Trusted Convergence and Knowing What We Know Together: Privacy-Aware Knowledge Discovery Across Neurodegenerative Disease Institutes

Ivan Savchuk [1] **Iona Biggart** [2,3] **Sergio D. Sierra M.** [3] **Ann-Kathrin Schalkamp** [2] **Ruxandra Mihai** [2,3] **Jonas Schulte-Schrepping** [1] **Matthias Becker** [4] **Caleb Webber** [5] **Cynthia Sandor** [2,3] **Hartmut Schultze** [6] **Anna C. Aschenbrenner** [1] **Joachim L. Schultze** [7] **Payam Barnaghi** [2,3]

Abstract

The rapid growth and disciplinary diversification of the neuroscience literature make it increasingly difficult for research institutes to identify where their expertise converges with that of peer institutions. This is a barrier to accelerating therapies for conditions such as Alzheimer's disease. Current AI tools rely on unverified web sources without secure, quality-controlled knowledge sharing. We present Giovanna, a domain-adapted retrieval-augmented generation (RAG) framework built on curated institutional corpora from two neurodegenerative disease research institutes, the German Center for Neurodegenerative Diseases (DZNE) and the UK Dementia Research Institute (UK DRI), designed to surface latent connections, support grounded hypothesis generation, and reveal cross-institutional research convergence within a trusted research environment. We contribute: (i) a neuroscience-specific embedding model fine-tuned on institutional corpora, achieving a recall@1 of 0.678 and MRR@20 of 0.780 superior to baseline; (ii) an embedding-sharing approach that identifies shared research themes without exchanging raw text; and (iii) an empirical comparison of reasoning and non-reasoning language models across query-complexity tiers, showing RAG over trusted institutional knowledge is essential for complex queries, while selective generative fine-tuning yields gains only on domain-specific synthesis tasks. We release Giovanna as a lightweight application for faster insight discovery in neuroscience both within and across research institutions.

[1]Systems Medicine, German Center for Neurodegenerative Diseases, Bonn, Germany [2]Department of Brain Sciences, Imperial College London, London, United Kingdom [3]UK Dementia Research Institute, Care Research and Technology Centre, London, United Kingdom [4]Modular High-Performance Computing and Artificial Intelligence, German Center for Neurodegenerative Diseases, Bonn, Germany [5]UK Dementia Research Institute, Care Research and Technology Centre, Cardiff, United Kingdom [6]Hewlett Packard Enterprise, Houston, Texas, USA [7]German Center for Neurodegenerative Diseases, Bonn, Germany. Correspondence to: Payam Barnaghi <p.barnaghi@imperial.ac.uk>.

AI for Science Workshop of the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

1 Introduction

The volume of published biomedical literature is growing at a rate that overtakes any individual researcher's capacity to synthesise it. In neurodegeneration research, this challenge carries direct clinical urgency: there remains a significant unmet need to identify new therapies and modifiable biomarkers that can prevent or alter the course of diseases such as Alzheimer's. Increasingly, evidence suggests that the most impactful discoveries will arise from convergence science, cross-disciplinary collaborations between researchers who may not yet be aware of the relevance of each other's work (Luo et al., 2025). Findings from molecular biology, clinical neuroscience, and population genetics must be integrated, yet tracking and comparing an exponentially expanding literature across these disciplines has become practically intractable. Conventional literature reviews and collaboration tools are constrained by human time resources, allowing contradictions and research gaps to persist undetected both within and across scientific fields.

Large language model (LLM)-based tools have opened new possibilities for scalable literature synthesis. General-purpose systems such as STORM (Shao et al., 2024) and web-connected AI search tools address the volume problem, but their reliance on open internet access is an increasing limitation as publisher paywalls expand and text-mining restrictions tighten. More critically, hallucination, the generation of plausible but factually incorrect content, remains a substantial risk in precision-demanding biomedical settings; Pal & Sankarasubbu (2024) found that Gemini 1.0 incorrectly answered approximately 40% of medical questions in a structured evaluation. Agentic, internet-connected workflows compound this risk by introducing second-hand hallucinations from unchecked sources, and impose dispro-

portionate computational and environmental overhead (Patterson et al., 2021; Strubell et al., 2019) for queries that are fundamentally domain-constrained and do not require general-purpose reasoning.

A separate but equally important challenge is *cross-institutional knowledge sharing*. Large institutes such as DZNE and UK DRI each maintain substantial publication corpora, yet identifying thematic overlap, complementary expertise, or duplicated effort across institutional boundaries typically requires manual coordination. No existing RAG framework provides a mechanism for privacy-aware inter- or intra-institute discovery. We address both challenges with Giovanna, an open-source RAG system built specifically for the neurodegenerative disease research community. Rather than relying on the internet, Giovanna grounds its responses in a curated, institutionally controlled publication database, keeping sensitive data within institutional boundaries while substantially reducing hallucination risk. To improve retrieval quality beyond what general-purpose embeddings provide, we fine-tune a domain-specific embedding model on institutional publication data and evaluate its impact across standard information retrieval metrics.

The main contributions of this paper include:

- **Domain-adapted embeddings.** We fine-tune an embedding model on neurodegenerative disease literature and demonstrate improved retrieval performance (recall, MRR) over general-purpose baselines (§3.4).

- **Privacy-aware, raw-text-free inter-institute discovery.** We introduce an embedding-sharing approach that enables DZNE and UK DRI to identify shared research topics and collaboration opportunities without exchanging raw publication text (§3.4).

- **Topic modelling and collaboration mapping.** Using domain-adapted embeddings, we perform topic modelling and author-level analysis to identify concrete cross-institutional collaboration opportunities (§3.9).

- **Open-source deployment.** Giovanna is currently deployed across 8,884 publications at DZNE and 2,672 at UK DRI, with strict version control ensuring reproducibility (§3.10).

- **Computational efficiency.** Giovanna operates without discrete GPU access and requires no retraining as the knowledge base grows, making it a practical and environmentally responsible alternative to large-scale agentic AI systems for domain-constrained retrieval tasks.

