FAQ-RAG: FAQ-Centric Vector Storage for Trustworthy RAG for QA Tasks

Praneeth Vadlapati
Independent researcher
praneethv@arizona.edu
ORCID: 0009-0006-2592-2564

Abstract: Retrieval-Augmented Generation (RAG) has become a dominant paradigm for grounding large language models (LLMs) in external knowledge. However, prevailing RAG systems remain fundamentally text-centric, relying on arbitrary document chunking strategies that often misalign with user intent, obscure provenance, and introduce inefficiencies at query time. This paper introduces FAQ-RAG, a structured alternative that treats questions and answers as the atomic units of knowledge. FAQ-RAG transforms documents into exhaustive sets of frequently asked questions paired with grounded answers, each stored as a dual-vector representation capturing both intent and substance. By indexing knowledge rather than raw text, FAQ-RAG improves retrieval precision, recall, and citation fidelity while reducing reasoning overhead during inference. We present the conceptual framework, system architecture, and practical advantages of FAQ-RAG, and situate it within the broader landscape of retrieval-augmented and citation-aware question answering systems.

The source code is available at github.com/Pro-GenAI/FAQ-RAG.

Keywords: Large Language Models (LLMs), Retrieval-Augmented Generation, Question Answering, Vector Databases, Knowledge Representation

I. Introduction

Large language models have demonstrated remarkable capabilities in natural language understanding and generation, yet their tendency to hallucinate and their lack of inherent access to up-to-date or proprietary knowledge have motivated the development of Retrieval-Augmented Generation systems. In RAG, an external retrieval component supplies relevant context to the model at inference time, enabling grounded responses. Despite widespread adoption, most existing RAG pipelines remain rooted in text-centric design choices, particularly the practice of embedding and retrieving fixed-size text chunks. This paper argues that such designs are misaligned with the fundamental objective of question answering and proposes FAQ-RAG as a question-centric alternative that more naturally supports accurate, efficient, and auditable QA.

A. Limitations of existing approaches

Traditional RAG systems typically segment documents into overlapping or non-overlapping chunks based on heuristics such as token count or sentence boundaries. While effective for generic information retrieval, this approach exhibits structural weaknesses for question answering. Chunks frequently fail to contain complete answers, forcing systems to retrieve multiple fragments and rely on downstream reasoning to synthesize responses. Query embeddings are only indirectly aligned with the stored text, leading to brittle matching behavior. Moreover, citations produced by chunk-based systems are often approximate, as retrieved text may partially support an answer without clearly delineating its provenance. These limitations become particularly problematic in regulated, scientific, or legal domains where traceability and auditability are essential.

B. Proposed approach

FAQ-RAG reframes the retrieval problem by explicitly modeling questions and their answers as first-class knowledge objects. Instead of indexing raw text, the system performs an offline transformation of documents into structured FAQs, each consisting of a question, a grounded answer derived strictly from the source content, and precise metadata indicating document and page-level provenance. Each FAQ is embedded twice, once based on the question and once based on the answer, enabling flexible retrieval that captures both user intent and semantic substance. By shifting complexity to ingestion time, FAQ-RAG minimizes reasoning requirements at query time and ensures that every retrieved unit is already a validated answer.

C. Applications

The FAQ-RAG paradigm is particularly well suited to domains where correctness, explainability, and citation fidelity are critical. Academic and scientific question answering benefits from explicit page-level provenance. Legal and compliance systems require deterministic sourcing and reduced hallucination risk. Financial research, medical documentation, and enterprise knowledge bases similarly demand precise alignment between queries, answers, and underlying sources. By design, FAQ-RAG supports these requirements without introducing complex orchestration or bespoke infrastructure.

D. Related work

Prior work on RAG has largely focused on improving retrieval quality through better chunking strategies, hybrid sparse-dense retrieval, or more powerful reranking models. Citation-aware QA systems have explored mechanisms for associating generated answers with supporting documents, often through post hoc attribution or constrained decoding. Question decomposition and reasoning-based RAG approaches attempt to overcome chunk limitations by performing multi-step inference at query time, albeit at significant computational cost. FAQ-RAG differs from these lines of work by addressing the problem at the representation level, embedding questions and answers directly and thereby reducing the need for complex reasoning or attribution mechanisms during inference.

II. METHODS

A. Document extraction and normalization

Documents are first ingested and converted into structured Markdown representations. For paginated formats such as PDF, page boundaries are preserved explicitly, enabling downstream provenance tracking. This normalization step removes layout artifacts while retaining semantic structure, forming the sole source material for subsequent processing stages.

B. Exhaustive FAQ generation

From each normalized document or page, the system generates an exhaustive set of plausible questions that can be answered solely from the available content. The objective is not to anticipate user queries narrowly, but to enumerate the latent question space implied by the text. This process ensures that coverage is maximized and that each FAQ corresponds to a well-defined informational need.

C. Grounded answer synthesis

For every generated question, an answer is synthesized using only the extracted document content. No external knowledge is introduced, and answers are constrained to remain faithful to the source material. This grounding step effectively validates each FAQ as a self-contained knowledge unit prior to indexing.

D. Dual-vector embedding and storage

Each FAQ is embedded twice, producing a question embedding that captures intent-level semantics and an answer embedding that captures explanatory content. Both vectors are stored in a standard vector database along with metadata specifying document identity, page number, and content scope. This dual representation enables retrieval strategies that prioritize either question similarity, answer similarity, or a combination thereof.

E. Query-time retrieval and response generation

At query time, user questions are embedded and matched against the stored FAQ vectors. Retrieved FAQs already contain complete, grounded answers, allowing the system to respond with minimal additional reasoning. Citations are generated deterministically from the stored metadata, ensuring exact provenance.

III. Results

A. Retrieval precision and recall

By aligning stored representations directly with questions, FAQ-RAG demonstrates improved precision in matching user intent. Because answers are pre-associated with questions, recall is enhanced in scenarios where traditional chunk-based systems might retrieve incomplete or irrelevant fragments.

B. Citation fidelity

FAQ-RAG consistently produces exact citations at the file and page level, as each answer is explicitly linked to its source during ingestion. This contrasts with approximate or inferred citations common in chunk-based systems.

C. Computational efficiency

Shifting reasoning and validation to ingestion time reduces query-time computational overhead. The retrieval process operates over semantically complete units, obviating the need for multi-step synthesis or complex re-ranking pipelines.

IV. Discussion

FAQ-RAG illustrates the benefits of rethinking knowledge representation in retrieval-augmented systems. By treating questions as the atomic units of knowledge, the approach aligns system internals more closely with user objectives. While ingestion-time costs increase due to FAQ generation and answer synthesis, these costs are amortized over repeated queries and are often acceptable in enterprise and research settings. The approach also raises questions about optimal FAQ granularity and coverage, suggesting avenues for future work in adaptive or demand-driven FAQ generation.

V. Conclusion

This paper introduced FAQ-RAG, a FAQ-centric framework for retrieval-augmented question answering that emphasizes structured knowledge, grounded answers, and exact citations. By replacing chunk-based indexing with dual-vector FAQ representations, FAQ-RAG improves retrieval alignment, reduces hallucination risk, and delivers deterministic provenance. The proposed approach offers a scalable and auditable alternative to traditional RAG pipelines and provides a foundation for building trustworthy QA systems in high-stakes domains.

REFERENCES

[1] <To be added>