EGT-KG: Evidence-Grounded Typed KG Retrieval for Practical Scientific QA with Small Language Models

Muran Yu[1], **Jiechao Gao**[1], **Yuandong Pan**[1], **Barney H. Miao**[1,2]
Andrew C. Lesh[1], **Kincho H. Law**[1], **Jie Wang**[1], **Michael D. Lepech**[1]
[1]Department of Civil and Environmental Engineering, Stanford University
[2]Department of Civil Engineering, University of Calgary
{yumuran, jiechao, ydpan, barneym, aclesh, law, jiewang, mlepech}@stanford.edu

Abstract

For emerging scientific research domains, local Small Language Models (SLMs) are becoming more attractive, as they offer stronger privacy control and more stable deployment pipelines than Large Language Models. However, in practice, scientific question-answering on SLMs often operates under inevitable constraints: small literature collections, fragmented evidence, limited context window and reasoning abilities. We propose the Evidence-Grounded Typed Knowledge Graph (EGT-KG), a retrieval framework to improve information retrieval with local SLMs. We assessed three question-answering settings: a vanilla Retrieval-Augmented Generation (RAG) workflow and two EGT-KG workflows: an automatically generated relation schema (AS) and an expert-defined relation schema (ES). Our experiments were evaluated with a six-dimensional evaluation framework (S3CRF: Soundness, Correctness, Completeness, Conciseness, Relevance, Fluency) on a Biopolymer-bound Soil Composite literature benchmark, showing that EGT-KG outperforms the vanilla RAG method in most settings, with the best improvement from *llama3:8b*: a Final Score of 70.37 (+14.67%) and 68.82 (+12.14%) by AS/ES EGT-KG variants.

1 Introduction

The rapid evolution of Large Language Models (LLMs) has changed the paradigm in scientific knowledge discovery and human-AI collaboration, including scientific question-answering (Zheng et al., 2025). However, the deployment of frontier LLMs for specialized research workflows can raise concerns about data privacy, operational costs, and inference latency (Pan et al., 2025; Cai et al., 2024). Therefore, Small Language Models (SLMs) have become increasingly attractive for domain-specific applications in constrained environments (Wang et al., 2025; Lu et al., 2024). Yet, SLMs are often unreliable for specialized scientific question-answering because they have limited internal knowledge, weaker reasoning ability, and lower tolerance for noisy context than frontier LLMs (Allen-Zhu and Li, 2024; Lee et al., 2026; Mallen et al., 2023). This challenge is particularly critical in emerging research domains, where knowledge is fragmented across a small set of knowledge sources and is not well represented in pretraining corpora.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) can ground local models in external documents, but conventional dense retrieval is often insufficient for scientific QA. Scientific questions may depend on relations among materials, procedures, conditions, observations, and results, while dense RAG mainly retrieves chunks by semantic similarity. As a result, retrieved passages may be broadly relevant but still miss the specific evidence needed for faithful answering.

Graph-based RAG methods address part of this issue by organizing information into entities and relations (Edge et al., 2024). However, general knowledge graphs may lack timeliness and fine-grained domain knowledge (Ding et al., 2024), and automatically constructed graphs can introduce noisy or ambiguous relations (Schäfer et al., 2024; Zhuang et al., 2025). Moreover, converting scientific narratives into triples may discard experimental context, conditional qualifiers, and quantitative details that are necessary for reasoning (Pujara et al., 2017).

We propose the Evidence-Grounded Typed KG (EGT-KG) framework to improve RAG with local SLMs in scientific question-answering settings. Our method extends graph-based retrieval with two principles: First, graph structures should guide retrieval rather than replace the original evidence. Second, the relation structure should be informative enough to be useful in retrieving. Thus, we construct a reified knowledge graph with evidence nodes and classified relations, allowing the system

EMNLP 2026 Industry Track, pages 2333–2354
October 24-29, 2026 ©2026 Association for Computational Linguistics

to use relation labels as a retrieval signal.

We evaluate our framework on Biopolymer-bound Soil Composite (BSC)-centered materials QA benchmark built from a 30-paper corpus. The knowledge of BSC is distributed across a compact but specialized literature collection, making it a realistic test case for small-corpus scientific QA.

We compared a vanilla RAG workflow with two EGT-KG variants: an automatically generated schema (AS) and an expert-designed schema (ES). The results show that EGT-KG enhanced retrieval improves answer quality for most cases, particularly in factual retrieval questions. The core contributions of this paper are summarized as follows: First, to enhance RAG in limited resource settings, we propose an Evidence-Grounded Typed KG framework that combines relation classification with provenance-aware evidence nodes for scientific question-answering with local SLMs. Second, we performed a controlled comparison on the BSC-centered benchmark between different workflows, systematically studying the trade-off between schema granularity and answer quality. Third, we introduce an S3CRF evaluation setting that combines multidimensional scoring (**S**oundness, **C**orrectness, **C**ompleteness, **C**onciseness, **R**elevance, and **F**luency) and retrieval performance analysis, illustrating how answer quality is affected by the six criteria, query expansion, fallback rate, and evidence-window usage.

2 Related Works

Retrieval. Retrieval-Augmented Generation (RAG) is a widely adopted approach to improving question-answering by providing external knowledge that is not stored in the model's parameters (Lewis et al., 2020). It can improve factual grounding and reduce hallucinations, demonstrating its potential for question-answering in professional domains such as medicine, education, and law (Fan et al., 2024). However, dense vector representations can lead to irrelevant documents, potentially ignoring explicit structured dependencies, properties, conditions, or procedures. In addition, computational overhead, memory usage, and RAG poisoning are of concern in the implementation of RAG system (Gupta et al., 2024; Greshake et al., 2023; Liu et al., 2025, 2024). Therefore, a cost-efficient, stable, and secure RAG system is demanded for sensitive tasks that require high precision.

Graph-RAG. Knowledge graph-enhanced retrieval methods mitigate the limitations of RAG by providing global descriptions and the hierarchical community structure of the corpus (Fu et al., 2025). Graph-RAG has shown that the graph structure can help multi-hop evidence discovery, avoiding the loss of target information in a single semantically similar chunk (Procko and Ochoa, 2024). However, current graph-based approaches compress the original text, which can potentially lead to information loss and further degrade response generation. Additionally, it also tends to utilize a large portion of triples in the knowledge graph. Without constraints, users may face context explosion and high reasoning costs (Xiang et al., 2025). Guo et al. (2024) proposed LightRAG to integrate a graph-based indexing approach to improve the performance of RAG in both efficiency and comprehension in information retrieval. However, the final passage that is passed to the answer model is still pre-processed by the language model, potentially introducing unverified noise.

Relation Classification. Relation classification is a well-studied research area in NLP since the last decade. Methods such as feature-based models, kernel-based models, neural-network models, and transformer encoders are utilized to assign a semantic relation to two nominals (Kambhatla, 2004; Bunescu and Mooney, 2005; Socher et al., 2012; Huang et al., 2020). However, these applications in relation classification focus on processing original text information rather than the relation node from the extracted triple in knowledge graph. Yu et al. (2020) proposed a relationship extraction method for construction of domain knowledge graphs using existing reference material from Wikipedia. Along with other semi-supervised relation classification methods (Agichtein and Gravano, 2000; Ravichandran and Hovy, 2002; Pantel and Pennacchiotti, 2006), these classification processes require the collection of additional corpora from the internet. Applying these methods to our framework will introduce additional computational costs, which is stressful for local SLMs.

3 Methodology

We propose an Evidence-Grounded Typed Knowledge Graph retrieval framework for domain-specific question-answering on SLMs. This framework aims to improve the retrieval quality, where the final answer will be grounded on original evi-

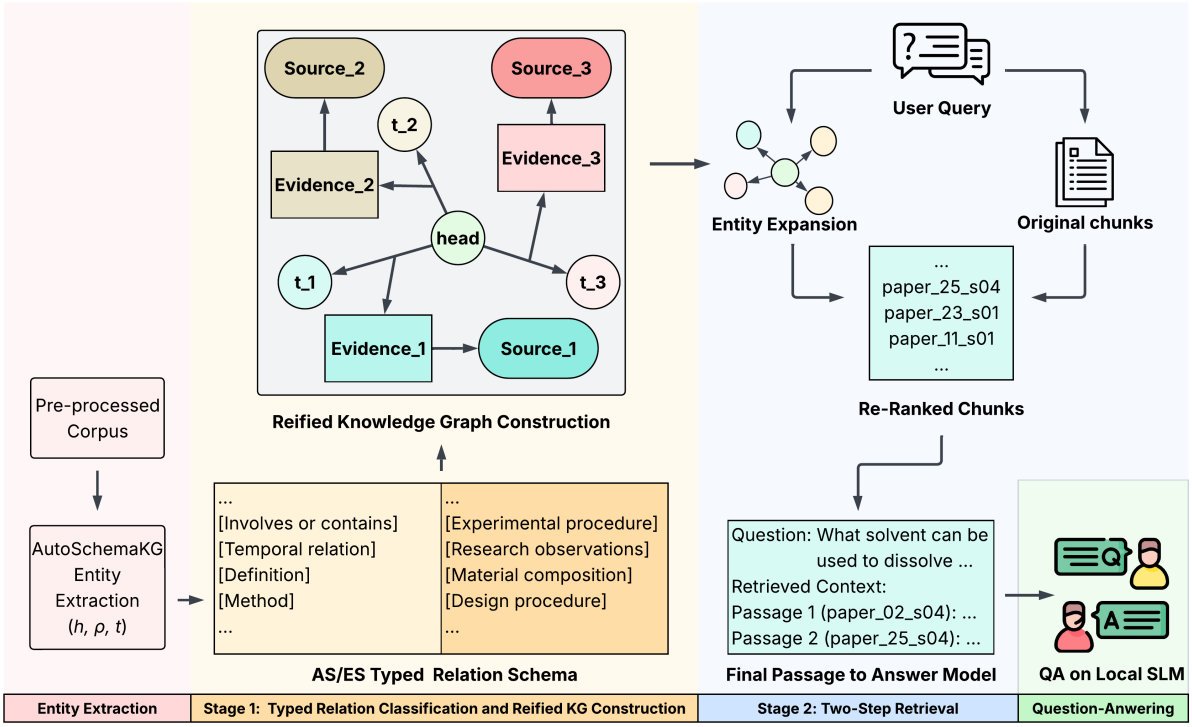

Figure 1: Overview of the EGT-KG framework. The framework consists of four main sections: entity extraction, typed relation classification and reified KG construction, two-step retrieval with re-ranking, and final QA.

dence nodes rather than pre-processed information.

