KG-LLM Workshop / LREC 2026

Efficient KG-Augmented RAG with Reusable Graph Community Summaries

Maha Karkout, Maria Khodorchenko, Nikolay Butakov, Denis Nasonov

ITMO University, Saint Petersburg, Russia

1 Research Background

- Dense RAG works well for localized factual queries but struggles with multi-section and cross-document evidence integration.
- GraphRAG adds relational structure through entities, relations, and communities, but full query-time graph reasoning is slow and can suppress fine-grained evidence.
- We propose a practical hybrid KG-RAG design that builds a knowledge graph offline, converts communities into reusable summaries, and retrieves these summaries together with text at inference time.
- Benchmarks: QASPER (document-bounded scientific QA) and ObliQA (cross-document regulatory QA).

2 Method

- Offline indexing: ontology-guided entity extraction, relation extraction, graph assembly, community detection, and community-level summarization.
- Online inference: retrieve Top-20 text segments and Top-5 community summaries, then answer from a unified context.
- QASPER uses per-paper graphs; ObliQA uses a global regulatory graph.

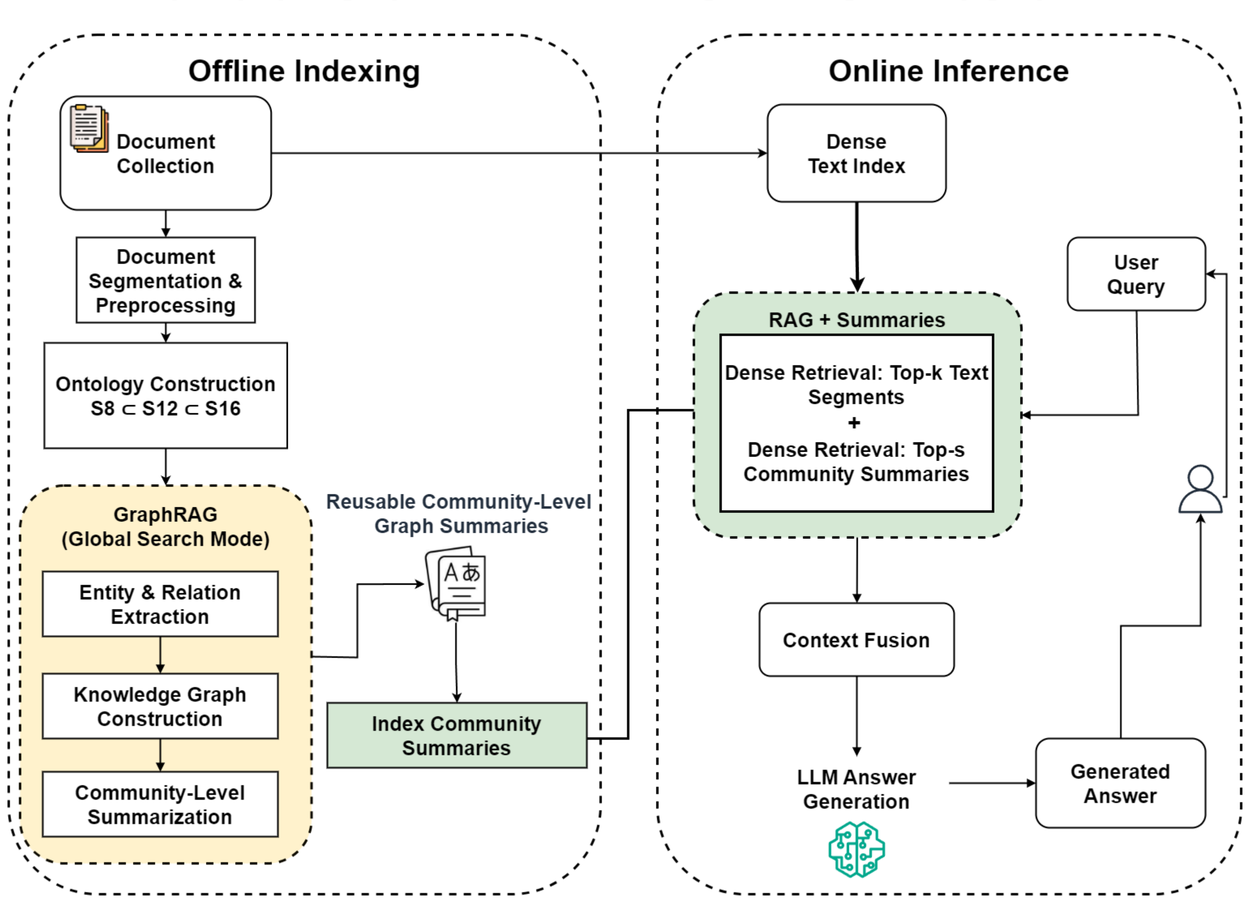

Fig. 1. KG-RAG pipeline: offline graph construction and community summarization, online retrieval of text segments and community summaries for answer generation.

3 Results & Discussion

Table 1. Main QA Performance (Overall score)

Dataset	Baseline RAG	GraphRAG	Hybrid KG-RAG
ObliQA	2.40	2.41	2.59
QASPER	2.41	2.13	3.14

Overall = mean of relevance, correctness, and completeness.

Table 2. Mean Query Latency (s/question)

Dataset	GraphRAG	Hybrid KG-RAG
ObliQA	311.29	11.94
QASPER	70.25	5.04

Key Takeaways

- Hybrid KG-RAG achieves the highest overall scores on both benchmarks.
- It avoids expensive query-time graph inference and remains much faster than full GraphRAG.
- On QASPER, the hybrid configuration shows a clear right-shift in score distributions and mostly positive per-question deltas.