2 Related Work

The challenge of synthesising an exponentially growing biomedical literature has driven increasing interest in retrieval-augmented systems that ground language model outputs in curated, domain-specific knowledge. Knowledge-intensive natural language processing tasks benefit from RAG over purely parametric models (Lewis et al., 2020), yet many current AI-scientist tools are trained and evaluated on large general corpora derived from the open web (Shao et al., 2024), limiting their utility in specialised research settings where data access is controlled, and domain precision is critical. Domain-specific pre-training and fine-tuning have consistently improved performance on biomedical text. BioClinicalBERT, fine-tuned on MIMIC-III clinical notes (Alsentzer et al., 2019), and PubMedBERT, pre-trained from scratch on PubMed abstracts (Gu et al., 2021), both demonstrated that general-purpose embeddings underperform in clinical and biomedical settings. Importantly, fine-tuning on a narrower, higher-quality corpus has been shown to outperform broad biomedical pre-training on targeted downstream tasks (Alsentzer et al., 2019). This motivated our study of contrastive learning (CL) approaches—unsupervised Simple Contrastive Sentence Embeddings (SimCSE) (Gao et al., 2021), self-supervised CL and supervised CL—to fine-tune embedding models using DZNE and UK DRI publication abstracts. Neuroscience presents a particularly specialised vocabulary spanning molecular biology, clinical phenotyping, and population genetics, amongst other scientific disciplines. To our knowledge, no neuroscience-specific embedding model is currently available in the public domain, a gap that Giovanna addresses.

While general-purpose language models demonstrate sufficient parametric knowledge for simple fact retrieval, RAG has been shown to be essential for knowledge-intensive queries that require synthesis across multiple sources or the identification of contradictions (Chen et al., 2024; Es et al., 2024). The benefit of retrieval scales with query complexity (Chen et al., 2024), motivating our systematic evaluation across simple, institutional, and multi-document synthesis query tiers. Reasoning models have shown gains on complex multi-step tasks (Guo et al., 2025), though whether this advantage persists when high-quality retrieved context is available remains an open question that our comparison addresses.

Topic modelling over scientific literature, ranging from classical approaches such as latent Dirichlet allocation (Blei et al., 2003) to neural methods such as BERTopic (Grootendorst, 2022), has enabled the discovery of latent research themes across large corpora. Giovanna builds on these foundations to track institutional research directions over time and identifies under-explored areas suitable for cross-institute collaboration. Data sharing between research institutes is constrained by confidentiality requirements and access restrictions. Federated learning has been applied in clinical settings to enable model training across institutions without exchanging raw data (Antunes et al., 2022); however, this approach requires substantial coordination overhead, with models trained locally and shared as non-

reversible weight updates. Sharing data as dense representations rather than raw text offers a lighter-weight alternative that preserves institutional privacy without requiring training on highly heterogeneous data (Noble et al., 2022).

We are now in an era in which convergence science, the deliberate integration of knowledge across disciplinary and institutional boundaries, is recognised as essential for tackling complex diseases such as neurodegeneration. The tools to support this integration are now within reach. Connecting research knowledge across institutes has the potential to accelerate discovery and increase the likelihood of medically significant findings by surfacing complementary expertise and reducing duplication of effort (Himmelstein et al., 2017). Giovanna is, to our knowledge, the first RAG framework designed explicitly to enable this form of privacy-preserving, cross-institutional convergence science in a controlled research environment.

3 Methods

3.1 Data Collection and Corpus Construction

All experiments used research papers affiliated with either the DZNE or UK DRI. The DZNE corpus comprised 8,884 full-text PDFs dated from 2008 to 2024, alongside an additional 9,814 plaintext abstracts with associated metadata (abstract, author, publication date, and DOI), retrieved from the DZNE Publication Database. The UK DRI corpus consisted of 2,672 full-text PDF papers dated from 2017 to 2024, identified via PubMed affiliation search. Abstract, author, publication date, and DOI metadata for each UK DRI publication was obtained via PubMed API queries using PubMed IDs.

3.2 Text Extraction and Pre-processing

Full-text extraction from PDFs was performed using GROBID (gro, 2008–2025), which parses documents into TEI-XML format and organises text semantically according to detected section headers, and stored as JSON files.

Several cleaning steps were applied uniformly across both corpora. Section headers were cleaned and standardised with a Mistral 7B Instruct model instructed to remove artifacts of PDF extraction. Text sections labelled as tables or figures were excluded. Unicode characters were normalised to ASCII equivalents, where possible, and removed otherwise. Language identification was performed at the chunk level using pycld2 (Al-Rfou and contributors, 2019). Chunks not detected as English, or for which no language was detected, were discarded. The GROBID-extracted sections were further segmented into sentences using pySBD (Sadvilkar & Neumann, 2020), and sentences longer than 1024 characters were removed to filter out remaining text artifacts.

For papers in both corpora, metadata records were matched to the corresponding full-text PDF publication. Where a paper contained both a GROBID-extracted abstract and a metadata-supplied abstract, the metadata version was retained as canonical; the GROBID-extracted version was discarded when character-level overlap exceeded a defined threshold, preventing duplicate content in the vector store.

3.3 Text Segmentation and Ingestion Into Database

The cleaned and segmented sentences from each GROBID-extracted text section were iteratively combined into chunks of fewer than 1024 characters, with a 50% overlap between consecutive chunks. Fixed-dimension embeddings of each chunk were generated using the embedding models described in Section 3.4. Cleaned text, metadata, and embeddings were loaded into a PostgreSQL database with the pgvector (pgvector contributors, 2024) extension.

3.4 Embedding Models for Information Retrieval

The following base embedding models were evaluated for information retrieval quality in neurodegenerative disease literature: All MiniLM L6 v2 (Reimers & Gurevych, 2019), All MPNet Base v2 (Song et al., 2020), PubMedBERT Base Embeddings (Mezzetti, 2023; Gu et al., 2021), and Qwen 3 Embedding 0.6B (Team, 2023; Bai et al., 2023).