The complete workflow for our framework is organized as follows. In the first stage, document chunks are processed to extract raw triples with short supporting evidence spans. Then, the extracted free-form relation labels are mapped into the two predefined typed schemas accordingly (AS and ES). These typed triples will be used to construct a reified knowledge graph with evidence nodes converted from raw evidence spans, linking the triples to unprocessed evidence text and source chunks. In the second stage, the reified knowledge graph guides query expansion and chunk candidates retrieval. Then the re-ranked chunk candidates and evidence window are passed to the answer model. Fig 1 provides an overview of the framework structure.

3.1 Graph triple extraction

The corpus set is built on a 30-paper corpus centered on Biopolymer-bound Soil Composite (BSC) and related emerging biopolymer-based materials. It includes 9 papers directly related to the evaluation questions, 9 BSC papers not directly related to the questions, 9 materials science papers similar to BSC, and 3 general reference papers. This corpus has three unique properties that make it a suitable test case. First, biopolymer-bound soil composites are an emerging material. So, a language model is unlikely to have much knowledge about them parametrically, which can be confirmed by baseline performance in Table 2. Therefore, the improvement of the framework can be attributed to retrieval quality from the corpus. Second, researchers who work on these materials designed the evaluation questions, so the questions are practically meaningful and cover aspects such as material behavior, material properties, and end-of-life treatment. Third, that same expertise made an expert-defined schema that can better capture useful information in KG. Appendix A.6 introduces the material and Table 16 lists the 30 papers.

We divided the collected corpus by document structure, preserving metadata such as section headings, resulting in 884 chunks averaging 1,387 characters. Thus, the retrieved evidence span can be traced back to its original document and local position. To ensure that the downstream comparisons are not confounded by different triple extraction passes, we perform triple extraction only once over the fixed chunk corpus and reuse the raw outputs for both EGT-KG variants.

Formally, let C denote the whole chunk set, and each $c \in C$ carries the metadata $M(c)$ including chunk ID, paper ID, source ID, title, chunk length,

and original text.

We utilize **AutoSchemaKG** (Chen and Bertozzi, 2023) as a triple-extraction engine. Specifically, we use AutoSchemaKG's extraction utilities to process each chunk in C and record structural information.

As a key feature, the extraction output preserves a head entity, a tail entity, a relation, and a short supporting evidence span with localization information rather than a processed summary or description of the triple. Thus, for each chunk $c \in C$, the extractor returns a set of raw triples T that can be represented as: $T(c) = \{h, \rho|e, t\}$, where h and t denote the head and tail entity, ρ is the free-form relation label, and e is a short evidence span that can be traced back to the source chunk. The evidence span e is appended with the relation ρ to be converted to an evidence node for graph construction, which enables Evidence Window slicing that can offer 22% context reduction with no cost in performance. The triple set T will be used by both AS/ES variants in subsequent experiments.

3.2 Typed relation classification

After triple extraction, we transform relation labels into a set of semantically meaningful relation types. We classify each distinct label once into a schema type or discard it, then apply the resulting label-to-type map to every triple with that label, so identical labels are always typed identically. During this process, we restrict the size of the knowledge graph by trimming noisy, inconsistent, and irrelevant triples for retrieval.

The two type schemas are designed to study how relation granularity itself can alter the process of query expansion and quality of responses. As shown in Table 1, the AS schema is automatically generated by answer models based on specific instructions for the BSC study, and the generation prompt is given in Appendix A.8. We selected the most prominent types and applied them to all AS EGT-KG variant studies. The ES schema was defined by two materials scientists with several years of research experience in relevant materials. They reviewed and identified the relation labels produced, and grouped them into 11 types. The AS–ES contrast therefore isolates the value of expert input, and the ES schema can potentially be expanded to broader materials science domains.

Formally, let ϕ denote the relation typing function, and τ denote the type schema: $\phi(T(c)) = (h, \tau_n|\rho|e, t)$,where τ_1 is the AS schema and τ_2 is the ES schema. Applying the relation typing function ϕ will classify each of the extracted triples t into a typed relation instance and be appended to the relation label. Thus, by differentiating relation types, our framework provides a high-level index to guide graph traversal and constrains query expansion, forcing the retrieval process to operate only on the most stable semantic categories.

3.3 Reification and provenance-aware knowledge graph construction

Leemhuis et al. (2022) states that although current methods of knowledge graph embedding perform well on total and functional relations, features such as partiality or non-functionality of relations are difficult to model correctly. So, they introduced the concept of Reification, which represents relations directly as objects in the embedding space rather than as geometric transformations from subject to object. For our framework, the purpose of reification is to preserve each relation instance as a same-level object linked to both its semantic arguments and its original support text. Therefore, rather than representing knowledge as direct entity-to-entity edges only, the reified knowledge graph stores two kinds of additional nodes besides the head and tail. The first is the evidence nodes, which contain the supporting evidence span e, and the second is the source nodes, which store the location of the evidence span in the original document. This reified design is central to our framework because it allows graph retrieval to remain structurally expressive while preserving the textual provenance required for faithful answer generation. This structure is especially important in scientific question-answering, where relation abstraction or description alone may omit crucial conditions, quantitative ranges, or contextual qualifiers in the original text information. So far, the first stage is completed, and the overview of stage 1 is shown in Appendix A.7.

3.4 Two-step retrieval with query expansion and re-ranking

At the beginning of the second stage, our EGT-KG framework follows a two-step retrieval strategy. In the first step, to improve recall, the answer model retrieves chunk candidates by using the reified knowledge graph to guide query expansion. In the second step, to improve precision, the retrieved chunk candidates are re-ranked according to their evidence-level relevance. The two-step retrieval design is guided by the hypothesis of our work, in which query expansion helps find the most relevant neighbor entities, while chunk candidates re-ranking helps retrieve the correct final passages.

In practice, the two-step retrieval works as fol-

Relation Type	Composition
AS schema (automatically generated)	Involves or Contains; Causes or Influences; Temporal Relation; Logical or Preposition; Research or Observation; Method; Definition; Production; Spatial Relation.
ES schema (expert designed)	Causes or Influences; Definition; Manufacture Procedure; Temporal Relation; Experimental Methods; Logical Reasoning; Research Observations; Experimental Procedures; Material Composition; Design Procedure; Biopolymer Properties.

Table 1: Typed relation schemas for knowledge graph construction.

lows. In the first step, given a user query q, we identify seed entities in KG through query-entity matching. Next, we perform a one-hop query expansion around the seed entities to collect related entities. The first-step retrieval score (S_1) is the weighted sum of the original-query score S_{ori} and the expanded-query relevance score S_{exp}, allowing the retrieval to benefit from both global and local views. This score will be used to select top-k chunk candidates. Formally, for a chunk candidate c and weight term α:

$$S_1(c|q) = \alpha S_{ori}(c|q) + (1-\alpha)S_{exp}(c|q) \quad (1)$$

For the second step, the top-k chunk candidates collected in step 1 are re-ranked by comparing the semantic similarity between the query and the content of the evidence nodes. The re-ranking process provides more precise support evidence within the retrieved material and determines if the final answer context should be constructed from localized evidence window or from full chunk fallback. Specifically, let S_{evi} denote the evidence-level score, and λ denote the weight for step 2, the re-ranking score S_2 is defined as:

$$S_2(c,e|q) = \lambda S_1(c|q) + (1-\lambda)S_{evi}(e|q) \quad (2)$$

For a complete illustration, Figure 3 in Appendix A.7 provides a detailed example.

3.5 Evidence-grounded answer generation

After two-step retrieval and re-ranking, the system constructs the final context passage from either top-k evidence windows or full-chunk fallbacks. All compared workflows, including the vanilla RAG and two EGT-KG variants, share the same answer-generation policy, instructing the answer model to answer the question directly with no unsupported inference, and explicitly acknowledging reference information insufficiency when retrieved chunks do not contain enough information. Fig 3 in Appendix A.7 shows the overview of the second stage.

Formally, given a question q and a final context passage P_q, the shared evidence-based prompt π, the answer model produces:

$$answer = SLM(q, P_q, \pi) \quad (3)$$

3.6 Evaluation with S3CRF

All generated responses are evaluated under a unified LLM-as-judge workflow. Due to the unique nature of our problem, current benchmarks are not best suited to evaluate emerging research domains with limited reference sources and open-ended questions. We adopt the S3CRF framework, scoring answers along six dimensions: soundness, correctness, completeness, conciseness, relevance, and fluency. The detailed evaluation criteria are shown in Table 13. Domain experts in BSC were involved in validating the benchmark design and evaluation criteria to improve the credibility of the assessment.

The six-dimensional scores (s_n) are combined into a weighted Final Score (S_{final}) to reflect the benchmark's emphasis on scientific question-answering demands, the overall Final Score with weights w_n is computed as: $S_{final} = \sum(w_n s_n)$.