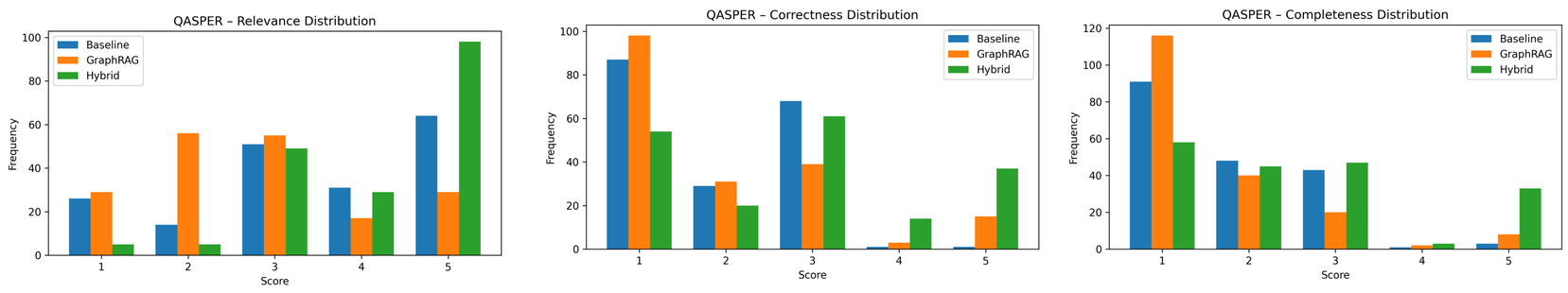

Fig. 2. Score distributions on QASPER (failure subset).

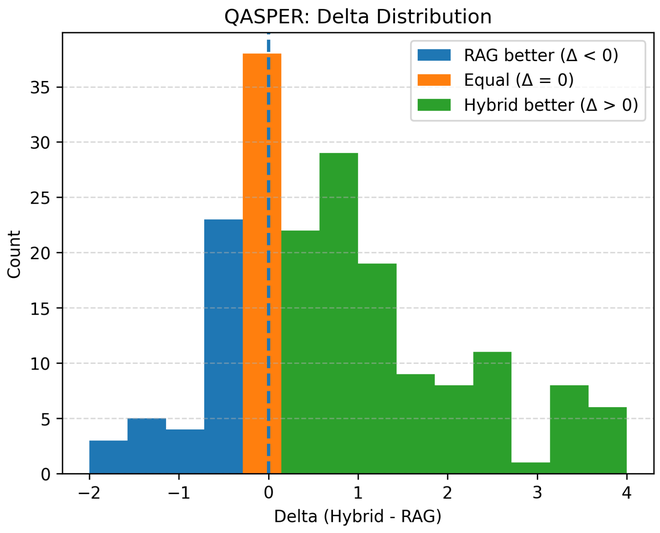

Fig. 3. Distribution of per-question overall-score deltas on QASPER (Hybrid – RAG).

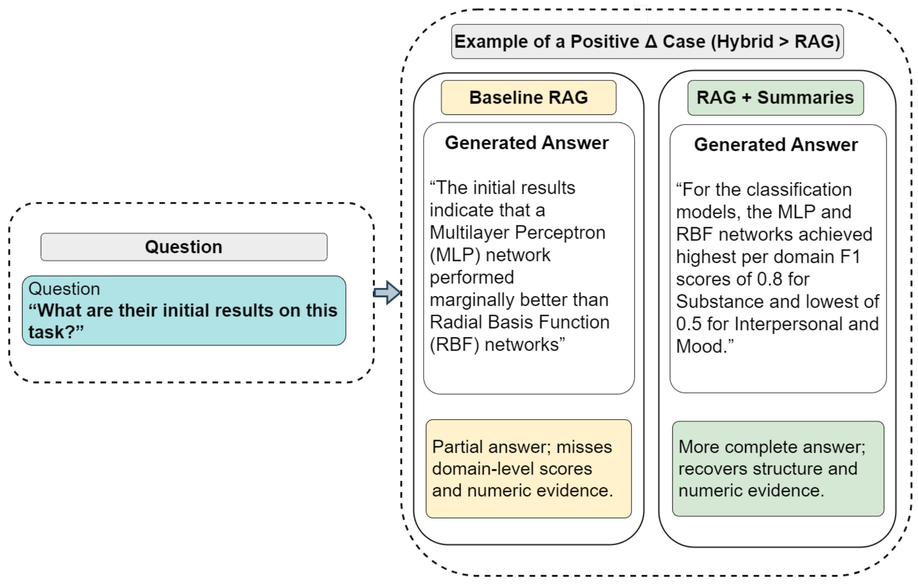

Fig. 4. Example positive case: the hybrid configuration recovers answer structure and quantitative details missing from baseline RAG.

4 Conclusions

Graph structure alone does not consistently guarantee better answers.

Reusable community summaries work best when they enrich dense retrieval rather than replace it.

Hybrid KG-RAG provides the best quality-efficiency trade-off across both retrieval paradigms.

Selected References

- Lewis et al. 2021; Edge et al. 2025.
- Dasigi et al. 2021; Gökhan et al. 2024.

This work was supported by the Russian Science Foundation, agreement no. 24-71-00115, https://rscf.ru/en/project/24-71-00115/.