Bridging the Menopause Information Gap: A Peer-Reviewed Multimodal RAG Framework with Hybrid Retrieval

Sara El-Ateif and Soufiane Elassri
National Higher School of Artificial Intelligence and Data Sciences (ENSIASD),
Ibn Zohr University, Taroudant, Morocco
(s.el-ateif@, soufiane.elassri.60@edu.)uiz.ac.ma

***Abstract*—We address the challenge of accessing reliable, personalized medical information for women experiencing menopause. Online data is often contradictory and specialist consultations remain cost-prohibitive, creating a critical need for scientifically validated guidance. We propose an intelligent system based on a Multimodal Retrieval-Augmented Generation (MRAG) architecture that centralizes and personalizes access to peer-reviewed menopause literature (PLOS ONE). Our approach uses Gemini 2.0 Flash to generate semantic captions for clinical imagery, enabling text-based retrieval of multimodal assets; a hybrid search mechanism combining vector similarity and BM25, optimized via cross-encoder reranking and semantic repacking; and an LLM-based query classifier routing requests with 96% accuracy. Empirical evaluation on 50 domain-specific queries yielded source faithfulness of 88%, response relevance of 90%, and contextual precision of 85%. A preliminary user study returned a System Usability Scale score of 82/100, indicating strong potential for digital women's health as a supplement to professional medical consultation.**

***Index Terms*—Digital Women's Health, Menopause, Multimodal Indexing, RAG, Information Retrieval.**

I. Introduction

Menopause is a universal physiological transition experienced by women, typically occurring between the ages of 45 and 55. It involves a significant decline in estrogen and progesterone levels, resulting in symptoms that range from vasomotor episodes and sleep disruption to mood instability and cognitive changes [1]. The clinical presentation varies considerably among individuals, which makes generic advice insufficient for personalized care.

Access to reliable online health information regarding menopause has not kept pace with the growing need. Searches for menopause-related symptoms frequently yield unverified blog posts, forum threads, and commercially motivated content. Furthermore, digital health tools are predominantly designed for English-speaking audiences, exacerbating inequalities in specialist care access.

Recently, Large Language Models (LLMs) have shown potential in bridging this information gap. However, their tendency to hallucinate renders them unsuitable as direct patient-facing tools without rigorous validation [3]. Retrieval-Augmented Generation (RAG) [2] addresses this by grounding answers in documents retrieved from a controlled knowledge base.

We built a specialized assistant that retrieves information from a curated corpus of 100 peer-reviewed menopause articles and processes both text and visual content. The novelty lies in the principled combination of these components into a single, locally-deployable application. Specifically, our contributions are: (1) a domain-restricted corpus from PLOS ONE [5] built via an automated Selenium pipeline; (2) a multimodal indexing pipeline using Gemini 2.0 Flash to caption clinical figures for text-based retrieval; (3) a hybrid retrieval engine combining BGE-large dense embeddings with BM25, followed by cross-encoder reranking and semantic repacking; (4) an LLM-based query classifier with 96% routing accuracy; and (5) a user-facing interface with four modules—RAG chatbot, symptom tracker, PDF report generator, and educational cards—evaluated via SUS.

II. Related Work

A. RAG in Healthcare

RAG [2] has become the default architecture for reducing hallucination in medical question answering. Handy et al. [4] applied this approach to the menopause domain using GPT-4, reporting faithfulness of 88.6% and relevance of 91.8%. Zhu et al. [6] proposed EMERGE, integrating RAG into EHR-based predictive modeling. A systematic review by Liu et al. [7] found that RAG-augmented systems outperform standard LLMs in diagnostic tasks with an odds ratio of 1.35.

B. Multimodal and Specialized Architectures

Medical documents are inherently multimodal. Xia et al. [8] demonstrated a 43.8% improvement in factual precision when their MMed-RAG system processed visual data alongside text. Lahiri et al. [9] applied multimodal RAG to Alzheimer's literature. On the retrieval optimization side, Sohn et al. [10] introduced RAG^2 to filter low-quality passages before generation.