4 Results

4.1 Experimental comparison setup

All compared workflows share the same strict evidence-grounded instruction prompt and the S3CRF multidimensional LLM-as-Judge evaluation framework. For all experiments, the answer models receive the top-k ($k = 6$) retrieved passages to construct the final prompt for each query (Appendix A.9). We report results for five local answer models evaluated by *gpt-4o* (OpenAI, 2024) and *gemini-2.5-pro* (Google, 2025) as judge models.

4.2 Overall performance across models

Table 2 summarizes the overall performance of baseline, vanilla RAG, and two EGT-KG variants. Generally, EGT-KG enhanced retrieval improves answer quality in most cases, but the improvement is clearly model-dependent. The highest absolute values of Final Scores are achieved by *llama3.1:8b* in the benchmark, and *gemma2:9b*

	Gemma2:9b		Llama3:8b		Llama3.1:8b		Mistral:7b		Qwen2.5:7b	
	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini
Baseline	35.96	24.60	28.02	24.80	29.19	24.95	28.65	22.29	33.51	24.43
Vanilla RAG	66.76	70.58	61.37	62.44	84.30	77.33	59.32	77.55	**86.01**	78.09
EGT-KG(AS)	72.22	72.62	**70.37**	66.11	83.75	78.44	60.34	**77.78**	82.82	**80.55**
EGT-KG(ES)	**73.12**	**74.04**	68.82	**68.21**	**86.02**	**81.72**	**60.57**	74.23	85.20	79.91

Table 2: Performance of workflows across answer models and judge models.

Model	Relation Type	Fallback	Window	AvgExpand	EE	Jaccard
Gemma2:9b	AS	0.6292	0.3708	10.40	3.57%	0.6095
Gemma2:9b	ES	0.6167	0.3833	10.00	3.83%	0.5678
Llama3:8b	AS	0.7500	0.2500	14.95	1.67%	0.5961
Llama3:8b	ES	0.7667	0.2333	16.90	1.38%	0.5855
Llama3.1:8b	AS	0.6042	0.3958	17.10	2.31%	0.5762
Llama3.1:8b	ES	0.6375	0.3625	16.55	2.19%	0.5427
Mistral:7b	AS	0.7792	0.2208	13.13	1.68%	0.6186
Mistral:7b	ES	0.7917	0.2083	12.80	1.63%	0.6037
Qwen2.5:7b	AS	0.7167	0.2833	24.15	1.17%	0.4930
Qwen2.5:7b	ES	0.7208	0.2792	25.05	1.11%	0.4847

Table 3: Retrieval performance analysis. Expansion efficiency (EE) is defined as $\mathrm{WindowRate}/\mathrm{AvgExpanded}$, and the Jaccard index measures the difference between original retrieved and re-ranked chunk candidates.

demonstrates the most stable improvements from EGT-KG enhanced retrieval. Besides, *llama3:8b* shows the greatest relative improvement compared to the vanilla RAG workflow, while *mistral:7b* and *qwen2.5:7b* exhibit smaller or less stable improvements compared to the other answer models.

The largest relative improvement appears for *llama3:8b* under *gpt-4o* evaluation, where the Final Score increases from 61.37 with vanilla RAG to 70.37 (+14.67%) with EGT-KG(AS) and 68.82 (+12.14%) with EGT-KG(ES). For *gemma2:9b*, it outperforms vanilla RAG under both judge models, presenting the most stable improvements as the granularity of the type schema increases. And *llama3.1:8b* reaches the highest overall scores with the ES schema (86.02 by *gpt-4o* and 81.72 by *gemini-2.5-pro*). By contrast, *mistral:7b* and *qwen2.5:7b* show smaller or less stable improvements, indicating that EGT-KG is not uniformly beneficial for all local models.

These results support two conclusions. First, our EGT-KG framework improved scientific question-answering by two-step retrieval from the reified knowledge graph in most settings. Second, generally, a finer-grained type schema can better guide answer models to retrieve the most relevant information from the original text to help with question-answering. However, given the differences in model structure and reasoning abilities, the benefit of additional granularity is conditional rather than universal.

As a supplementary generalization check, we evaluate EGT-KG on QASPER (Dasigi et al., 2021) with a cross-paper setting, where the system retrieves from all 281 papers without access to the gold paper ID. As shown in Appendix A.3, EGT-KG improves mean Answer F1 over vanilla RAG by 5.9% on average, with gains on four out of the five base models. We additionally evaluate on HotpotQA. The outcome depends entirely on how large a candidate pool retrieval must search: the framework adds almost nothing when the official distractor setting supplies ten candidates per question, but improves Answer F1 by $+0.142$ (*gemma2:9b*) once the same paragraphs are pooled into a single corpus (Appendix A.4).

4.3 Retrieval performance and error analysis

To further investigate why the EGT-KG framework can help more in some settings than others, we conducted a retrieval performance analysis in Table 3. The fallback rate is the fraction of the final k chunks for which no evidence window could be constructed, and AvgExpand is the mean number of expansion entities produced per question by one-hop traversal.The most positive cases (*gemma2:9b* and *llama3.1:8b*) show the lowest fallback rates

and highest evidence-window utilization (EE). By contrast, *mistral:7b* has the highest fallback rates (0.78-0.79) with limited improvements in Final Score. Compared to *gemma2:9b*, *mistral:7b* has a similar Jaccard index but a much higher fallback rate, suggesting that the model successfully re-ranked the original chunk candidates, but failed to extract a useful evidence window.

Also, by analyzing average expanded ($AvgExpand$) query entities, we found that retrieving more neighbor entities does not always guarantee better results. Compared to other answer models, *qwen2.5:7b* collected the largest number of neighbor entities through one-hop query expansion, expanding to 25.05 entities per question under ES with the lowest expansion efficiency. The lowest Jaccard index demonstrates *qwen2.5:7b* replaced more original chunk candidates than other answer models, indicating that the model failed to extract useful evidence from these new chunk entities, while these underutilized chunk candidates distracted model's attention from the most relevant information.

5 Discussion

5.1 EGT-KG enhanced retrieval helps by improving evidence access

In general, the experimental results suggest that the main value of the EGT-KG framework lies in improving the precision of textual evidence passed to the answer model. With the same prompt structure, the improvements observed in EGT-KG variants over the vanilla RAG workflow can plausibly be explained by better retrieval quality and evidence localization than by refined prompt engineering.

To be specific, the reified knowledge graph helps the answer model to identify relevant entities, expand query around structured neighbor entities, and re-rank chunk candidates based on evidence-level matching. Thus, the EGT-KG framework should be treated as a lightweight retrieval booster for SLMs in a limited resources setting.

5.2 Finer-grained schema is not always better

The comparison between the two variants demonstrated that increased granularity does not always lead to better performance. Although some models can consistently benefit from the finer-grained type schema, the coarse AS EGT-KG variants remain more effective for models that cannot fully interpret additional knowledge injected. This result reveals the trade-off between semantic precision and retrieval robustness. Finer schema granularity can surely provide more informative instruction that aligns the original query with reified knowledge graph, but on the other hand, it may introduce sparsity and classification noise, especially for local SLMs, which are typically less robust to noise or weakly aligned context than frontier LLMs. The coarse schema, with more general relation classes defined, may not be as specific as the finer one, but it provides a more stable retrieval path to SLMs.

5.3 Retrieval quality is mainly constrained by evidence localization

The retrieval analysis in section 4 indicates that the bottleneck of retrieval quality is evidence localization quality, rather than the number of retrieved chunk candidates. The framework successfully expands the range of retrieved chunk candidates via one-hop query expansion with 30% of the original chunk candidates replaced after re-ranking. However, the response quality highly depends on whether the answer model can locate a clear evidence window in retrieved chunks. The models with better performance in our framework gained lower fallback rates and higher evidence-window utilization rates, while the weaker models always fall back to full chunk information passing. Thus, for our specific scenario with limited resources, more graph expansion does not automatically improve results. Instead, the model's ability to convert chunk information into grounded evidence should be prioritized when determining whether the expanded retrieval is helpful or noisy.

6 Conclusion

We presented EGT-KG, an evidence-grounded typed KG retrieval framework for practical scientific question-answering with local small language models. Our findings suggest that the EGT-KG framework improves answer quality over vanilla RAG in most settings, especially on dimensions related to factual correctness, reasoning soundness, completeness, and relevance. This observation is especially important for SLMs when scientific QA systems must operate on specialized corpora with fragmented evidence and limited model capacity. Future work will focus on reducing fallback rate, balancing multi-hop expansion with limited model context length, improving evidence window utilization, and testing the framework on more domains of scientific research.

7 Limitations

The current study has several limitations. Firstly, the relatively high fallback rate indicates that there is still a large portion of retrieved chunk candidates that cannot be converted into a clear evidence window. Therefore, the potential of our provenance-aware retrieval is not fully exploited. The current results may be improved with a better solution for evidence localization for SLMs. Secondly, our query expansion relies on one-hop neighbor entity expansion. While this choice takes noise management and SLM's limited reasoning ability into consideration, it reduces our framework's capability in handling complex queries that require multi-hop reasoning or cross-document inference. Third, the current validation is conducted on a BSC-centered benchmark. Although our corpus setting is appropriate as a test case for open-ended question-answering with fragmented and specialized scientific knowledge, the present results should not be generalized to a broader scope of research domains without further validation. The additional evaluation on QASPER has partially shown generalization, but its question set frequently uses pronouns instead of precise terminology. In cross-paper settings, the performance of EGT-KG is underestimated because of the nature of the question set. Fourth, our work focuses on improving retrieval for local SLMs rather than on comparing against frontier LLMs as answer generators. Thus, the results should be interpreted as evidence that better retrieval quality can strengthen feasible local SLMs in constrained resource settings, rather than a claim of general superiority over LLMs. Finally, for the designed open-ended questions, the LLM-as-judge method remains sensitive to judge model preferences. The use of two judges improves robustness, but the current Final Score difference should be treated as a relative performance indicator rather than as a measure of actual scientific usefulness.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*.