3.4.1 Domain-Specific Embedding Fine-Tuning

To improve retrieval performance on institutional literature, PubMedBERT Base Embeddings was further fine-tuned on sentences from 9,814 DZNE publication abstracts. Three fine-tuning methods were evaluated:

- **Unsupervised SimCSE**, where each sentence serves as both the anchor and the positive example, and all other in-batch sentences are negatives.
- **Self-supervised CL**, where an abstract sentence is the anchor, another sentence from the same abstract is the positive example, and all other in-batch sentences are negatives.
- **Supervised CL**, where a paper title is the anchor, a sentence from the paper's abstract is the positive example, and all other in-batch sentences not from the same abstract are negatives.

All learning approaches used the SentenceTransformers `MultipleNegativesRankingLoss` loss function with `gather_across_devices=True`. Training hyperparameters are reported in Appendix Tables 4 and 5. The best-performing fine-tuning approach was subsequently used to fine-tune the best-performing baseline embedding model.

3.5 Retrieval-Augmented Generation

RAG was implemented using the Langchain library's `history_aware_retriever` function. At inference time, a user query is encoded by a `PGVectorStore` with default parameters, using the same embedding model as the target vector store. The top 20 most relevant chunks with a cosine similarity of at least 0.3 to the query are retrieved by maximum cosine similarity. Retrieved chunks are concatenated with the query into a prompt and passed to the generator model.

3.6 Benchmarking Set Preparation

Institute-Specific Question Set To evaluate performance on institutional knowledge, a multiple-choice question (MCQ) dataset was constructed from the publication corpora (Figure 1). Discussion sections from a random sample of 1,000 DZNE publications and titles from a random sample of 1,500 UK DRI publications were passed to a Mistral 24B model (Langchain, `temperature=0`) with a prompt instructing generation of questions with one correct and three incorrect answer choice letters. Questions were manually reviewed for quality, yielding a final set of 794 DZNE and 486 UK DRI MCQs.

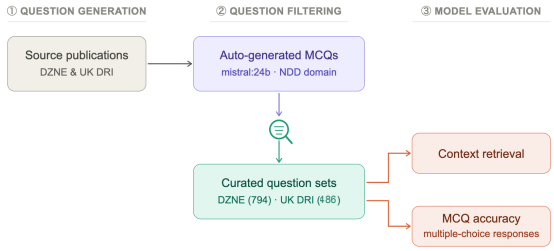

Figure 1. NDD benchmark creation pipeline. MCQs were generated from DZNE and UK DRI publication discussion sections using Mistral 24B, manually filtered into curated sets (DZNE: 794, UK DRI: 486), and used to evaluate models on context retrieval and multiple-choice accuracy.

Publicly-Available Neuroscience Question Set A complementary evaluation set was drawn from four publicly available resources (Figure 2), including questions from BioASQ (yes/no answers), PubMedQA (yes/no/maybe answers), the LAB-Bench benchmark (LitQA2 subcomponent), and the textbook Introduction to Behavioral Neuroscience (chapter-end questions) (Tsatsaronis et al., 2015; Jin et al., 2019; Laurent et al., 2024; Kirby et al., 2024). Questions were filtered for neuroscience relevance using a predefined keyword list (see Appendix B).

3.7 Model Evaluation Tasks

3.7.1 Information Retrieval Task

A `PGVectorStore` retriever with `k=20` and otherwise default parameters was provided with an DZNE or UK DRI question and answer choices and tasked with retrieving the discussion section context used to generate that question. 20 most-similar contexts were retrieved, and evaluation performance was evaluated using the recall@k = 1, 3, 5 and mean reciprocal rank (MRR)@20 metrics. The vector store used for retrieval contained chunks from 3,824 DZNE publications, excluding abstracts.

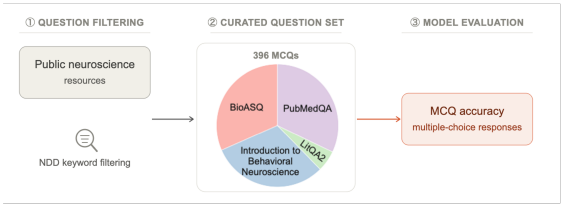

Figure 2. Public benchmark creation pipeline. Publicly available neuroscience resources were filtered using neurodegenerative disease (NDD) keywords to construct a set of 396 multiple-choice questions spanning BioASQ, PubMedQA, Introduction to Behavioral Neuroscience, and LitQA2. Model performance was evaluated based on multiple-choice response accuracy.

3.7.2 Multiple Choice Response Task

We evaluated the full RAG system across multiple configurations using the benchmarking dataset described in Section 3.6. Performance was quantified as the proportion of correctly generated letter answer choices, extracted from the model responses via regex-matching.

3.8 Topic modelling

3.8.1 Topic discovery

All abstracts from the DZNE and UK DRI datasets were embedded separately using the All MiniLM L6 v2 model. First, fine-grained topics were generated with the BERTopic topic modelling pipeline, employing `umap.UMAP(random_state=0)` and `CountVectorizer` with `ngram_range=(1,3)` and `stop_words="english"`, while keeping all other parameters at their default settings. A topic hierarchy was constructed using the `hierarchical_topics` function. The resulting topic clusters were then merged up to the largest possible cluster distance threshold below 1, yielding the final set of topics.

3.8.2 Inter-institute topic comparison

Topic embeddings were extracted from the trained DZNE and UK DRI models and compared across institutions. For each topic i at each institute, an overlap score S_i was calculated as the total number of topic keywords K shared with any topic j in the comparison topic list:

$$S_i = \sum_j |K_i \cap K_j|.$$

Topics with overlap score threshold ≤ 1 were classified as unique, and topics with higher scores were classified as overlapping.

Descriptive topic names were generated using GPT-4o based on the top 10 keywords for institution-specific topics and on the overlapping keywords for the overlapping topics.