C. Gap Analysis

Most systems handle either text or images, but rarely both in a tightly integrated pipeline. Hybrid dense–sparse retrieval is seldom implemented in the medical domain, and women's health—menopause in particular—is almost entirely absent from the RAG literature. To the best of our knowledge, no

existing system combines curated multimodal retrieval, hybrid search, and end-user tools (symptom tracking, PDF reporting) into a single, domain-specific, locally-deployable application. Our contribution lies in bridging these distinct components into a comprehensive digital health tool.

III. Methodology

Fig. 1 gives an overview of the system. A user query first passes through an LLM classifier that determines whether direct generation is sufficient or whether retrieval is needed (see Section III-C). When retrieval is triggered, the query hits both a dense vector index and a BM25 index built from our curated corpus; results are merged, reranked by a cross-encoder, repacked to remove redundancy, and passed to the generator.

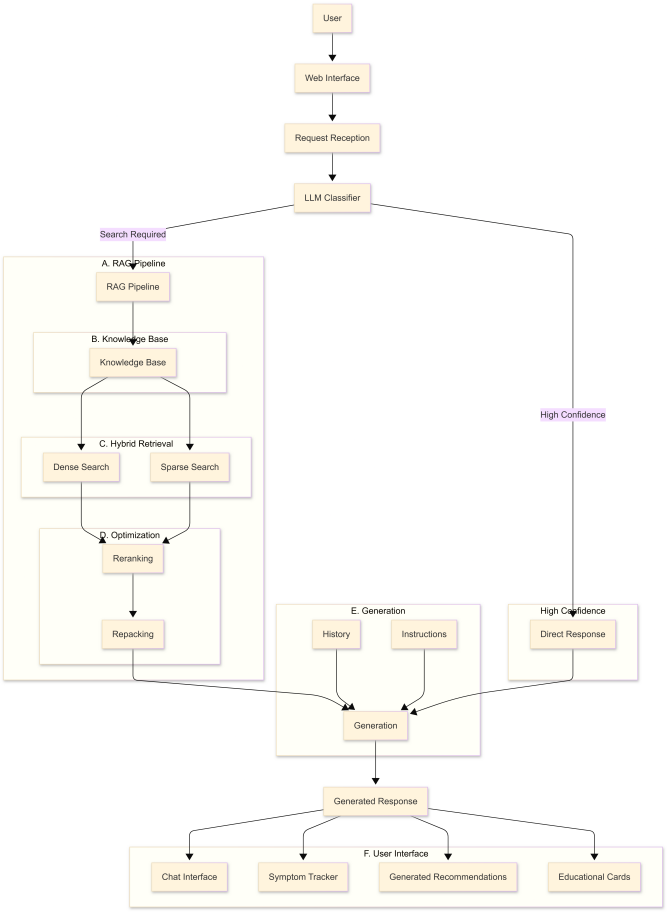

Fig. 1. System architecture. Queries are classified by an LLM and routed either to direct generation or the full RAG pipeline. Hybrid retrieval (dense + sparse) feeds into cross-encoder reranking and semantic repacking before generation.

A. Curated Clinical Knowledge Base

The corpus consists of 100 peer-reviewed articles on menopause from PLOS ONE, collected via a Selenium-based scraper. PLOS ONE was chosen for its open-access policy, PubMed and Scopus indexing, and broad clinical scope spanning hormonal mechanisms, vasomotor symptom management, psychological impact, and cardiovascular outcomes.

We acknowledge that restricting the corpus to a single publisher risks content bias toward English-language, Western-demographic perspectives and may omit consensus guidelines from bodies such as the International Menopause Society. Multi-source expansion is a priority for future work. The knowledge base prioritizes reliability and clinical traceability; every retrieved passage links to a peer-reviewed source, embodying "Clinical Safety by Design."