Maria I Allende, B Alan Davis, Joshua E Miller, Eric L Christiansen, Michael D Lepech, and David J Loftus. 2020. Hypervelocity impact performance of biopolymer-bound soil composites for space construction. *Journal of Aerospace Engineering*, 33(2):04020001.

Mohamed K Ayeldeen, Abdelazim M Negm, and Mostafa A El Sawwaf. 2016. Evaluating the physical characteristics of biopolymer/soil mixtures. *Arabian Journal of Geosciences*, 9(5):371.

Adrian Biggerstaff, Gerald Fuller, Michael Lepech, and David Loftus. 2021. Determining the yield stress of a biopolymer-bound soil composite for extrusion-based 3d printing applications. *Construction and Building Materials*, 305:124730.

Adrian Biggerstaff, Michael Lepech, Gerald Fuller, and David Loftus. 2022a. A shape stability model for 3d printable biopolymer-bound soil composite. *Construction and Building Materials*, 321:126337.

Adrian Biggerstaff, Michael Lepech, and David Loftus. 2022b. Determining the structuration of biopolymer-bound soil composite. *Materials and Structures*, 55(7):190.

Adrian O Biggerstaff, Michael Lepech, and David Loftus. 2025. A study on the flow behavior and thixotropy of biopolymer-bound soil composite. *Journal of Materials in Civil Engineering*, 37(2):04024487.

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 724–731.

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. 2024. Medusa: Simple llm inference acceleration framework with multiple decoding heads. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Ilhan Chang and Gye-Chun Cho. 2014. Geotechnical behavior of a beta-1, 3/1, 6-glucan biopolymer-treated residual soil. *Geomech. Eng*, 7(6):633–647.

Ilhan Chang, Jooyoung Im, Awlia Kharis Prasidhi, and Gye-Chun Cho. 2015a. Effects of xanthan gum biopolymer on soil strengthening. *Construction and Building Materials*, 74:65–72.

Ilhan Chang, Minkyung Jeon, and Gye-Chun Cho. 2015b. Application of microbial biopolymers as an alternative construction binder for earth buildings in underdeveloped countries. *International journal of polymer science*, 2015(1):326745.

Bohan Chen and Andrea L. Bertozzi. 2023. Autokg: Efficient automated knowledge graph generation for language models. *Preprint*, arXiv:2311.14740.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *Preprint*, arXiv:2105.03011.

Linyi Ding, Sizhe Zhou, Jinfeng Xiao, and Jiawei Han. 2024. Automated construction of theme-specific knowledge graphs. *arXiv preprint arXiv:2404.19146*.

Cassandra A Dove, Fiona F Bradley, and Siddharth V Patwardhan. 2016. Seaweed biopolymers as additives for unfired clay bricks. *Materials and Structures*, 49(11):4463–4482.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Chenhan Fu, Guoming Wang, Rongxing Lu, and Siliang Tang. 2025. Global discovery: A global graph-rag approach for query-focused multimodal summarization across multiple pdf papers. In *Knowledge Science, Engineering and Management: 18th International Conference, KSEM 2025, Macao, China, August 4–7, 2025, Proceedings, Part V*, page 1–8, Berlin, Heidelberg. Springer-Verlag.

Google. 2025. Gemini 2.5 pro.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 79–90, New York, NY, USA. Association for Computing Machinery.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *arXiv preprint arXiv:2410.12837*.

Wenti Huang, Yiyu Mao, Zhan Yang, Lei Zhu, and Jun Long. 2020. Relation classification via knowledge graph enhanced transformer encoder. *Knowledge-based systems*, 206:106321.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction. In *Proceedings of the ACL interactive poster and demonstration sessions*, pages 178–181.

Hamid Reza Khatami and Brendan C O'Kelly. 2013. Improving mechanical properties of sand using biopolymers. *Journal of Geotechnical and Geoenvironmental Engineering*, 139(8):1402–1406.

A Kumar, RK Mishra, K Verma, SM Aldosari, CK Maity, S Verma, R Patel, and VK Thakur. 2023. A comprehensive review of various biopolymer composites and their applications: from biocompatibility to self-healing. *Materials Today Sustainability*, 23:100431.

Seongyun Lee, Yongrae Jo, Minju Seo, Moontae Lee, and Minjoon Seo. 2026. Lost in the noise: How reasoning models fail with contextual distractors. *Preprint*, arXiv:2601.07226.

Mena Leemhuis, Özgür L Özçep, and Diedrich Wolter. 2022. Knowledge graph embeddings with ontologies: reification for representing arbitrary relations. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 146–159. Springer.

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 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, Leo Yu Zhang, and Yang Liu. 2025. Prompt injection attack against llm-integrated applications. *Preprint*, arXiv:2306.05499.

Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1831–1847.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D Lane, and Mengwei Xu. 2024. Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.

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 (volume 1: Long papers)*, pages 9802–9822.

Barney H Miao, Yiwen Dong, Andreas Theissler, Andrew C Lesh, David J Loftus, and Michael D Lepech. 2024a. Biosys: efficient quality control system for manufacturing of sustainable biopolymer composites. In *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 11–21.

Barney H Miao, Yiwen Dong, Andreas Theissler, Andrew C Lesh, David J Loftus, and Michael D Lepech. 2026a. Ai-powered non-destructive testing for smart manufacturing of carbon-negative biopolymer-bound soil composite. *Communications Engineering*.

Barney H Miao, Robert J Headrick, Zhiye Li, Leonardo Spanu, David J Loftus, and Michael D Lepech. 2024b. Development of biopolymer composites using lignin: A sustainable technology for fostering a green transition in the construction sector. *Cleaner Materials*, 14:100279.

Barney H Miao, Robert J Headrick, Zhiye Li, Leonardo Spanu, David J Loftus, and Michael D Lepech. 2024c. Life cycle assessment and design of lignoblock: A lignin bound block on the path towards a green transition of the construction industry. *Journal of Cleaner Production*, 474:143610.

Barney H Miao, Robert J Headrick, Zhiye Li, Leonardo Spanu, David J Loftus, and Michael D Lepech. 2024d. Life cycle assessment and design of lignoblock: A lignin bound block on the path towards a green transition of the construction industry. *Journal of Cleaner Production*, 474:143610.

Barney H Miao, Andrew C Lesh, David J Loftus, and Michael D Lepech. 2025a. Taking advantage of biological binders to solidify granular material: manufacture and recyclability of lignin-based biopolymer composites. In *Biologically Inspired Materials, Processes, and Systems (BIMPS) 2025*, volume 13430, pages 81–94. SPIE.

Barney H Miao, Daniel Woo, Darius Javan, Edward J Garboczi, Robert J Headrick, Andrew C Lesh, Zhiye Li, David J Loftus, and Michael D Lepech. 2025b. Recycling of lignin-based biocomposites: Improving sustainability and enhancing material strength. *Resources, Conservation and Recycling*, 215:108104.

Barney H Miao, Daniel Woo, Andrew C Lesh, David J Loftus, and Michael D Lepech. 2026b. Solvent selection enables sustainable and affordable lignin biocomposite for cement-free construction. *Cleaner Materials*, page 100371.

Sravan Muguda, Samuel John Booth, Paul Neil Hughes, Charles Edward Augarde, Celine Perlot, Agostino Walter Bruno, and Domenico Gallipoli. 2017. Mechanical properties of biopolymer-stabilised soil-based construction materials. *Géotechnique letters*, 7(4):309–314.

OpenAI. 2024. Chatgpt.

Guanzhong Pan, Vishal Chodnekar, Abinas Roy, and Haibo Wang. 2025. A cost-benefit analysis of on-premise large language model deployment: Breaking even with commercial llm services. *arXiv preprint arXiv:2509.18101*.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics*, pages 113–120.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, science, engineering, and technology (AIxSET)*, pages 166–169. IEEE.

Jay Pujara, Eriq Augustine, and Lise Getoor. 2017. Sparsity and noise: Where knowledge graph embeddings fall short. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1751–1756, Copenhagen, Denmark. Association for Computational Linguistics.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual meeting of the association for Computational Linguistics*, pages 41–47.

Yi Ren, Yanan Zhong, Yang Yang, Hongfeng Huo, Lei Zhang, Jijuan Zhang, Kai Huang, and Zhongfeng Zhang. 2024. Green recyclable biocomposite prepared from lignin and bamboo. *Journal of Cleaner Production*, 449:141710.

Aled D Roberts, William Finnigan, Emmanuel Wolde-Michael, Paul Kelly, Jonny J Blaker, Sam Hay, Rainer Breitling, Eriko Takano, and Nigel S Scrutton. 2019. Synthetic biology for fibers, adhesives, and active camouflage materials in protection and aerospace. *MRS communications*, 9(2):486–504.

Aled D Roberts and Nigel S Scrutton. 2023. Starcrete: A starch-based biocomposite for off-world construction. *Open Engineering*, 13(1).

Ales Deakin Roberts, DR Whittall, Rainer Breitling, Eriko Takano, Jonny J Blaker, Sam Hay, and Nigel S Scrutton. 2021. Blood, sweat, and tears: extraterrestrial regolith biocomposites with in vivo binders. *Materials Today Bio*, 12:100136.

