

A PRACTICAL DOMAIN-SPECIFIC RETRIEVAL-AUGMENTED GENERATION FRAMEWORK FOR HEADACHE RESEARCH

L. Francesco Iannone[1], M. Lovino[2], M. Romozzi[3|4], C. Vollono[4], P. Calabresi[3|4], C. Benatti[2], L. Pani[2|5], F. Tascedda[1]

[1]Department of Life Sciences, University of Modena and Reggio Emilia, Italy; [2]Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Italy; [3]Dipartimento Universitario di Neuroscienze, Università Cattolica del Sacro Cuore, Rome, Italy; [4]Neurologia, Dipartimento di neuroscienze, Organi di Senso e Torace, Fondazione Policlinico Universitario Agostino Gemelli IRCCS, Rome, Italy; [5]Department of Psychiatry and Behavioral Sciences, University of Miami, USA

Background. Headache medicine is characterized by rapidly expanding and heterogeneous scientific literature, particularly on pharmacological treatments. This creates practical difficulties for clinicians and researchers who need rapid, reliable, and source-traceable answers to focused and detailed scientific questions. Although large language models can support biomedical summary, their unguided use is limited by the risk of inaccurate, outdated, or non-verifiable statements. Retrieval-Augmented Generation (RAG) offers a practical solution by linking generative models to a selected, personalized and controlled scientific corpus. Herein we describe the development and practical use of a domain-specific RAG system designed to support headache research.

Methods. A personalized RAG workflow was developed using a local library of headache-related studies selected and controlled by the researcher. Documents were ingested through a Python-based pipeline that extracted full text and metadata, corrected or integrated bibliographic information when available, removed duplicates, and organized the corpus into structured libraries with searchable metadata.Full texts were divided into overlapping chunks, embedded using OpenAI text-embedding-3-small, and stored in a Supabase PostgreSQL database with pgvector indexing. Rule-based topic classification enabled optional category-restricted retrieval. A FastAPI backend exposed retrieval, health, and version endpoints and was deployed through Render.com, while GitHub was used for version control and reproducible deployment. Customized GPT agents were connected to the retrieval API through an OpenAPI schema and instructed to retrieve evidence from the indexed corpus before generating answers. This architecture was designed to keep clinicians and researchers in full control of the documents, metadata, and information sources used by the system.

Results. The system provided a low-cost, modular, and traceable environment for querying selected headache literature eliminating the so-called hallucinations. In practical use, users can ask focused questions on mechanisms, diagnostic criteria, pharmacological treatments, or methodological issues. The agent retrieves relevant passages from the organized, personal library, summarizes the evidence, reports methodological details when available, and links outputs to local source material. This supports rapid preliminary evidence mapping, paper identification, and hypothesis generation.

Conclusions. A personalized, domain-specific RAG system may provide a practical tool for literature-grounded scientific work in headache medicine. Its main value is not to replace systematic reviews or expert interpretation, but to accelerate targeted retrieval, improve traceability, organize scientific libraries, and reduce unsupported generation. An important direction of ongoing development is the transition toward fully local and free/open-source RAG systems.