension is scored on a 1–5 scale and averaged to obtain an overall quality score (*Mean*). In addition, citation quality is measured using Precision, Recall, and F1 with respect to gold references. Table 2 summarizes our evaluation results.

Impact of Retrieval. We first compare base models with and without retrieval augmentation. Surprisingly, LLM evaluation scores are slightly higher when retrieval is disabled. This suggests that long prompts containing multiple retrieved passages can make generation more difficult for smaller models. However, without retrieval the models are unable to produce verifiable citations as expected, resulting in

Model	Orig Acc↑	Orig F1↑	OS Acc↑	OS F1↑	Zero Acc↑	Zero F1↑
BioBERT †	68.08	52.72	–	–	–	–
OS-8B *	–	–	**76.4**	–	–	–
Llama 3.2 3B Instruct	62.40	40.93	49.11	49.04	64.41	58.97
Llama 3.1 8B Instruct	**75.60**	**54.06**	45.91	43.35	**64.77**	**59.74**
LoRA FT 2ep 10k context	68.00	46.69	58.01	55.83	58.13	57.23
Full FT 5ep 16k context	62.80	38.69	58.01	**56.21**	60.62	57.24

Table 3: Single-paper QA evaluation results (PubMedQA dataset). Best values are **bolded**. † results from (Jin et al., 2019); * results from (Asai et al., 2024).

Model	R1 F1↑	R2 F1↑	RL F1↑	BERTScore F1↑	SMOG Index↑	Compression Ratio↑
CATT (Cachola et al., 2020)	**44.9**	**22.6**	**37.3**	–	–	**47.3**
Llama 3.2 3B Instruct	1.21	0.011	1.21	54.55	24.45	20.72
Llama 3.1 8B Instruct	1.21	**0.014**	1.21	54.57	**24.81**	20.79
Full FT 5ep 16k context	**1.31**	**0.014**	**1.30**	**54.74**	23.26	**22.43**

Table 4: Text compression evaluation results (SciTLDR dataset). Best values are **bolded**. R1, R2, and RL denote ROUGE-1, ROUGE-2, and ROUGE-L F1 scores. Scores are averages over test samples.

a citation F1 score of 0. Retrieval therefore remains essential for grounding responses in evidence.

Fine-Tuning Effects. Fine-tuning substantially improves both answer quality and citation grounding. Our fully fine-tuned 3B model approaches the performance of a substantially larger 8B model. Low-Rank Adaptation (LoRA) improves performance relative to the base model but occasionally produces repetitive outputs or unstable citation formatting, which negatively affects organization scores. Across experiments, training duration and context length have limited impact on overall performance, though longer training slightly degrades organization for LoRA models.

These results on multi-paper QA highlight a key trade-off: while retrieval enables citation grounding, it can introduce noise that negatively affects answer organization and fluency, particularly for smaller models with limited context handling capacity.

4.1.2. Single-Paper Question Answering

We evaluate single-paper question answering using the biomedical PubMedQA (Jin et al., 2019) dataset to further assess retrieval robustness and domain generalization. The task requires predicting *yes*, *maybe*, or *no* answers to biomedical questions. We report our evaluation results for the best-performing LoRA and full fine-tuned models, as well as pretrained Llama 3.1 8B and 3B models as comparison. The results are summarized in Table 3.

We follow the retrieval-based evaluation protocol introduced in ScholarQABench (Asai et al., 2024) and consider the following three evaluation settings.

Original Task (Orig). In the original task setup, models are provided with gold context passages from the relevant paper abstracts. Under this setup, larger models achieve higher accuracy, indicating stronger reasoning ability when reliable evidence is available. The LoRA-adapted 3B model shows notable improvement over its base version despite not being trained on PubMedQA, suggesting that lightweight adaptation can improve domain transfer.

Retrieval Task (OS) In the retrieval variant, models must identify relevant passages from the corpus. The performance drop in the retrieval setting reflects a domain mismatch between the unarXive corpus (primarily computer science and physics) and PubMedQA (biomedical domain). This setup allows us to explicitly evaluate robustness under domain shift. Nevertheless, we show that the fine-tuned models outperform their pretrained counterparts, indicating that training improves the model's ability to filter irrelevant context by ignoring irrelevant retrievals and leveraging retrieved content.

Zero-Context Task (Zero). Finally, we evaluate models without any retrieved evidence. In this setting, answers therefore rely entirely on the models' parametric knowledge. Pretrained models perform slightly better than fine-tuned ones, suggesting that task-specific training may introduce mild

specialization that reduces general-domain recall. This highlights the trade-off between fine-tuning for domain-specific QA and preserving general domain knowledge.

4.2. Text Compression Evaluation

To evaluate summarization capabilities, we benchmark the system on the SciTLDR (Cachola et al., 2020) dataset, which focuses on extreme compression of scientific papers. The task requires generating a one-sentence summary from the abstract, introduction, and conclusion (AIC) sections. Table 4 summarizes our evaluation results.