Henning Roedel, Isamar Rosa Plata, Michael Lepech, and David Loftus. 2015. Sustainability assessment of protein-soil composite materials for limited resource environments. *Journal of Renewable Materials*, 3(3):183–194.

Henning Roedel, Isamar Rosa, Maria I Allende, Michael D Lepech, David J Loftus, and Edward J Garboczi. 2019. Prediction of ultimate compressive strength for biopolymer-bound soil composites (bsc) using sliding wingtip crack analysis. *Engineering Fracture Mechanics*, 218:106570.

I Rosa, MD Lepech, and DJ Loftus. 2018. Multiscale modeling and testing of protein-bound regolith and soils. In *16th Biennial International Conference on Engineering, Science, Construction, and Operations in Challenging Environments*, pages 580–590. American Society of Civil Engineers Reston, VA.

Isamar Rosa, Henning Roedel, Maria I Allende, Michael D Lepech, and David J Loftus. 2020. On designing biopolymer-bound soil composites (bsc) for peak compressive strength. *Journal of Renewable Materials*, 8(8):845–861.

Isamar Rosa, Henning Roedel, Michael D Lepech, and David J Loftus. 2015. Creation of statistically equivalent periodic unit cells for protein-bound soils. In *ASME International Mechanical Engineering Congress and Exposition*, volume 57526, page V009T12A058. American Society of Mechanical Engineers.

Henning Schäfer, Ahmad Idrissi-Yaghir, Kamyar Arzideh, Hendrik Damm, Tabea M.G. Pakull, Cynthia S. Schmidt, Mikel Bahn, Georg Lodde, Elisabeth Livingstone, Dirk Schadendorf, Felix Nensa, Peter A. Horn, and Christoph M. Friedrich. 2024. Biokgrapher: Initial evaluation of automated knowledge graph construction from biomedical literature. *Computational and Structural Biotechnology Journal*, 24:639–660.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2025. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Trans. Intell. Syst. Technol.*, 16(6).

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. 2025. When to use graphs in rag: A comprehensive analysis for graph retrieval-augmented generation. *arXiv preprint arXiv:2506.05690*.

Haoze Yu, Haisheng Li, Dianhui Mao, and Qiang Cai. 2020. A relationship extraction method for domain knowledge graph construction. *World Wide Web*, 23(2):735–753.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17744–17761.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2025. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. *arXiv preprint arXiv:2510.10114*.

A Appendix

A.1 Additional performance analysis

Table 4 and 5 together show the workflow performance for simple questions and complex questions. For simple questions, both EGT-KG variants outperform the vanilla RAG workflow in 9 out of 10 comparisons, with ES EGT-KG variants achieving the best result in 8 of them. However, the benefit becomes much less stable on complex questions, suggesting that our framework primarily improves precision in evidence localization for simple fact retrieval questions. For complex questions with multi-hop reasoning and global evidence aggregation, the benefits from finer-grained schemas are reduced. This finding further confirms the limitation discussed before: the one-hop query expansion reduces model's capability in processing complex queries that require multi-hop reasoning or cross-document inference. Notably, the baseline workflow (plain QA with SLMs) achieves relatively higher Final Scores on complex questions than on simple factual questions. A plausible explanation is that the complex questions in our benchmark are broader and more discussion-oriented. Thus, even without external retrieval, SLMs can still obtain partial credit on complex questions by generating generally relevant responses. Table 6 reports the raw scores for the six S3CRF dimensions. As expected, the gains in soundness (+3.03%), correctness (+4.52%), completeness (+2.78%), and relevance (+3.20%) are noticeably larger than conciseness (+2.22%) and fluency (+1.71%). It suggests that the main benefit of our framework comes from improving core answer-quality dimensions related to reasoning, factual accuracy, coverage, and query alignment, rather than surface-level language style.

A.2 Component ablation

As Table 8 shows, no single component makes a dominant contribution to the overall improvement, because its measured effect flips between the judge models. To be specific, the re-ranker carries the effect under *gemini-2.5-pro* (+2.03 over query expansion), whereas the same addition is negative under *gpt-4o* (-0.66), and the evidence window on its own contributes essentially nothing over query expansion under either judge (+0.02 and -0.03). The stable partial gains is C (query expansion + rerank): because evidence re-ranking operates on the candidate chunks that query expansion supplies, the two components together consistently improve over vanilla RAG under both judges (+2.61 from *gpt-4o*, +3.81 from *gemini-2.5-pro*), even though the marginal contribution of re-ranking alone changes sign. The assembled pipeline is therefore the best configuration to deploy. The full framework is the highest-scoring configuration under both judges (A $\rightarrow$ E: +4.29 from *gpt-4o*; +3.86 from *gemini-2.5-pro*), and no partial configuration beats it. The remaining improvements come from the interaction between evidence re-ranking and the evidence window. The mechanism is straightforward: The evidence window can suppress noise and highlight the most relevant sentence. But if only the evidence window is deployed, when the gold evidence is not present in the retrieved candidate chunks, the answer model is still stuck for lack of ground truth — which is exactly why the evidence window alone changes nothing. Once re-ranking has moved the chunk containing the gold evidence into the final top-k, the evidence window pays off: it removes the unnecessary surrounding text, cutting mean context by 22% ($8,196 \rightarrow 6,399$ characters) at no cost in score, and lets the answer model focus on the evidence that matters.

A.3 Supplementary QASPER Evaluation

To further study whether the EGT-KG framework generalizes beyond the BSC corpus, we conduct a supplementary evaluation on QASPER (Dasigi et al., 2021). QASPER is a scientific question-answering benchmark built from NLP research papers. Each question is associated with a source paper and is answered using evidence from that paper.

The official QASPER evaluation protocol provides the target paper ID and restricts retrieval to chunks from that paper. This per-paper setting is appropriate for evaluating reading comprehension once the relevant paper is known, but it does not fully reflect many practical scientific RAG scenarios, where a user asks a question without specifying which paper contains the answer. We therefore use a cross-paper retrieval setting: the system is not given the gold paper ID and must retrieve evidence from all papers in the development split. In our setup, retrieval is performed over 5,253 chunks from 281 papers, and evaluation is conducted on all 1,005 development questions.

We compare three settings. The first is a question-only baseline, where the models answer questions without retrieved context. The second is vanilla cross-paper RAG, which retrieves chunks

	Gemma2		Llama3		Llama3.1		Mistral		Qwen2.5	
	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini
Baseline	22.36	15.93	17.28	12.73	18.73	17.45	16.20	13.55	22.36	16.63
Vanilla-RAG	69.40	72.06	56.56	62.40	88.70	82.56	58.22	77.68	**89.00**	85.80
EGT-KG(AS)	74.58	78.18	75.96	74.70	89.72	84.62	57.30	**79.00**	83.86	86.21
EGT-KG(ES)	**77.36**	**79.98**	**78.72**	**78.00**	**89.78**	**86.52**	**58.88**	75.90	88.42	**87.00**

Table 4: Performance comparison across answer models and judge models for simple questions ($question_1$)

	Gemma2		Llama3		Llama3.1		Mistral		Qwen2.5	
	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini	gpt-4o	gemini
Baseline	49.55	33.28	38.76	36.88	39.65	32.45	41.10	31.03	48.02	32.23
Vanilla-RAG	64.12	**69.10**	**66.18**	**62.48**	79.90	72.10	60.42	**77.42**	**83.02**	70.38
EGT-KG(AS)	**69.86**	67.06	64.78	57.52	77.78	72.26	**63.38**	76.56	81.78	**74.98**
EGT-KG(ES)	68.88	68.10	58.92	58.42	**82.26**	**76.92**	62.26	72.56	81.98	72.82

Table 5: Performance comparison across answer models and judge models for complex questions ($question_2$)

	Soundness		Correctness		Completeness		Conciseness		Relevance		Fluency	
	gpt	gemini	gpt	gemini	gpt	gemini	gpt	gemini	gpt	gemini	gpt	gemini
Baseline	29.0	18.6	19.4	15.1	29.0	18.6	60.1	58.3	33.1	23.9	75.8	89.2
Vanilla-RAG	70.3	71.5	68.8	69.7	60.3	50.8	83.8	91.8	78.2	86.6	88.5	**96.7**
EGT-KG(AS)	71.8	**74.1**	72.2	71.2	**61.4**	51.8	86.7	**92.9**	81.1	89.5	90.6	96.5
EGT-KG(ES)	**72.4**	73.9	**73.9**	**72.2**	**61.4**	**53.6**	**86.8**	91.0	**82.0**	**90.1**	**91.2**	**96.7**

Table 6: Dimensional performance comparison across workflows.

Setting	**gemma2**	**qwen2.5**	**llama3.1**	**llama3**	**mistral**	**Mean**
Q-only baseline	0.1338	0.1370	0.1471	0.1231	0.1253	0.1333
Vanilla RAG	0.1432	0.1590	0.1729	**0.1737**	0.1449	0.1587
EGT-KG	**0.1554**	**0.1641**	**0.1840**	0.1710	**0.1659**	**0.1681**
EGT-KG vs. Vanilla	+8.5%	+3.2%	+6.4%	-1.6%	+14.5%	+5.9%

Table 7: Supplementary QASPER results. Answer F1 is reported over the full development split of 1,005 questions.

Judge	**A** vanilla dense	**B** A + query expansion	**C** B + rerank	**D** B + window	**E** full
gpt-4o	72.01	75.28 (+3.27)	74.62 (-0.66)	75.25 (-0.03)	**76.30** (+4.29)
gemini-2.5-pro	66.60	68.38 (+1.78)	70.41 (+2.03)	68.40 (+0.02)	**70.46** (+3.86)

Table 8: Component ablation, pooled Final Score over the five answer models (40 questions $\times$ 5 sampling rounds, $n = 200$ per cell). A is vanilla dense retrieval ($\alpha = 1.0$, no re-ranking, no evidence window); B adds knowledge-graph query expansion ($\alpha = 0.7$); C and D add evidence re-ranking and the evidence window to B, respectively. Parentheses give the increment over A for B and E, and over B for C and D.

