Building Knowledge Graph of Public Neuroimaging Datasets Using Retrieval-Augmented Generation (RAG)

Fan-Yu Yen[1], and Qianqian Fang[1,*]

[1]*Department of Bioengineering, Northeastern University, 360 Huntington Avenue, MA 02115, USA*
yen.f@northeastern.edu

Abstract: Data-sharing becomes increasingly common among neuroimaging and fNIRS researchers. As datasets grow in size and complexity, finding data resources that meet specific research requirements becomes increasingly challenging. In this work, we integrate retrieval-augmented generation (RAG) with the NeuroJSON.io data-sharing platform to enable advanced natural language-based dataset discovery and automatic knowledge graph construction. The system indexes over 1,500 datasets and supports complex natural language queries that traditional keyword-based search struggles to address. The resulting knowledge graph also provides users a visual overview of the relationship among complex datasets.

Introduction: Data sharing platforms are increasingly used to support reproducible research, secondary analysis, and benchmark development. NeuroJSON.io[1] hosts structured neuroscience datasets that can be reused for replication or downstream methodology studies. However, as the number of datasets grows, identifying relevant datasets becomes increasingly challenging because users must often rely on metadata browsing or keyword search, followed by manual inspection of records. This limits efficient dataset discovery. In this study, we present a retrieval-augmented generation (RAG) framework integrated withNeuroJSON.io that supports natural language-based search over a large amount of datasets.

Methods: We use the LightRAG[2] framework to index over 1,500 datasets hosted on NeuroJSON.io. The metadata of each dataset record is converted into a structured markdown document containing study and subject-related information to accelerate the retrieval process. During the indexing stage, the RAG pipeline parses dataset metadata and identifies relationships between key entities, which is powered by Ollama[3] using the Qwen3:32b[4] large language model (LLM). To improve output consistency and reduce latency, we disable the model's extended reasoning mode. Both the data ingestion pipeline and the downstream question-answering inference run on an NVIDIA A100 GPU.

Results: The RAG pipeline successfully extracts structured metadata from the ingested datasets and constructs a knowledge graph with 17,759 nodes and 27,862 edges (Fig. 1a). This graph not only provides the LLM with structured context to locate relevant connections but also offers users a graphical overview of relationships among all datasets. Users can subsequently query the RAG using natural language questions. Examples of representative queries and their corresponding answers are shown in Fig. 1b.

Conclusion: We demonstrate a natural language-based search interface integrated with NeuroJSON.io, allowing users to more precisely and efficiently locate relevant datasets. By combining RAG with automatic knowledge graph construction, this approach lowers the barrier to dataset discovery in large-scale neuroscience data repositories. Future works will incorporate structured metadata validation to further refine retrieval accuracy by programmatically verifying quantitative criteria.

Fig. 1 (a) A subset of the knowledge graph showing fNIRS-related entities and relationships. (b) Example natural language queries and RAG-retrieved datasets.

References:

1. Fang Q*, Yen F, Lin Y, Xu E, NeuroJSON.io – a community portal for sharing neuroimaging and biophotonics data. *Optica Biophotonics Congress: Biomedical Optics Congress 2024, JS4A.34, FL* (2024)
2. Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2024). LightRAG: Simple and Fast Retrieval-Augmented Generation (Version 3). arXiv.
3. Ollama. "Ollama." Ollama.com, ollama.com/.
4. Yang, An, et al. "Qwen3 Technical Report." ArXiv.org, 2025, arxiv.org/abs/2505.09388.