Generated summaries are compared against the gold TLDR statement summaries reviewed by authors and peer reviewers. Overlap metrics were computed using ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004), as well as BERTScore (Zhang et al., 2020). Following prior work (Cachola et al., 2020), we report the maximum target score for each metric to reduce variability across reference summaries. In addition, we compute a compression ratio between the source text and generated summaries. Because individual summaries consist of a single sentence, traditional readability metrics (*Flesch, Flesch-Kincaid, SMOG*) cannot be applied per example. Thus, we concatenate outputs and compute aggregate readability scores[3] to approximate lexical and syntactic complexity.

We compare our fine-tuned 3B model with baseline Llama models and the specialized CATT model used in the original SciTLDR benchmark. The results show modest improvements from fine-tuning. ROUGE and BERTScore increase slightly, and the compression ratio improves relative to the base model. However, generated summaries remain longer than the target TLDR statements, and readability metrics suggest relatively dense language. These findings highlight the difficulty of extreme scientific summarization and confirm that models specifically trained for the task continue to outperform general-purpose LLMs.

5. Conclusion

We presented a lightweight retrieval-augmented framework for scholarly assistance that combines task-aware routing, hybrid retrieval, and compact language models within a unified architecture. Rather than relying on increasingly large proprietary systems, our work investigates under which conditions improved retrieval design can compensate for reduced model scale across scholarly tasks.

Our evaluation results show that small instruction-tuned models can approach the performance of larger systems when paired with appropriate retrieval strategies, particularly for tasks requiring grounded, citation-based answers. Fine-tuning further improves robustness in multi-document question answering and helps mitigate the impact of partially irrelevant context. However, these gains are not uniform: model capacity remains a key factor for reasoning quality, especially in complex or domain-shifted settings.

Importantly, our findings highlight that improvements in retrieval introduce inherent trade-offs. While more precise retrieval can improve grounding and interpretability, it may reduce recall or introduce longer, noisier contexts that disproportionately affect smaller models. In addition, retrieval pipelines incur additional engineering complexity and may generalize poorly when the underlying corpus does not match the target domain, as observed in our biomedical QA evaluation. These results emphasize that retrieval and model scale should be viewed as complementary components rather than interchangeable solutions.

Overall, our study suggests that progress in retrieval quality is as critical as progress in model scaling for building practical and efficient scholarly assistants. Future work will focus on improving retrieval robustness, expanding domain coverage, and incorporating automated verification of generated outputs. We also plan to explore efficiency-oriented improvements, such as more accurate routing strategies and lightweight reranking, alongside broader evaluations that assess usability, readability, and trust in real-world scholarly workflows.

6. Limitations

First, although the proposed design integrates structured scholarly metadata from knowledge graph and textual evidence within a single pipeline, a standardized benchmark for question answering over the SemOpenAlex knowledge graph is currently unavailable. Thus, current evaluation focuses on the text-based components, while the KG-Fact module is presented primarily as an architectural capability. Constructing reliable SPARQL-question pairs for large scholarly knowledge graphs remains an open challenge. Second, retrieval quality remains a bottleneck. Dense vector search may return passages that are only loosely related to the query, reducing grounding quality and potentially leading to weak or partially supported citations.

Finally, the current system does not include a post-generation verification step to validate citations against source documents. Incorporating reference validation or retrieval-based fact checking would be an important step toward more reliable scholarly assistants.

[3] https://github.com/cdimascio/py-readability-metrics

7. Ethical Considerations

All datasets used are publicly available under research-friendly licenses, e.g., ScholarQABench-Multi (Asai et al., 2024) is released under the ODC-BY license, with some constituent datasets subject to their own licensing terms. Our 165K-paper datastore unarXive (Saier et al., 2023) comprises open-access content compliant with text and data mining permissions. The system is designed as an assistive tool for scholarly tasks and is not intended for use in high-stakes domains without human oversight. We acknowledge potential biases in the underlying literature (e.g., publication bias, domain skew) that may influence system behavior. Additionally, our approach encourages the use of smaller or lightweight models (e.g., 3B parameters) over larger models to reduce computational and environmental impact.

8. Bibliographical References

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, Dan Weld, Doug Downey, Wen-tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs.

Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know your audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI 2024, New York, NY, USA.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, Zheng Cheng, Zifeng Zhao, Linfeng Zhang, and Guolin Ke. 2025. Sciassess: Benchmarking LLM proficiency in scientific literature analysis. In *Findings of the Association for Computational Linguistics*, NAACL 2025, pages 2335–2357.

Harrison Chase. 2022. LangChain.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazar'e, Maria Lomeli, Lucas Hosseini, and Herv'e J'egou. 2024. The faiss library. *ArXiv*, abs/2401.08281.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2024, pages 9194–9211.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, ICLR 2022.

Mohamad Yaser Jaradeh, Markus Stocker, and Sören Auer. 2020. Question answering on scholarly knowledge graphs. In *Proceedings of the 24th International Conference on Theory and Practice of Digital Libraries*, TPDL 2020, pages 19–32.

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 2019, pages 2567–2577.