Retrieval pool	Vanilla RAG	EGT-KG	Δ
Official distractor (10)	0.4857	0.5015	+0.016
Pooled corpus (1,977)	0.3165	0.4585	**+0.142**

Table 9: HotpotQA Answer F1 under two retrieval regimes, 200 development questions, *gemma2:9b*.

from the full cross-paper pool using dense retrieval. The third is our EGT-KG framework, which augments retrieval with KG-expanded query terms, evidence-aware reranking, and evidence-window

construction. We report Answer F1 using the official QASPER evaluator.

The results show that retrieval is essential in this more realistic cross-paper setting, where vanilla RAG improves mean Answer F1 from 0.1333 to 0.1587 (19.0%) over the question-only baseline. EGT-KG further improves mean Answer F1 to 0.1681, a 5.9% relative gain over vanilla RAG, with positive gains for four out of five tested models. Additionally, we performed a simple test on official per-paper setting on *gemma2:9b* ($\alpha = 0.9, \lambda = 0.5, k = 8$). The EGT-KG framework (0.367) outperform vanilla RAG (0.334) by 10.2%. The improvement on scores are derived from more relevant chunk candidates for both workflows.

A.4 Cross-domain evaluation on HotpotQA

We evaluated on 200 HotpotQA development questions for generalization, inducing the relation schema from that corpus's own relation-label distribution.

The outcome depends entirely on how large a candidate pool the retrieval has to search. In the official distractor setting, each question is supplied with ten candidate paragraphs with two gold paragraphs, and the framework adds nothing ($+0.016$): dense retrieval already finds the gold paragraphs, so there is no localization problem left to solve. Pooling the same questions' paragraphs into a single 1,977-paragraph corpus reverses this ($+0.142$). Growing the pool from 10 to 1,977 candidates costs vanilla RAG 34.8% of its Answer F1 and EGT-KG 8.6%, and the complete gold-paragraph set is retrieved for 36.0% of questions by vanilla RAG against 56.0% by EGT-KG.

A.5 Small-corpus construction setting

Another practical advantage of our framework is the feasibility of being built on a small corpus in a limited resource scenario. Rather than requiring large-scale indexing or computationally intensive global graph structure, our EGT-KG framework constructs a knowledge graph from a compact benchmark of 30 papers. This design decision mimics the real-world situation in the emerging scientific domain of question-answering, where users may only have access to a small collection of relevant papers and may not be willing to afford the cost of heavy knowledge graph construction. Thus, our contribution to EGT-KG is to improve scientific QA while preserving the possibility of practical deployment. The experiment results show that even the lightweight knowledge graph, which is further trimmed by type schema, can provide useful and explainable retrieval to enhance response generation for scientific QA.

A.6 Introduction to BSC

Biopolymer-bound soil composite (BSC) is a nature-inspired material that valorizes waste-stream biopolymers from major industries (e.g., paper pulping, biofuels, and agriculture) as binders for granular materials. Unlike conventional materials such as ordinary Portland cement concrete, BSC is potentially carbon-negative due to its use of carbon-rich biopolymers, enabling sequestration of carbon in the composite (Miao et al., 2024c). It is also fully recyclable, reduces construction waste, and offers compressive strength comparable to concrete, making it a promising low-cost, sustainable, cement-free alternative for the built environment.

A.7 Detailed Pipeline Diagrams

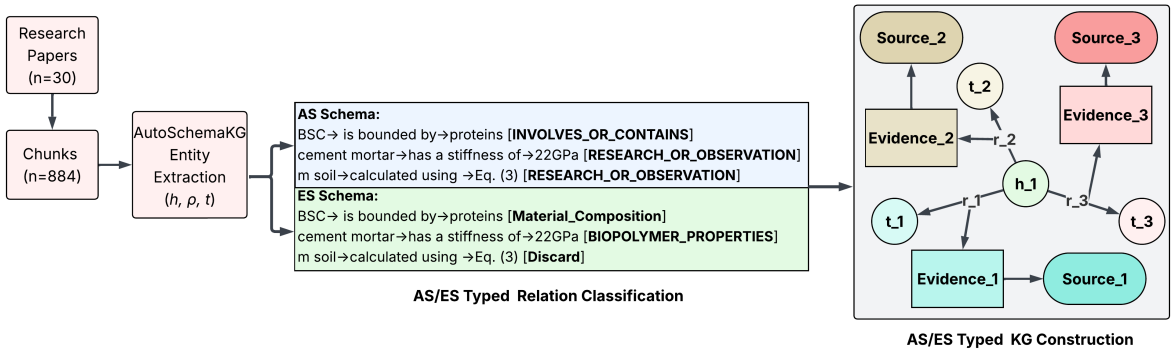

Figure 2: Stage 1 of EGT-KG: triples extraction, relation classification, and KG construction.

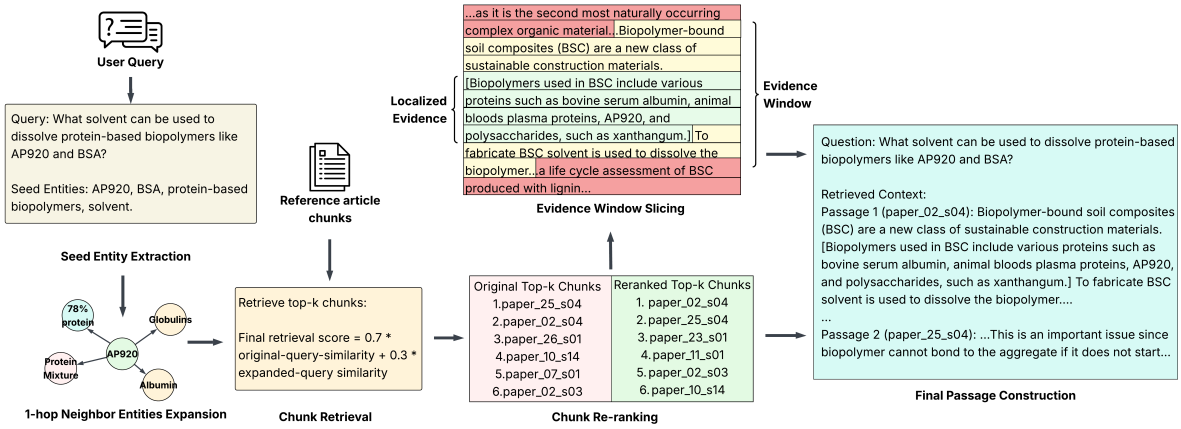

Figure 3: Stage 2 of EGT-KG: query expansion, context retrieval and re-ranking, evidence-aware context slicing, and passage construction.

A.8 Prompts

Entity Extraction Prompt

```
{
    "en": {
        "system":
"You are an expert knowledge graph constructor specializing in materials science, civil engineering, and
biopolymer-bound soil composites (BSC). You always respond with a valid JSON array and nothing else.",
        "triple_extraction":
"You are extracting entity-relation triples from academic text about Biopolymer-bound Soil Composites (BSC) and
related materials science topics.

Task: Extract all important entity-relation triples from the given passage. For each triple, identify the head
entity, tail entity, and the relationship between them. Use a concise, descriptive relation label that accurately
captures the nature of the relationship.

Output Format: Return a JSON array of objects. Each object must have exactly three keys:
\"Head\": the source entity (a specific noun or noun phrase, not a pronoun)
\"Relation\": must follow the format \"relation_label | brief evidence from text that is summarized or directly cited
from the original text\."
relation_label: a concise phrase describing the relationship (e.g., \"is defined as\", \"increases\", \"is composed
of\", \"was tested using\", \"causes\", \"is located in\") After the pipe symbol (|), include a brief quote or
paraphrase from the source text that supports this triple. Prefer direct citation from the original text when
possible; otherwise provide a concise summary that stays faithful to the original wording
\"Tail\": the target entity (a specific noun or noun phrase, not a pronoun)
Example Output:
 {\"Head\": \"BSC\", \"Relation\": \"is defined as | biopolymer-bound soil composites are a mixture of granular soil
aggregate, biopolymer, and water\", \"Tail\": \"mixture of soil aggregate, biopolymer and water\"}......

Constraints: Extract ALL relevant triples from the passage. Head and Tail must be specific entities, not pronouns or
vague references. The relation_label should be a concise, natural-language description of the relationship. The
evidence after | should be a brief supporting quote or paraphrase. Prefer direct citation from the original text. If
no triples can be extracted, return exactly: []. Do not include any text other than the JSON array\n\nHere is the
passage.
    }
}
```

Figure 4: Entity extraction prompt

RAG QA Prompt for all workflows

```
SYSTEM_PROMPT = (
    "You are a research assistant specializing in Biopolymer-bound Soil Composites (BSC) and materials science. You
must answer strictly from the provided context from research papers. If the context is insufficient, state it
clearly. Do not add unverified knowledge or infer missing details."
)

USER_PROMPT = """Answer the following question using ONLY the context provided below.

## Question
{question}

## Retrieved Context

{context_block}

## Instructions
- Your first sentence must directly answer the question.
- Do not discuss topics the question did not ask about.
- Use only information explicitly supported by the retrieved context.
- Do not infer or add missing details.
- If one passage is sufficient to answer the question, prioritize it and ignore less relevant passages.
- Cite only the most relevant paper ID(s) for key claims.
- Be concise: 2-5 sentences is ideal.
- If the retrieved passages do not contain enough information, say so clearly.
"""
```

Figure 5: RAG QA prompt for all workflows, containing original question, retrieved context, and instructions

AS Schema Prompt

```
System: You are an expert knowledge graph designer specializing in materials science, civil engineering, and
biopolymer-bound soil composites (BSC). You always respond with a valid JSON array and nothing else.

Task: Propose 20 relation types that together cover the relationships expressed in this literature. A relation type
is a broad, semantically coherent category into which many specific relation labels (for example "is used for",
"increases", "is a type of") can be grouped.

Requirements
- Each type must be broad enough to absorb many surface forms, and narrow enough to be distinguishable from the
others.
- The types should be mutually exclusive: a given relation should fall naturally into exactly one.
- Together they should cover the relationships a researcher would need in order to reason about this material:
composition, causation, procedure, measurement, definition, and comparison, among others.
- Name each type in UPPER_SNAKE_CASE.
- Do not include a catch-all or "other" category.
Output Format: Return a JSON array of 20 objects and nothing else:
[{"type": "TYPE_NAME", "description": "one line stating what belongs in
  this type", "examples": ["example relation label", "another label"]}, ...]
```

Figure 6: Prompt used to induce the AS relation schema.