B. Multimodal Indexing Pipeline

We used the *Unstructured* library for document ingestion, partitioning PDFs into three content streams:

1) **Text** is divided at semantic boundaries using 512-token chunks with 64-token overlap.
2) **Tables** are converted to Markdown and summarized by Gemini 2.0 Flash [12], enabling key statistics to be matched to natural-language queries.
3) **Images** are base64-encoded and passed to Gemini 2.0 Flash [11], [12], prompted to generate medically specific RAG-optimized captions. A manual review of 20 random images found 90% caption accuracy; two inaccurate captions were manually corrected.

This pipeline is shown in Fig. 2. All three content streams are embedded into the same vector space.

C. LLM Query Classifier

Before retrieval is triggered, each query is classified by a zero-shot Gemini 2.0 Flash prompt [11], [12]. Factual, clinical, or symptom-specific queries are routed to the full RAG pipeline; greetings or out-of-domain queries (e.g., *"Hello, can you help me?"*) are handled directly or politely declined. On a 50-query evaluation set, the classifier achieved 96% routing accuracy. All faithfulness and relevance metrics in Section IV are computed exclusively on RAG-routed queries.

D. Vectorization and Hybrid Retrieval

1) Embedding Model Selection: We evaluated four open-source embedding models on a 50-query subset (Table I). BAAI/bge-large-en-v1.5 [21] achieved the highest precision (92.3%) at the cost of the longest indexing time (5h 12min). Since indexing is a one-time offline operation, this trade-off was acceptable; smaller models sacrifice over 13 precision points—too large a penalty for a health domain.

TABLE I
Embedding Model Comparison (Corpus: 100 articles; Test: 50 queries)

Model	Prec. (%)	Index Time
BAAI/bge-large-en-v1.5 [21]	**92.3**	5h 12min
intfloat/e5-large-v2 [22]	88.9	3h 45min
all-mpnet-base-v2 [23]	81.4	2h 30min
all-MiniLM-L6-v2 [24]	78.6	1h 10min