3.9 Author Similarity Analysis

A similarity score for each pair of authors at the DZNE was computed based on the cosine similarities of their publication abstracts embedded with All MiniLM L6 v2.

3.9.1 Author Similarity Metric

To arrive at a directed author similarity measure of a query author to a target author, we first find the mean of the cosine similarities between each of the query author's papers and closest paper of the target author. Let $A = \{a_1, a_2, \ldots, a_m\}$ be the set of m paper abstract embeddings written by query author A, and $B = \{b_1, b_2, \ldots, b_n\}$ be the set of n paper abstract embeddings written by target author B. Let $\cos_\text{sim}(a_i, b_j)$ denote the cosine similarity between a_i and b_j. We define the directed similarity of author A to author B as

$$S_{A \to B} = \frac{1}{m} \sum_{i=1}^{m} \max_{1 \leq j \leq n} \cos_\text{sim}(a_i, b_j).$$

$S_{A \to B}$ captures how well the work of author A is represented within the work of author B, while $S_{B \to A}$ captures the reverse. To obtain a symmetric similarity measure between authors A and B, we define

$$S_{AB} = \frac{1}{2} \left(S_{A \to B} + S_{B \to A} \right).$$

We used the directed similarity measure in Table 2 and the symmetric similarity measure in Figure 6.

3.10 Deployment and User Interface

Giovanna was deployed as a full-stack, containerised, locally-hosted application at both the DZNE and UK DRI. An interactive user interface was developed for user query, topic keyword-based search, and 2D-embedding visualizations.

4 Results

4.1 Performance of Embedding Models

Table 1 reports retrieval performance across all evaluated embedding models. The best performing model at both recall@1 and MRR@20 was Qwen 3 Embedding 0.6B with unsupervised fine-tuning, achieving a recall@1 of 0.678 and MRR@20 of 0.780. The baseline Qwen 3 Embedding 0.6B without fine-tuning was the second strongest model overall; notably, it marginally outperformed its fine-tuned counterpart at recall@3 and recall@5, achieving 0.885 and 0.921, respectively, compared to 0.864 and 0.911, suggesting that unsupervised fine-tuning sharpens precision at the top rank at a small cost to broader coverage. Across all model families, the unsupervised fine-tuning approach improved retrieval performance relative to the respective base model, while the self-supervised fine-tuning approach reduced performance in all cases. General-purpose models performed poorest overall; all-mpnet-base-v2 achieved the lowest recall@1 of 0.379, highlighting the limitation of domain-agnostic embeddings for specialised neuroscience retrieval.

4.2 Performance of generative models and RAG systems

The performance of the full RAG system across different configurations is shown in Figure 3. For the DZNE and UKDRI question sets, all RAG configurations outperformed the baseline. For the DZNE questions, Qwen 3 30b with the Qwen 3 Embedding 0.6B retriever showed the biggest gains (97.0% and 97.1% for the configurations with the base and fine-tuned embedding model, respectively) compared to baseline Qwen 3 30b (80.1%). For the UK DRI questions, the same configuration also achieved the best performances of 95.2% and 94.4% for base and fine-tuned Qwen 3 Embedding 0.6B models, respectively, compared to a 74.9% baseline. We saw minimal performance differences (<1%) for the fine-tuned Qwen 3 Embedding 0.6B model compared to RAG with base Qwen 3 Embedding 0.6B on UK DRI and DZNE questions.

For the public benchmarks, we observed mixed results. On the BioASQ, LitQA2, and PubMedQA sets, RAG did not improve and sometimes degraded performance. The large generator models Qwen 3 30b and Llama 3.3 70b on the BioASQ set with yes/no questions were most negatively affected. Inspection of the Qwen 3 30b model's reasoning suggests that when context does not contain the key named entity needed to answer the question, the model favors the "No" response (Appendix E). Overall, RAG-enabled models responded "No" roughly twice as frequently as their corresponding baselines, resulting in a shift either closer to the ground truth yes/no response distribution (Gemma 3 4b and Llama 3.1 8b) or away from that distribution (Qwen 3 30b and Llama 3.3 70b) (Appendix F). Use of the Qwen 3 Embedding 0.6B fine-tuned model attenuated the negative impact in many RAG configurations evaluated on the public benchmarks. We attribute this result to the fine-tuned model's consistently lower similarity score assignment compared to the base model, resulting in greater filtering of irrelevant context when combined with a simi-

Table 1. Embedding model performance comparison. Best performance highlighted in bold.

Model	Recall@1 (Accuracy)	Recall@3	Recall@5	MRR@20
All MiniLM L6 v2 (No fine-tuning)	0.429	0.650	0.727	0.563
all-mpnet-base-v2 (No fine-tuning)	0.379	0.597	0.670	0.509
PubMedBERT Base Embeddings (No fine-tuning)	0.450	0.681	0.763	0.586
PubMedBERT Base Embeddings (Unsupervised fine-tuning)	0.523	0.746	0.800	0.649
PubMedBERT Base Embeddings (Self-supervised fine-tuning)	0.334	0.589	0.683	0.490
PubMedBERT Base Embeddings (Supervised fine-tuning)	0.443	0.703	0.775	0.589
Qwen 3 Embedding 0.6B (No fine-tuning)	0.645	**0.885**	**0.921**	0.770
Qwen 3 Embedding 0.6B (Unsupervised fine-tuning)	**0.678**	0.864	0.911	**0.780**