A.9 Parameter sensitivity

λ	0.40	0.50	0.65	0.80
Final Score	66.47	68.37	**73.54**	69.12
k	3	6	10	15
Final Score	63.80	**71.40**	68.20	66.30

Table 10: Sensitivity to the re-ranking weight λ and to the number of retrieved chunks k. Reported results use $\lambda = 0.65$ and $k = 6$.

A.10 Performance by relation type

Relation type	KG%	Ret%	ratio	Δ
Causes or Influences	36.4	34.5	0.95	+3.96
Research Observations	17.1	16.2	0.95	+4.98
Material Composition	11.9	12.0	1.01	+2.03
Manufacture Procedure	10.3	9.8	0.95	+4.72
Logical Reasoning	7.5	7.7	1.03	+2.83
Definition	5.7	8.4	1.47	+1.70
Experimental Methods	5.0	5.8	1.16	+5.03
Biopolymer Properties	3.8	2.4	0.63	+5.49
Temporal Relation	1.0	1.2	1.20	+5.59
Design Procedure	0.9	1.5	1.67	−0.20
Experimental Procedures	0.4	0.5	1.25	+2.78

Table 11: Performance by relation type under the ES schema. KG% is the type's share of the graph, Ret% is the share of the evidence actually retrieved. Δ is the Final Score improvement over vanilla RAG.

A.11 Distribution of relation types in the reified knowledge graph

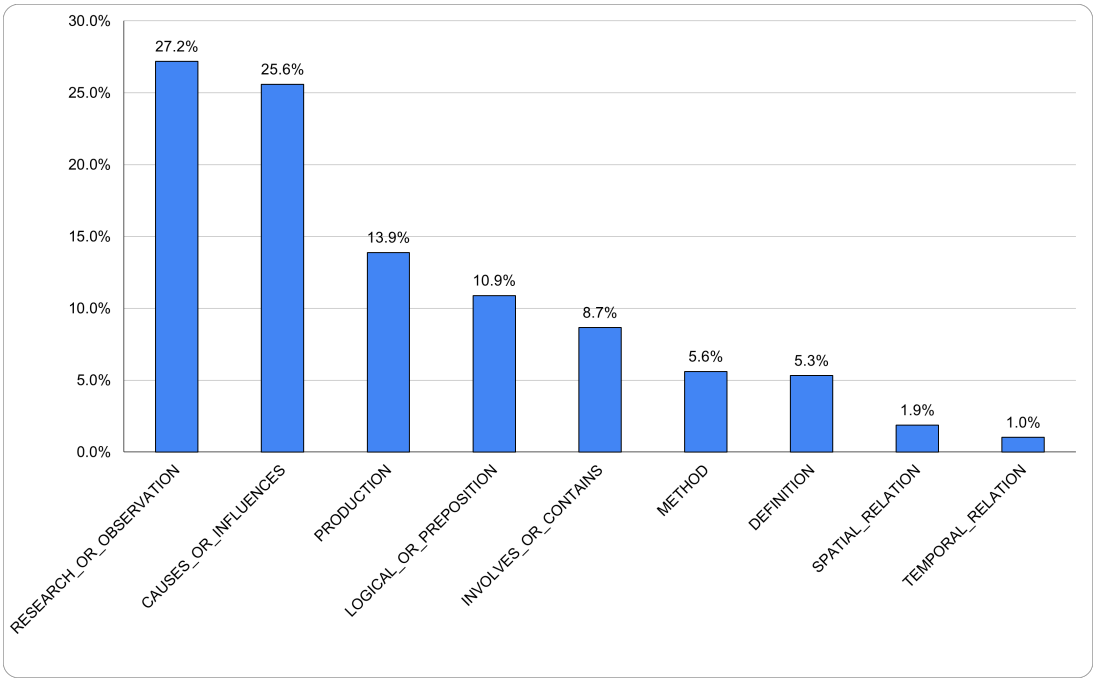

Figure 7: EGT-KG(AS) relation schema distribution

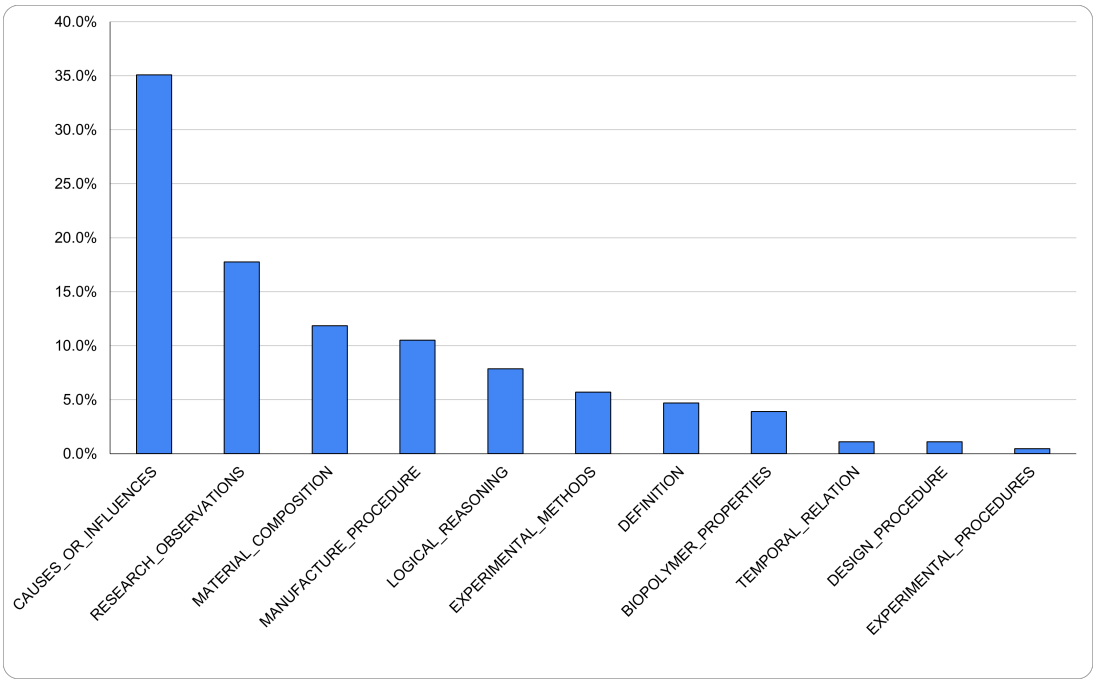

Figure 8: EGT-KG(ES) relation schema distribution

A.12 Evaluation Criteria

Metric	Definition	Primary focus
Soundness	Logical validity of reasoning from question to answer, including appropriate use of referenced material.	Reasoning validity
Correctness	Factual accuracy compared to the reference material; hallucinated facts reduce the score.	Factual grounding
Completeness	Coverage of all relevant aspects of the question.	Information coverage
Conciseness	Appropriate detail and brevity without unnecessary verbosity.	Brevity and clarity
Relevance	Alignment with the user's query.	Query alignment
Fluency	Grammatical, coherent, and readable language.	Readability

Table 12: S3CRF evaluation metrics.

Dimension	Score	Description
Soundness $w_{snd} = 0.275$	0	Incorrect logical connection between question and answer.
	1	Poor reasoning, conclusion is not directly supported.
	2	Weak reasoning, several logical flaws in answer.
	3	Acceptable reasoning with partially unsupported claims.
	4	Strong reasoning with minor gaps.
	5	Perfect logical reasoning with proper use of evidence.
Correctness $w_{cor} = 0.275$	0	Incorrect or fabricated response.
	1	Significant factual errors or hallucinations.
	2	Multiple minor factual errors or hallucinations.
	3	Generally correct with acceptable inaccuracies.
	4	Mostly accurate with trivial errors.
	5	Accurate and verifiable answer supported by reference documents.
Completeness $w_{cmp} = 0.15$	0	Fails to address the question.
	1	Incomplete response with superficial coverage.
	2	Missing coverage of important information.
	3	Addresses main point but misses some details.
	4	Covers most important aspects.
	5	Comprehensively addresses all aspects.
Conciseness $w_{con} = 0.05$	0	Completely unnecessary length.
	1	Severely verbose or oversimplified.
	2	Relatively too long/short, affecting readability.
	3	Somewhat verbose or oversimplified with acceptable readability.
	4	Mostly appropriate length with minor issues.
	5	Perfect balance of detail and brevity.
Relevance $w_{rel} = 0.2$	0	Completely off-topic response.
	1	Mostly irrelevant with trivial relevant discussion.
	2	Partially relevant, significant off-topic content.
	3	Mostly relevant but some off-topic content.
	4	Highly relevant with minor off-topic content.
	5	Directly and precisely addresses the question.
Fluency $w_{flu} = 0.05$	0	Incoherent.
	1	Poor organization, very hard to read.
	2	Choppy expression or hard to follow.
	3	Acceptable expression with some disjointed sections.
	4	Good coherence with minor awkwardness.
	5	Excellent flow, reads naturally.

