NOMAD RAGBOT: An AI assistant for navigating distributed community knowledge

Team: Ragalicious

Esma B. Boydas[1], Carlos Madariaga[2], Bernadette Mohr[1], Sherjeel Shabih[1], Joseph F. Rudzinski[1],

[1] *Physics Department and CSMB, Humboldt-Universität zu Berlin, Germany*
[2] *Bundesanstalt für Materialforschung und -prüfung (BAM), Berlin, Germany*

Introduction

Community knowledge is often fragmented across various websites, forums, and chat platforms, making it difficult to obtain reliable and verifiable information. While LLMs can find and summarize such content, their answers can be hard for non-experts to validate, may include outdated information, and rarely provide clear paths to authoritative sources. To address these issues, we developed NOMAD RAGBOT. This retrieval-augmented assistant grounds responses in verified documentation, using the NOMAD ecosystem, a collection of research data management software [244], as a representative use case. By combining semantic retrieval with controlled generation, the system delivers accurate, source-linked answers that help users confidently explore relevant resources.

Results

We implemented a retrieval augmented generation workflow [124] over a corpus of more than 50 documentation sites of varied sizes and structures. The system ingests markdown or HTML files, converts them into vector representations using a custom embedding infrastructure based on the Nomic-Embed-Text model [245] deployed on a remote server, and stores these embeddings in ChromaDB [246] for efficient semantic retrieval. A modular design separates document processing, indexing, retrieval, and conversational management, enabling straightforward extension to new data sources.

User queries are embedded and matched against the vector store through a retriever–reranker pipeline. Retrieved passages are scored semantically and the most relevant chunks are passed to a large language model (LLM) via an adaptive prompt template that emphasizes factual, context-based responses. This design substantially reduces hallucination and improves clarity compared to direct LLM querying.

The system supports multi-turn dialogue through a conversational memory module that constructs context-aware prompts by combining prior chat history with the retrieved evidence. The backend operates as a REST API that manages requests, retrieval, and generation via an `/ask` endpoint, allowing external integration such as the Discord interface.

To improve retrieval accuracy, we implemented a context-aware dynamic chunking strategy that respects markdown heading hierarchies. Each chunk preserves section and subsection titles, with overlapping boundaries to maintain continuity. This structure-aware segmentation improves retrieval precision and ensures that generated answers remain contextually grounded.

Finally, to support objective improvement and facilitate the rapid diagnosis of failure in retrieval and answer generation, we added an evaluation dashboard that runs a gold-standard question set, reports pass rates under adjustable thresholds, filters by source, and provides searchable test items.

We demonstrated our prototype using a Gradio web interface [38] with a simple prompt box, example prompts to get users started, and separate response and citation fields. The prototype performed well in the preliminary vetting stage, providing coherent and relevant answers, even for more advanced queries about complex software development topics in NOMAD. Observable limitations included: 1. a lack of prioritization towards more established or developed data sources, and 2. occasional inaccurate combinations of information from distinct sources.

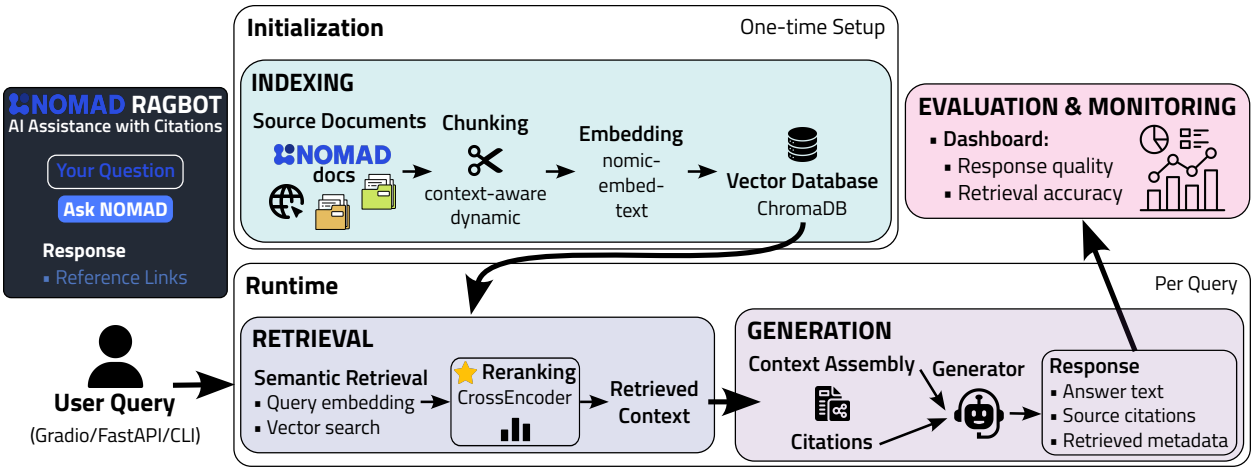

Figure 74: The system performs (1) offline indexing with context-aware chunking, (2) retrieval with CrossEncoder reranking, and (3) LLM-based generation with citations. Performance is monitored via an evaluation dashboard. The figure was generated using icons from Flaticon.com.

Future Work

Despite the immediate usability of the prototype, several straightforward yet impactful improvements remain. These include prioritizing primary documentation sources, enriching retrieval with structured metadata and code-aware chunks, and expanding the evaluation set with task-oriented metrics. Following these technical enhancements, we plan to extend coverage to additional NOMAD sources, including community discussions such as Discord chats, which introduce new challenges for context-preserving data chunking. Ultimately, NOMAD RAGBOT will be integrated into the NOMAD user interfaces and community platforms.

Open source Materials

The source code is available on GitHub:
Explanation and demo video available on Zenodo.

Author Contributions

E.B.B.: Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Writing – original draft.
C.M.: Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Writing – original draft.
B.M.: Software, Validation, Visualization, Writing - review and editing.
S.S.: Methodology, Software, Validation, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review and editing.
J.F.R.: Conceptualization, Methodology, Software, Data curation, Visualization, Writing – original draft, Writing – review and editing, Supervision, Project administration.

Acknowledgements

This work was supported by the NFDI consortium FAIRmat - Deutsche Forschungsgemeinschaft (DFG) - Project 460197019