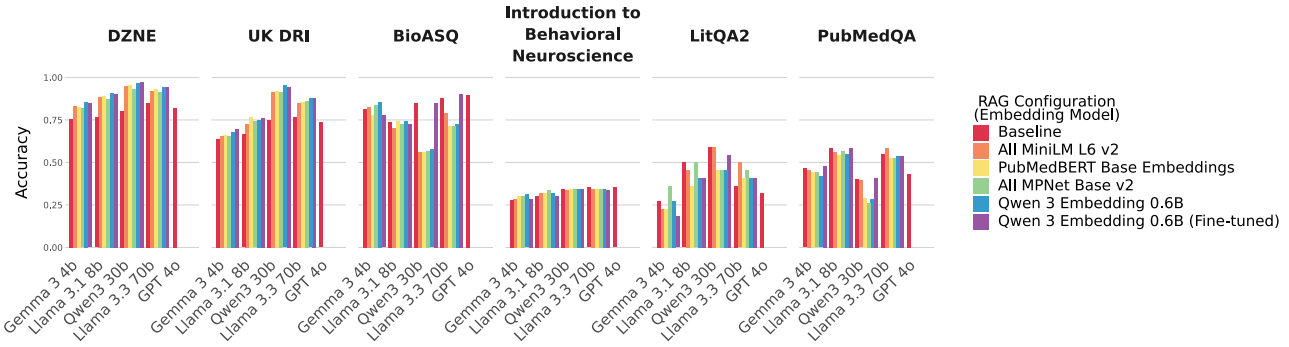

Figure 3. **Accuracy of large language models on biomedical MCQ benchmarks across three retrieval configurations.** Five generative models were evaluated on two curated neurodegenerative disease question sets (DZNE, n=794; UK DRI, n=486) and four public benchmarks (BioASQ, Introduction to Behavioral Neuroscience, LitQA2, PubMedQA) under the conditions of no retrieval (Baseline) and embedding-model-assisted retrieval (All MiniLM L6 v2, PubMedBERT Base Embeddings, All MPNet Base v2, base or fine-tuned Qwen 3 Embedding 0.6B). GPT-4o results are reported under baseline conditions only.

larity score threshold in RAG (Appendix G). Performance on Introduction to Behavioral Neuroscience remained uniformly low across all configurations (∼28–35%), reflecting that these student-focused questions on foundational neuroscience concepts are unlikely to be addressable through literature retrieval. Together, these results demonstrate that RAG provides reliable gains for institutional question sets where the retrieval corpus is closely matched to the evaluation domain, but does not generalise to external benchmarks requiring broad or foundational knowledge.

4.3 Topic modelling across institutes

The topic modelling analysis identified multiple key research themes both within and across institutes. By incorporating publication dates, temporal trends in topic prevalence were examined (Figure 4). This analysis reveals a marked increase in research activity in areas such as multiple sclerosis, dementia care, and neuronal synaptic activity in recent years. Figure 5 illustrates the overlap and distinction of topics between institutes, alongside representative examples of institute-specific themes. For instance, research on sleep and circadian rhythm was found to be unique to UK DRI, while gut microbiome research was specific to DZNE. In contrast, a broad range of topics, including neuro-COVID, Huntington's Disease, and dementia care were shared across institutes, highlighting areas of common research focus.

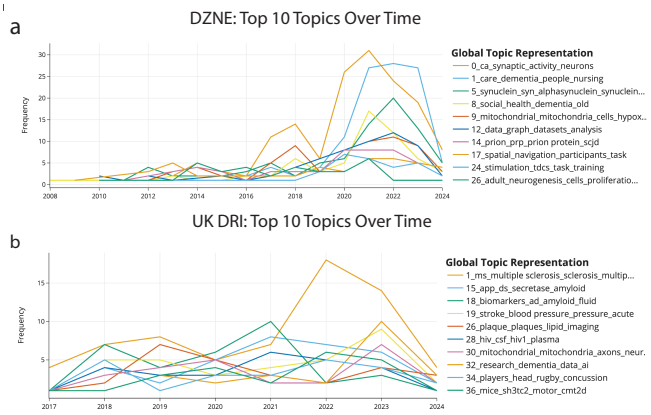

Figure 4. Temporal evolution of the top 10 research topics identified using topic modelling for (a) DZNE and (b) UK DRI. Each line represents the yearly frequency of a topic, with labels indicating the dominant keywords describing each topic. Both institutions show increasing activity in neurodegeneration-related themes over time, with notable surges in specific topics (e.g., multiple sclerosis and dementia) in recent years, reflecting shifting research priorities.

4.4 Author retrieval and clustering

Using the abstract embeddings from DZNE, we built a tool for uncovering potential author collaborations. We first

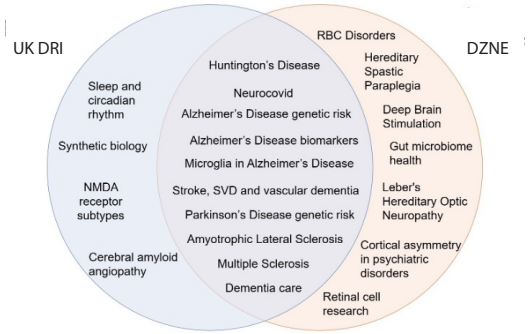

Figure 5. Overlap of the top 10 most frequent research areas across both institutes within each category (unique to DZNE, unique to UK DRI, and common to both institutes), as identified by the topic modelling approach (see Section 3 for details).

computed and visualized DZNE authors' co-publication and pairwise paper abstract-based similarity scores together in a heatmap (Figure 6), revealing both the existing pattern of co-publication at DZNE, as well as a "collaboration potential" represented by the author-to-author similarity score. While some author groups at DZNE showed high levels of collaboration and co-publication, other author clusters demonstrated high "collaboration potential" but little actual collaboration as represented by publication output. Thus, to facilitate personalized author discovery, we propose an author-to-author query tool (Table 2).

Figure 6. Heatmaps illustrating co-publication (upper triangular portion) and inter-author similarity (lower triangular portion) for all DZNE-affiliated authors with five or more publications. We observe regions characterised by high similarity and high co-publication (green-outlined square), as well as regions exhibiting high similarity despite low levels of co-publication (red-outlined square).

4.5 Giovanna Application

We developed a user interface equipped with a RAG-based chat functionality (Figure 7). All system responses are accompanied by the DOI and section name information of the retrieved contexts, allowing the user to fact-check language model outputs by accessing the relevant scientific publications.