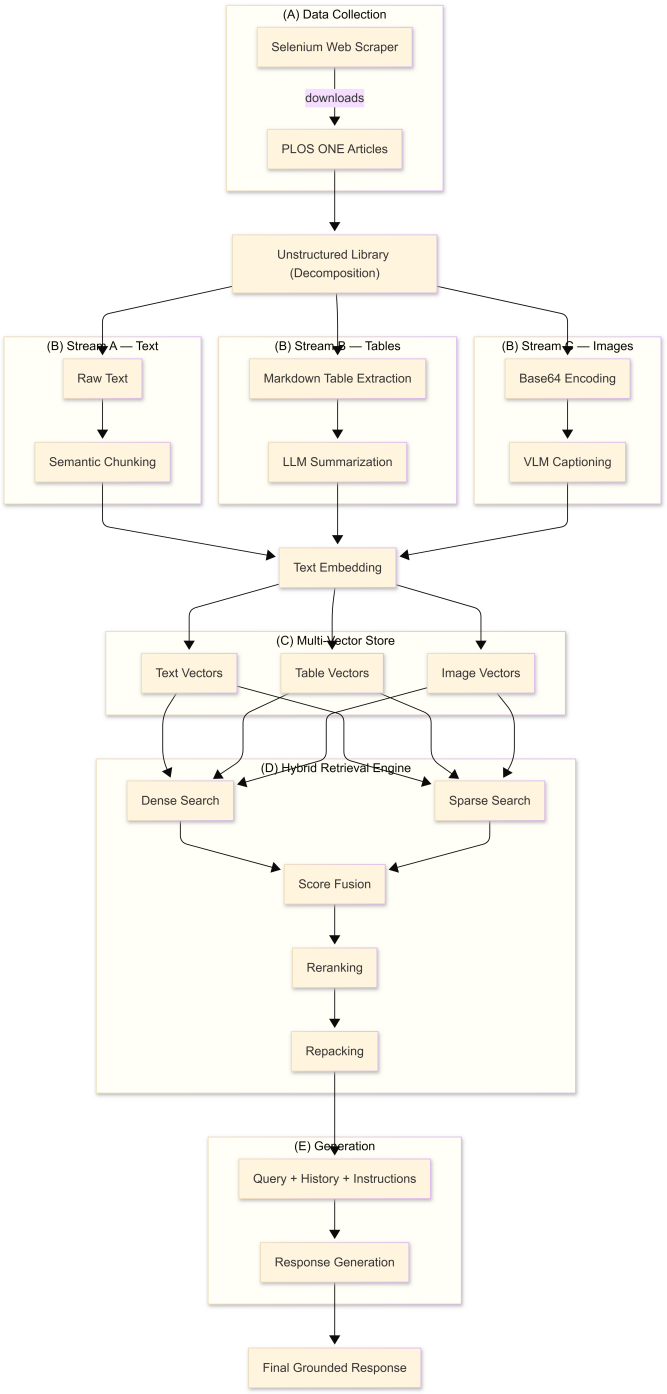

Fig. 2. Multimodal processing pipeline. PDFs are split into text, table, and image streams. Text is chunked (512 tokens, 64-token overlap); tables are summarized by an LLM; images are captioned by Gemini 2.0 Flash. All streams feed into a shared ChromaDB vector store.

2) Hybrid Retrieval: Menopause queries combine colloquial language (e.g., "hot flashes") with clinical terminology (e.g., "vasomotor episodes"). Dense retrieval handles semantic similarity but can miss exact medical terms; BM25 captures term-level matches but lacks contextual understanding. We combine both via a weighted score:

$$S_{\text{final}} = \alpha \cdot S_{\text{dense}} + (1 - \alpha) \cdot S_{\text{sparse}} \tag{1}$$

where $\alpha = 0.7$ was selected via grid search over $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$ on a 20-query validation set. The BM25 component follows the standard formulation [13]:

$$\text{BM25}(t, d) = \text{IDF}(t) \cdot \frac{\text{tf}(t, d) \cdot (k_1 + 1)}{\text{tf}(t, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \tag{2}$$

where $k_1 = 1.5$ and $b = 0.75$ are standard defaults. Table II presents the evaluation of retrieval strategies, with the α grid search results merged as a footnote. The hybrid approach outperformed dense-only by 2.4 points in precision while recovering recall lost to vocabulary mismatch.

TABLE II
COMPARATIVE EVALUATION OF RETRIEVAL STRATEGIES. $\alpha = 0.7$ SELECTED VIA GRID SEARCH (BEST: PREC. 94.1%, REC. 89.5%).

Approach	Prec. (%)	Rec. (%)	Time
Hybrid (Dense+BM25)	**94.7**	**89.2**	12.8s
Dense only (BGE-large)	92.3	82.1	7.3s
BM25 only	76.8	91.4	5.5s
TF-IDF	71.2	85.3	3.2s

E. Reranking and Repacking

The top-k retrieved chunks pass through the ms-marco-MiniLM-L-6-v2 cross-encoder that scores query–document pairs jointly. Among three cross-encoder models evaluated, ms-marco-MiniLM-L-6-v2 achieved the best NDCG (0.847) and MRR (0.792) at 13.2s latency. It identifies medical nuances that bi-encoders miss, justifying the latency trade-off.

Repacking applies semantic clustering (using BGE-large vectors) to ensure diverse, non-redundant context before generation. This step achieves a Diversity Score of 9.2/10 versus 6.8/10 for naive top-k selection (Table III). Coherence and diversity were scored on a 1–10 scale by two independent annotators (inter-rater agreement: $\kappa = 0.81$).