Table 13: Rubric for S3CRF evaluation (0-5).

A.13 Simple/complex questions in the question set and papers used in corpus

ID	Question
Q1	What solvent can be used to dissolve protein-based biopolymers like AP920 and BSA?
Q2	Give me an overview of the processes needed to dissolve the biopolymers (e.g. time and any other considerations)
Q3	What is the maximum amount (in percentage) of biopolymer that can be dissolved into solution?
Q4	What parameter do we need to consider in order to maximize the amount of biopolymer in the composite?
Q5	What type of compaction method is used to manufacture BSC? Why is this important?
Q6	Why is BSC a sustainable building material?
Q7	What happens with saturation ratios more than 1.0?
Q8	How do we desiccate BSC?
Q9	What is the porosity of BSC?
Q10	How can I maximize the strength of these biocomposites?
Q11	What biopolymers can be used to make BSC?
Q12	What is the microstructure of BSC?
Q13	How can the structure of BSC be modeled?
Q14	How can I prevent shrinkage cracking in BSC?
Q15	How can BSC structure be imaged?
Q16	How can BSC be recycled?
Q17	How is the strength of BSC characterized?
Q18	What form of sample is used for BSC characterization?
Q19	How can the sustainability of BSC be improved?
Q20	How does the use of BSA contrast with AP920?

Table 14: Simple question set used in QA evaluation.

ID	Question
Q1	How can I select the most optimal solvent to dissolve a given biopolymer? How can I choose a solvent that not only effectively dissolves the biopolymer but also enhances the resulting bridge strength or interfacial bond strength between the aggregate and the biopolymer binder?
Q2	How can each step in the manufacture of BSC be modified to improve both manufacturability and the overall macroscale properties of the composite? What potential issues could arise at each step, and how might they affect the final properties of the manufactured composite?
Q3	Explain the relationship between biopolymer concentration and biopolymer solution viscosity during dissolution, and how the "workability limit" physically constrains the maximum binder content.
Q4	If higher biopolymer content typically results in a stronger BSC, could we reduce the amount of compaction applied to the specimens (i.e., increase the void space in the material)? Would this reduction in compaction and the corresponding increase in void space, which the biopolymer can fill, affect any other material or design parameters?
Q5	How can higher levels of compaction enhance the compressive strength of BSC, and what are the resulting impacts on design parameters?
Q6	In comparison to hydrophilic BSCs which use water as a solvent, does BSC using hydrophobic biopolymers with a higher potential for carbon sequestration (they are more carbon rich) lead to always having a lower carbon footprint?
Q7	If a novel biopolymer is being used and its particle density is not yet characterized, is it preferable to prepare an initial mix above or below the saturation ratio? What are the potential advantages and disadvantages of each approach, and which strategy would minimize risk in terms of workability, mechanical performance, and consistency of the resulting BSC?
Q8	How do the desiccation requirements differ between protein-based BSCs (which use water as a solvent) and lignin-based BSCs (which use DMSO as a solvent)? If the desiccation time for a BSC using a particular solvent is excessively long, what strategies can be employed to accelerate the process without compromising the composite's mechanical integrity or interfacial bonding?
Q9	Can microscale characterization techniques, such as X-ray micro-CT, be used to accurately determine the porosity of BSC? How can the porosity of BSC be increased without compromising its mechanical or physical properties?
Q10	Rank the most effective methods for improving the compressive strength of biopolymer-bound soil composites, and provide a detailed explanation for each method. In your evaluation, consider not only the potential gains in compressive strength but also factors such as practicality, cost-effectiveness, scalability, and environmental footprint.
Q11	Which biopolymers are most promising in terms of environmental impact, availability, and practicality. If I want to explore or develop a new variant of BSC, what are some promising biopolymers that can be used?
Q12	What phenomena can be observed from microstructural images of BSC (phenomenon and type of image).
Q13	What types of data are required to develop an accurate model of BSC behavior? Which material, structural, and processing parameters should be prioritized for measurement in future investigations to improve model reliability and predictive capability?
Q14	Does a staggered desiccation approach, in which the BSC is first desiccated under ambient conditions and then oven-dried directly reduce the formation of shrinkage cracks? Please hypothesize how this potential approach can be designed (number of steps, time per step, conditions at each step)
Q15	Can SEM be used in the same capacity/function as Micro-CT in calculating statistical descriptors (lineal path function and two-point probability correlation functions).
Q16	Explain the process of recycling BSCs at the end of their lifecycle to reclaim the biopolymer. Also hypothesize a more effective way for recycling other than this approach.
Q17	If I wanted to perform a study on the strength of a biopolymer bridge between two aggregate particles how would I go about this? Propose a plan.
Q18	For macroscale considerations, are there any limits or standards for minimum size of a test specimen? What are these standards based on?
Q19	Propose the most viable and potential studies that can be carried out. When proposing plans or studies, please consider how we can improve the affordability or accessibility of existing and future variants of BSC .
Q20	Is BSA or AP920 more practical for construction in extraterrestrial environments, both in the short term and the long term?

Table 15: Complex question set used in QA evaluation.

ID	Paper name	Relation to questions
1	On Designing Biopolymer-Bound Soil Composites (BSC) for Peak Compressive Strength (Rosa et al., 2020)	Directly related
2	Development of biopolymer composites using lignin: A sustainable technology for fostering a green transition in the construction sector (Miao et al., 2024b)	Directly related
3	Sustainability Assessment of Protein-Soil Composite Materials for Limited Resource Environments (Roedel et al., 2015)	Directly related
4	Multiscale Modeling and Testing of Protein-Bound Regolith and Soils (Rosa et al., 2018)	Directly related
5	Green recyclable biocomposite prepared from lignin and bamboo (Ren et al., 2024)	General Material
6	Life cycle assessment and design of LignoBlock: A lignin bound block on the path towards a green transition of the construction industry (Miao et al., 2024d)	Directly related
7	Recycling of lignin-based biocomposites: Improving sustainability and enhancing material strength (Miao et al., 2025b)	Directly related
8	Taking advantage of biological binders to solidify granular material: manufacture and recyclability of lignin-based biopolymer composites (Miao et al., 2025a)	Directly related
9	Solvent selection enables sustainable and affordable lignin biocomposite for cement-free construction (Miao et al., 2026b)	Directly related
10	Toward Cement-Free Construction: Microstructural Evaluation of Engineered Biopolymer Composite	Directly related
11	Hypervelocity Impact Performance of Biopolymer-Bound Soil Composites for Space Construction (Allende et al., 2020)	Not related to questions
12	Determining the yield stress of a Biopolymer-bound Soil Composite for extrusion-based 3D printing applications (Biggerstaff et al., 2021)	Not related to questions
13	A shape stability model for 3D printable biopolymer-bound soil composite (Biggerstaff et al., 2022a)	Not related to questions
14	Determining the structuration of biopolymer-bound soil composite (Biggerstaff et al., 2022b)	Not related to questions
15	A Study on the Flow Behavior and Thixotropy of Biopolymer-Bound Soil Composite (Biggerstaff et al., 2025)	Not related to questions
16	BioSys: Efficient Quality Control System for Manufacturing of Sustainable Biopolymer Composites (Miao et al., 2024a)	Not related to questions
17	AI-powered non-destructive testing for smart manufacturing of carbon-negative biopolymer-bound soil composite (Miao et al., 2026a)	Not related to questions
18	Prediction of ultimate compressive strength for biopolymer-bound soil composites (BSC) using sliding wingtip crack analysis (Roedel et al., 2019)	Not related to questions
19	Creation of Statistically Equivalent Periodic Unit Cells for Protein-Bound Soils (Rosa et al., 2015)	Not related to questions
20	A comprehensive review of various biopolymer composites and their applications: From biocompatibility to self-healing (Kumar et al., 2023)	General Material
21	Synthetic biology for fibers, adhesives, and active camouflage materials in protection and aerospace (Roberts et al., 2019)	General Material
22	Blood, sweat, and tears: extraterrestrial regolith biocomposites with in vivo binders (Roberts et al., 2021)	Similar materials to BSC
23	Effects of Xanthan gum biopolymer on soil strengthening (Chang et al., 2015a)	Similar materials to BSC
24	Evaluating the physical characteristics of biopolymer/soil mixtures (Ayeldeen et al., 2016)	Similar materials to BSC
25	Geotechnical behavior of a beta-1,3/1,6-glucan biopolymer-treated residual soil (Chang and Cho, 2014)	Similar materials to BSC
26	Application of Microbial Biopolymers as an Alternative Construction Binder for Earth Buildings in Underdeveloped Countries (Chang et al., 2015b)	Similar materials to BSC
27	Improving Mechanical Properties of Sand Using Biopolymers (Khatami and O'Kelly, 2013)	Similar materials to BSC
28	Mechanical properties of biopolymer-stabilised soil-based construction materials (Muguda et al., 2017)	Similar materials to BSC
29	Seaweed biopolymers as additives for unfired clay bricks (Dove et al., 2016)	Similar materials to BSC
30	StarCrete: A starch-based biocomposite for off-world construction (Roberts and Scrutton, 2023)	Similar materials to BSC

Table 16: Papers included in the corpus and their relation to the question sets.