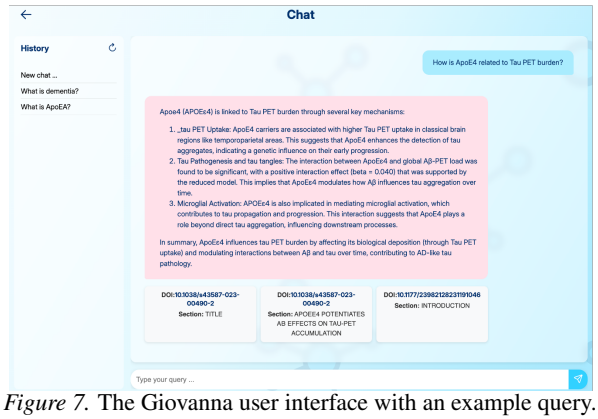

Figure 7. The Giovanna user interface with an example query.

5 Discussion

Giovanna demonstrates the utility of domain-specific retrieval and generation for institutional neuroscience research, showing that a specialist, institute-wide AI system can meaningfully support both research discovery and cross-institute collaboration. Domain-adapted retrieval consistently outperforms general-purpose baselines, and RAG grounding proves essential where parametric knowledge alone is insufficient. Together, these findings support the case for specialist, privacy-aware AI systems over general-purpose tools in high-stakes research environments.

The retrieval evaluation results show a consistent benefit of unsupervised domain adaptation. Qwen 3 Embedding 0.6B fine-tuned on neurological domain text achieved the highest retrieval performance (recall@1: 0.678 vs. 0.645 for its base counterpart). General-purpose models (All MiniLM L6 v2, all-mpnet-base-v2) performed poorest, confirming that domain-agnostic embeddings are poorly suited to neuroscience retrieval. Unsupervised fine-tuned variants outperformed their base equivalents across all model families, suggesting that lightweight contrastive adaptation provides a meaningful retrieval signal without labelled data. Supervised fine-tuning also improved over base models, though to a lesser extent, likely reflecting the challenge of obtaining sufficient high-quality labelled pairs in a specialised domain (Gu et al., 2021; Alsentzer et al., 2019).

RAG consistently improved performance on institutional question sets over purely parametric models, though gains were inconsistent on public benchmarks. No single generative model dominated across all datasets: GPT-4o achieved the highest accuracy on Introduction to Behavioral Neuroscience, Qwen 3 30b performed best on DZNE, UK DRI, and LitQA2, and Llama 3.1 8b on PubMedQA. For the institutional question sets, Qwen 3 30b in a RAG configuration performed strongest at DZNE, achieving 97.1% accuracy, the best result among all settings. At UK DRI, the RAG-enabled models consistently showed performance gains over

Table 2. Most similar authors to query Author A having no co-publications with Author A. The similarity score is defined in 3.9.1.

Result author	Similarity score	Result most similar paper
Author B	0.59	Baseline CSF/Serum-Ratio of Apolipoprotein E and Rate of Differential Decline in Alzheimer's Disease
Author C	0.58	Baseline CSF/Serum-Ratio of Apolipoprotein E and Rate of Differential Decline in Alzheimer's Disease
Author D	0.58	Clinico-genetic findings in 509 frontotemporal dementia patients
Author E	0.56	Development of a sensitive trial-ready poly(GP) CSF biomarker assay for C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis

baseline across all models. We saw consistent added benefit of larger-dimension embeddings for the institutional question sets, where the Qwen 3 Embedding variants outperformed smaller embedding models in all RAG setups. However, in public datasets, RAG frequently harmed model performance by introducing unrelated context. Here, the fine-tuned Qwen 3 Embedding 0.6B showed increased robustness compared to other embedding models, with smaller score declines compared to baseline.

The embedding-sharing approach proposes a fundamentally different model for cross-institutional knowledge sharing: research groups discover shared topics and complementary expertise by exchanging fixed-dimension vectors and topic keywords rather than raw text, enabling reduced data exposure without the coordination overhead of federated learning (Antunes et al., 2022). Applied across DZNE and UK DRI, topic-overlap analysis (Figure 5) reveals shared engagement in Huntington's disease, dementia, and neurological consequences of COVID-19, while also identifying high-value uninvestigated directions such as sleep and circadian-rhythm dysregulation and gut microbiome involvement in neurodegeneration. Author-level clustering surfaces thematically aligned researchers, enabling targeted collaboration and more efficient use of funding. The approach is domain-agnostic and scales to any institution with a substantial publication corpus.

The computational design of Giovanna reflects a proportionality principle: using a model appropriately sized for the task. It runs locally on consumer hardware without discrete GPUs, and expanding the knowledge base only requires generating embeddings for new publications with no retraining. This contrasts with the inference cost of large-scale API-served models such as GPT 4, which consumed between 3.41 kWh and 13.24 kWh in a recent set of experiments on the USA Computing Olympiad database (Woo, 2025). For domain-constrained retrieval, such costs are avoided: smaller-parameter models with grounding substitute for scale, making purpose-built retrieval systems a more environmentally responsible alternative to large agentic pipelines.

Several limitations should be acknowledged. Evaluation via multiple-choice questions, while reproducible, does not fully capture performance on the open-ended queries researchers pose in practice. The institute-specific MCQ datasets were generated using Mistral 24B, which may have introduced systematic biases in question construction. The domain-specific embedding model was fine-tuned exclusively on DZNE abstracts; transferability to other institutions has not been formally evaluated, though the scale of the training corpus partially mitigates this concern. Privacy constraints limited this study primarily to open-source generative models; findings may not transfer to the latest proprietary closed-weight alternatives.

Future work will incorporate the full UK DRI publication corpus, extend the embedding-sharing protocol to additional institutes and disease domains, and conduct human domain expert evaluation of the MCQ benchmark. Agentic extensions and literature-grounded hypothesis generation are natural next steps towards making Giovanna a fully autonomous AI-augmented scientific discovery platform.