TABLE III
COMPARATIVE EVALUATION OF REPACKING STRATEGIES

Strategy	Coherence	Diversity
Semantic clustering	**8.7/10**	**9.2/10**
Top-k selection	7.3/10	6.8/10

F. Patient-Facing Application Interface

The framework is deployed as an interactive application comprising four modules: (1) a **Conversational RAG Interface** for free-form medical queries; (2) a **Longitudinal Symptom Tracker** collecting 0–10 severity ratings across seven primary menopausal symptoms; (3) a **Clinical Report Generator** that synthesizes tracked symptoms and AI-grounded recommendations into a downloadable PDF, bridging clinical dialogue and enabling patients to present structured symptom histories to physicians; and (4) **Educational Dashboards** offering curated clinical cards from the corpus. A persistent disclaimer defines the system as an educational support tool, not a substitute for professional medical diagnosis.

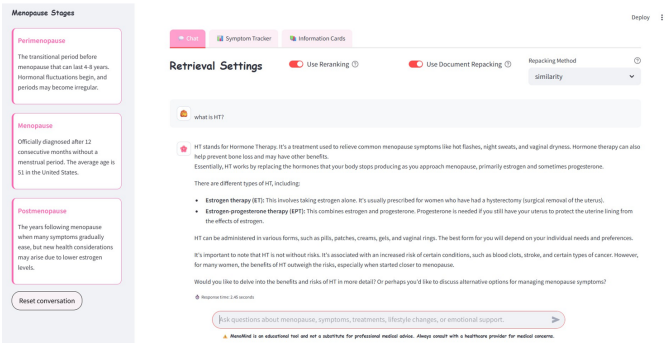

Fig. 3. Application interface with four tabs: (1) RAG chatbot (top-left), (2) symptom severity tracker (top-right), (3) PDF report generation panel (bottom-left), and (4) educational cards (bottom-right).

IV. Experimental Evaluation

We evaluated the system using two complementary approaches: (1) automated metrics via the RAGAS framework on a held-out query set, and (2) a preliminary user study using the System Usability Scale.

A. Quantitative Metrics (RAGAS)

RAGAS was run on 50 manually written queries covering clinical themes of the corpus (symptoms, mechanisms, treatments, lifestyle, and psychological impact). The query set includes colloquial and clinical phrasings as well as three complexity tiers: simple factual (40%), multi-aspect (40%), and comparative (20%). Three metrics were computed:

Faithfulness measures the fraction of claims supported by retrieved context:

$$F = \frac{|S_{\text{claims}} \cap S_{\text{context}}|}{|S_{\text{claims}}|} \tag{3}$$

Answer Relevance computes average cosine similarity between the original query and N synthetic questions generated from the answer:

$$\text{Relevance} = \frac{1}{N}\sum_{i=1}^{N} \cos(\vec{v}_q, \vec{v}_{q_{\text{gen},i}}) \tag{4}$$

Context Precision evaluates noise remaining in retrieved passages after reranking:

$$\text{CP@}k = \frac{\sum_{k=1}^{K}(\text{precision@}k \times v_k)}{\text{Total Relevant Items}} \tag{5}$$

We note that RAGAS uses an LLM as judge [17], which introduces known circular dependency. Confidence intervals were estimated by bootstrapping over 1,000 resamples. Results are summarized in Tables IV and V.

TABLE IV
Quantitative Performance on Test Dataset ($N = 50$ queries; 95% CI from bootstrapping)

Metric	Score	95% CI
Faithfulness	88%	±3.1%
Answer Relevance	90%	±2.8%
Context Precision	85%	±3.6%
Avg. Response Time	15.2s	—

TABLE V
RAGAS Metrics by Query Complexity Tier ($N = 50$)

Query Type	Faith.	Relev.	Ctx. Prec.
Simple factual (40%)	92%	93%	89%
Multi-aspect (40%)	86%	89%	83%
Comparative (20%)	81%	84%	78%
Overall (weighted)	**88%**	**90%**	**85%**

Error Analysis. For the comparative query *"Is HRT safer than SSRIs for mood symptoms?"*, the system retrieved relevant passages but generated a partially hedged response introducing a claim not explicitly supported by context, lowering faithfulness to 0.75—consistent with the lower average for comparative queries in Table V.