Sebastian Joseph, Lily Chen, Jan Trienes, Hannah Göke, Monika Coers, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. FactPICO: Factuality evaluation for plain language summarization of medical evidence. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, ACL 2024, pages 8437–8464.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm,

Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. 2023. The semantic scholar open data platform.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP 2023, pages 611–626.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, NeurIPS 2020.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025. Scilitllm: How to adapt llms for scientific literature understanding. In *Proceedings of the Thirteenth International Conference on Learning Representations*, ICLR 2025.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*, ICLR 2019.

Junyu Luo, Junxian Lin, Chi Lin, Cao Xiao, Xinning Gui, and Fenglong Ma. 2022. Benchmarking automated clinical language simplification: Dataset, algorithm, and evaluation. In *Proceedings of the 29th International Conference on Computational Linguistics*, COLING 2022, pages 3550–3562, Gyeongju, Republic of Korea.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, ACL 2023, pages 9802–9822, Toronto, Canada.

Gary Marcus. 2022. The galactica ai model was trained on scientific knowledge – but it spat out alarmingly plausible nonsense. Accessed: 2025-09-17.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *Proceedings of the First Conference on Language Modeling*, COLM 2024.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL 2002, page 311–318, USA.

Tarek Saier, Johan Krause, and Michael Färber. 2023. unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network. In *Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries*, JCDL 2023, pages 66–70.

Junhong Shen, Neil A. Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolò Fusi. 2024. Tag-llm: Repurposing general-purpose llms for specialized domains. In *Proceedings of the 41st International Conference on Machine Learning*, ICML 2024, pages 44759–44773.

Sotaro Takeshita, Simone Paolo Ponzetto, and Kai Eckert. 2024. ROUGE-K: do your summaries have keywords? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics*, *SEM 2024, pages 69–79.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

David Wadden, Kejian Shi, Jacob Morrison, Alan Li, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2025. Sciriff: A resource to enhance language model instruction-following over scientific literature. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2025, pages 6072–6109.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR 2020.

9. Language Resource References

Asai, Akari and He, Jacqueline and Shao, Rulin and Shi, Weijia and Singh, Amanpreet and Chang, Joseph Chee and Lo, Kyle and Soldaini, Luca and Feldman, Sergey and D'arcy, Mike and Wadden, David and Latzke, Matt and Tian, Minyang and Ji, Pan and Liu, Shengyan and Tong, Hao and Wu, Bohao and Xiong, Yanyu and Zettlemoyer, Luke and Neubig, Graham and Weld, Dan and Downey, Doug and Yih, Wen-tau and Koh, Pang Wei and Hajishirzi, Hannaneh. 2024. *OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs*.

Auer, Sören and Barone, Dante A. C. and Bartz, Cassiano and Cortes, Eduardo G. and Jaradeh, Mohamad Yaser and Karras, Oliver and Koubarakis, Manolis and Mouromtsev, Dmitry and Pliukhin, Dmitrii and Radyush, Daniil and Shilin, Ivan and Stocker, Markus and Tsalapati, Eleni. 2023. *The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge*.

Sören Auer and Allard Oelen and Muhammad Haris and Markus Stocker and Jennifer D'Souza and Kheir Eddine Farfar and Lars Vogt and Manuel Prinz and Vitalis Wiens and Mohamad Yaser Jaradeh. 2020. *Improving Access to Scientific Literature with Knowledge Graphs*.

Cachola, Isabel and Lo, Kyle and Cohan, Arman and Weld, Daniel. 2020. *TLDR: Extreme Summarization of Scientific Documents*.

Färber, Michael and Ao, Lin. 2022. *The Microsoft Academic Knowledge Graph enhanced: Author name disambiguation, publication classification, and embeddings*.

Färber, Michael and Lamprecht, David and Krause, Johan and Aung, Linn and Haase, Peter. 2023. *SemOpenAlex: The Scientific Landscape in 26 Billion RDF Triples*.

Jin, Qiao and Dhingra, Bhuwan and Liu, Zhengping and Cohen, William and Lu, Xinghua. 2019. *PubMedQA: A Dataset for Biomedical Research Question Answering*.

Rodney Kinney and Chloe Anastasiades and Russell Authur and Iz Beltagy and Jonathan Bragg and Alexandra Buraczynski and Isabel Cachola and Stefan Candra and Yoganand Chandrasekhar and Arman Cohan and Miles Crawford and Doug Downey and Jason Dunkelberger and Oren Etzioni and Rob Evans and Sergey Feldman and Joseph Gorney and David Graham and Fangzhou Hu and Regan Huff and Daniel King and Sebastian Kohlmeier and Bailey Kuehl and Michael Langan and Daniel Lin and Haokun Liu and Kyle Lo and Jaron Lochner and Kelsey MacMillan and Tyler Murray and Chris Newell and Smita Rao and Shaurya Rohatgi and Paul Sayre and Zejiang Shen and Amanpreet Singh and Luca Soldaini and Shivashankar Subramanian and Amber Tanaka and Alex D. Wade and Linda Wagner and Lucy Lu Wang and Chris Wilhelm and Caroline Wu and Jiangjiang Yang and Angele Zamarron and Madeleine Van Zuylen and Daniel S. Weld. 2023. *The Semantic Scholar Open Data Platform*.

Saier, Tarek and Krause, Johan and Färber, Michael. 2023. *unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network*.