6 Conclusion

Giovanna is designed to support research collaboration at scale, which requires tools that make complementary knowledge visible without centralising it. By letting each institute contribute a compact, privacy-aware representation of its research landscape, joint analyses can be composed without either party relinquishing control over their underlying data, functioning as shared infrastructure for collective intelligence, encouraging collaboration through transparent evidence of where interests converge and where gaps remain.

In its current form, Giovanna operates as a specialist research tool, grounding queries in verified institutional literature, surfacing collaboration opportunities, and reducing the burden of large-scale literature synthesis. The trajectory toward AI-assisted hypothesis generation and inter-institute deployment points to a system capable of functioning as a scientific knowledge-discovery assistant, whose outputs are auditable, reproducible, and anchored in high-quality domain knowledge rather than unconstrained generation.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here. The presented system is designed to support, not replace, domain experts in neurodegenerative disease research, and the embedding-sharing approach operates on published manuscripts to preserve institutional data privacy and support inter- and intra-institute collaboration.

Code Availability

The code used to generate the analysis for this work is available at: https://gitlab.dzne.de/systemsmedicine/giovanna-analysis.

Acknowledgments

This research is supported by the German Center for Neurodegenerative Diseases and the UK Dementia Research Institute.

References

Grobid. https://github.com/grobidOrg/grobid, 2008–2025.

Al-Rfou and contributors. pycld2: Python bindings for Compact Language Detector 2 (CLD2). https://github.com/aboSamoor/pycld2, 2019.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., and Eskofier, B. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13 (4):1–23, 2022.

Bai, J. et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022, 2003.

Chase, H. Langchain. https://github.com/langchain-ai/langchain, 2023.

Chen, J., Lin, H., Han, X., and Sun, L. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024.

Es, S., James, J., Anke, L. E., and Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pp. 150–158, 2024.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6894–6910, 2021.

Grootendorst, M. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. In *ACL*, 2021.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife*, 6: e26726, 2017.

Jin, Q., Dhingra, B., Cohen, W. W., and Lu, X. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Kirby, E. D., Glenn, M. J., Sandstrom, N. J., and Williams, C. L. Introduction. In *Introduction to Behavioral Neuroscience*. OpenStax, 2024. URL https://openstax.org/books/introduction-behavioral-neuroscience/pages/1-introduction.

Laurent, J. M., Janizek, J. D., Ruzo, M., Hinks, M. M., Hammerling, M. J., Narayanan, S., Ponnapati, M., White, A. D., and Rodriques, S. G. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2):305–315, 2025.

Mezzetti, D. Embeddings for medical literature. https://medium.com/neuml/embeddings-for-medical-literature-74dae6abf5e0, 2023.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, pp. 10110–10145. PMLR, 2022.

Pal, A. and Sankarasubbu, M. Gemini goes to med school: exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 21–46, 2024.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. Carbon emissions and large neural network training. arxiv. *arXiv preprint arXiv:2104.10350*, 2021.

Pedregosa, F. et al. Scikit-learn: Machine learning in python. *JMLR*, 2011.

pgvector contributors. pgvector: Open-source vector similarity search for postgres. https://github.com/pgvector/pgvector, 2024. Version 0.5.1.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP*, 2019.

Sadvilkar, A. and Neumann, M. pysbd: Pragmatic sentence boundary disambiguation. In *Proceedings of the Second Workshop on NLP for Conversational AI*, pp. 147–152. Association for Computational Linguistics, 2020.

Shao, Y., Jiang, Y., Kanell, T., Xu, P., Khattab, O., and Lam, M. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, 2024.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*, 2020.

Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3645–3650, 2019.

Team, A. C. A. Qwen embedding models (0.6b and 8b). https://huggingface.co/Qwen, 2023. Embedding models based on Qwen architecture, including 0.6B and 8B parameter variants.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138, 2015.

Wolf, T. et al. Transformers: State-of-the-art natural language processing. *EMNLP*, 2020.

Woo, N. H. A comparative study of ai and human programming on environmental sustainability. *Scientific Reports*, 15:39182, 2025. doi: 10.1038/s41598-025-24658-5.

A Implementation Details

A.1 Hardware

All experiments were conducted on a system with both CPU and GPU resources. Specifically, for training, we used with 1x NVIDIA A100 with 80G RAM (UK DRI) and up to 8x NVIDIA H100 with 80G RAM each (DZNE). Inference was performed on the Apple Silicon M3 (DZNE) or M4 (UK DRI) chipset, which uses a unified memory architecture where computation is distributed across CPU and Neural Engine cores, without reliance on discrete GPU hardware.

A.2 Software Environment

All experiments used Python 3.10 managed via Conda. Embedding generation and model fine-tuning used the Sentence Transformers (Reimers & Gurevych, 2019) and Transformers (Wolf et al., 2020) libraries; retrieval pipelines were implemented with Langchain (Chase, 2023). Topic modelling used BERTopic (Grootendorst, 2022) with Scikit-Learn clustering (Pedregosa et al., 2011).

A.3 RAG Models

Table 3. Models used in the RAG system.

Model	**Type**	**Parameters**	**Notes**
DeepSeek R1 1.5b	Generative	1.5B	Lightweight reasoning model
LLaMA 3.1 8b	Generative	8B	Efficient general-purpose LLM
LLaMA 3.3 70b	Generative	70B	Large-scale high-performance LLM
Qwen 3 30b	Generative	30B	Strong multilingual reasoning
GPT-4o	Generative	Proprietary	API-based, closed-source model
all-MiniLM-L6-v2	Embedding	22M	Sentence-transformer embedding model

B Neurodegenerative Disease-Related Keywords for Public Benchmark Curation

Keywords: brain, neuro, cognitive, cognition, neuronal, neurological, neuroscience, psychosis, psychotic, schizophrenia, bipolar, epilepsy, memory, dementia, alzheimer, aphasia, consciousness, parkinson, migraine, autism, adhd, multiple sclerosis, neurodegenerative, eeg, mri, fmri, electrophysiology, stroke, seizure, cortex, cortical, nerve, axon, spinal, neuron, synapse, neurotransmitter, psychology, mental, psychiatry, hallucination, delusion, affective, amygdala, hippocampus.

C Training Specifics

Table 4. Hyperparameters for training PubMedBERT Base Embeddings.

Hyperparameter	Value
Device	6 NVIDIA H100 GPUs
Per-device training batch size	32 (unsupervised/self-supervised), 64 (supervised)
Epochs	1
Precision	fp32
Random State	0

Table 5. Hyperparameters for training Qwen 3 Embedding 0.6B.

Hyperparameter	Value
Device	8 NVIDIA H100 GPUs
Per-device training batch size	16
Epochs	1
Precision	bf16
Random State	0

D Benchmarking dataset specifics

Table 6. Evaluation datasets specifics

Dataset	Number of Questions	Origin
DZNE	794	Mistral 24B-generated MCQs from DZNE paper discussion sections
UK DRI	486	Mistral 24B-generated MCQs from UK DRI paper titles
BioASQ	125	BioASQ (13th edition), Task b, training data[1]
PubMedQA	127	HuggingFace dataset[2]
LitQA2	22	HuggingFace dataset[3]
Introduction to Behavioral Neuroscience	122	Online textbook[4], (chapters 1–19)

E RAG Reasoning Trace Example

Example RAG reasoning trace for Qwen 3 30b with All MiniLM L6 v2 evaluated on BioASQ.

Query: "Is SLC22A3 expressed in the brain?"

Reasoning excerpt: "Okay, let's try to figure out if SLC22A3 is expressed in the brain based on the given context. ... The question is about SLC22A3. **The context doesn't mention SLC22A3 at all. The closest is SLC18B1. Since the question is about SLC22A3, and there's no information provided about its expression in the brain, the answer would be no.** But wait, maybe I missed something. Let me check again. The context starts by talking about studies on astrocytes and their expression of certain proteins. It mentions that some proteins are not detected (ND). However, SLC22A3 isn't referenced. The answer would be B: no, because **the context doesn't state that SLC22A3 is expressed in the brain. The absence of mention implies that either it's not expressed or the studies didn't detect it. Since the question is about whether it's expressed, and the context doesn't confirm it, the answer is B.**"

Model Response: "B" [No]

Correct Answer: "A" [Yes]

F BioASQ Benchmarking Answer Distributions

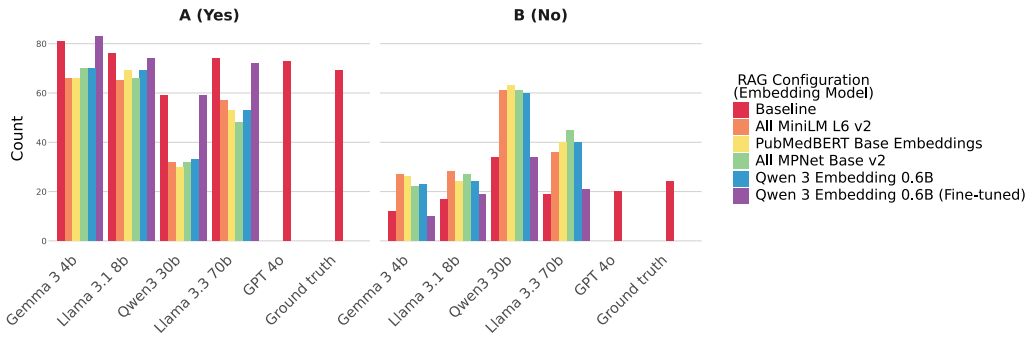

Figure 8. Letter answer choice counts for all model configurations evaluated on the BioASQ question set. The corresponding accuracy results are shown in Figure 3. Invalid letter answers comprised <1% of all responses.

[1] https://participants-area.bioasq.org/datasets/
[2] https://huggingface.co/datasets/qiaojin/PubMedQA
[3] https://huggingface.co/datasets/futurehouse/lab-bench
[4] https://openstax.org/details/books/introduction-behavioral-neuroscience

G Qwen Embedding 0.6B Cosine Similarity Distributions for Base and Fine-Tuned Models

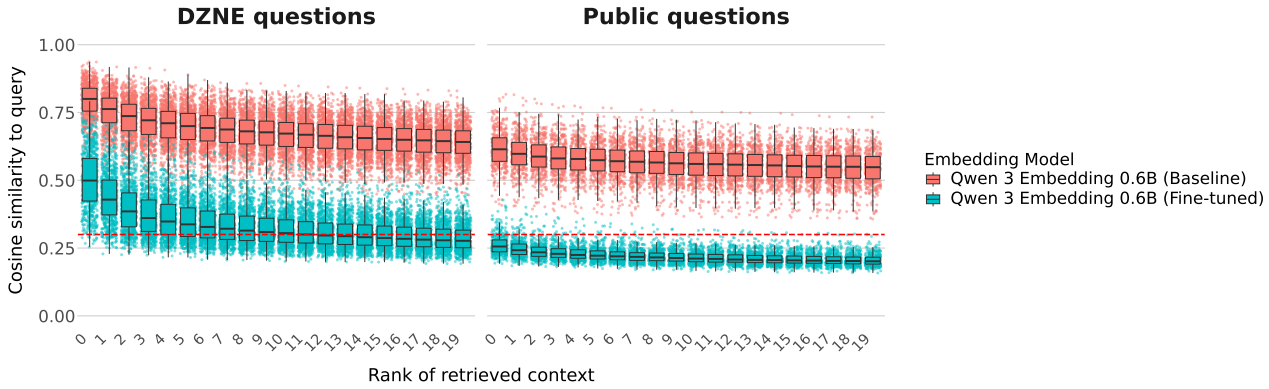

Figure 9. Cosine distance values of the retrieved contexts at each of the 20 ranks are show, comparing the Qwen 3 Embedding 0.6B base and fine-tuned models. Retrieval was performed using the same set of DZNE contexts as introduced in 3.7.1 with either the DZNE question set (left subplot) or public-source question set (right subplot) used as the query. The RAG cosine similarity score threshold used in experiments is shown (dashed red line).