B. User Study

We recruited five participants: three women (aged 45–58) experiencing menopausal symptoms, and two health informatics students (aged 26–31). Participants completed three task-based sessions followed by the SUS questionnaire [18]. The average SUS score was 82/100 ("Excellent", Grade A [19]). Table VI summarizes both quantitative scores and qualitative observations. We acknowledge that $N = 5$ is a preliminary sample establishing a baseline for future larger-scale clinical validation.

V. Discussion

The high faithfulness observed (Table IV) is a direct consequence of restricting the corpus to peer-reviewed articles and applying cross-encoder reranking. The strong relevance scores reflect hybrid retrieval's ability to handle terminology variation in patient language. Comparative queries remain the most challenging tier (Table V), motivating future work on multi-hop reasoning via knowledge graph integration.

The total end-to-end latency of 15.2s includes retrieval (∼12.8s) and reranking (∼13.2s); pipeline parallelism reduces

TABLE VI
USER STUDY RESULTS ($N = 5$; SCALE 1–5 FOR USABILITY METRICS)

Component	Score	Strengths	Limitations
Usefulness	4.8/5	Highly relevant	—
Clarity	4.6/5	Accessible language	Occasionally technical
Navigation	4.5/5	Intuitive interface	—
Chatbot	—	Empathetic, structured	Struggles with vague queries
Symptom Tracker	—	Useful PDF export	No longitudinal trend view
Multimodal	—	Tables well integrated	Image display needs work
SUS	**82/100**	**Grade A**	**Excellent**

the theoretical sequential time of ∼26s to the observed value. In a medical decision-support context, accuracy and factual grounding take precedence over conversational speed; the system targets asynchronous, thoughtful patient inquiry rather than rapid chat. Future work will explore quantized reranking models to reduce this overhead.

Table VII compares our results with Handy et al. [4] as the primary external benchmark. Our system achieves comparable performance to GPT-4 while remaining locally deployable and incorporating multimodal data.

TABLE VII
COMPARISON WITH STATE-OF-THE-ART SYSTEMS

System	Faith.	Relev.	Modalities	Domain
Handy et al. [4] (GPT-4)	88.6%	91.8%	Text only	Menopause
Ours	**88%**	**90%**	Text + Tables + Images	Menopause

Limitations. The corpus is static and requires manual re-execution of the ingestion pipeline for updates. Image retrieval relies on text captions rather than direct visual reasoning. The user study ($N = 5$) limits statistical generalization. RAGAS evaluation uses an LLM as judge, creating a known circular dependency [17].

VI. CONCLUSION

We built a multimodal RAG assistant for menopause health information centered on source quality over scale. Restricting retrieval to a curated PLOS ONE corpus, combining dense and sparse search with an empirically validated fusion weight ($\alpha = 0.7$), reranking with a cross-encoder, and captioning clinical images with Gemini 2.0 Flash yielded 88% faithfulness and 90% relevance on a domain-specific query set. A usability study returned a SUS score of 82/100. An LLM query classifier routes requests with 96% accuracy, and per-complexity analysis reveals comparative queries as the most challenging frontier.

Future work will pursue multilingual support to reach non-English-speaking users. We plan to integrate a biomedical knowledge graph such as PrimeKG [20] to support multi-hop clinical reasoning. Additional priorities include longitudinal symptom tracking with trend detection, a native multimodal model for direct visual reasoning, and quarterly automated corpus updates. To mitigate the inherent bias of LLM-as-judge evaluation (RAGAS), future validation phases will incorporate partial human evaluation with medical professionals to establish clinical ground truth. The evaluation dataset will also be expanded to a larger and more diverse query set to better characterize system robustness across edge-case and colloquial phrasings.

Ethical Note. The preliminary user study was conducted following institutional guidelines, with informed consent obtained from all participants and no retention of personal health data. Symptom tracking is processed locally without persistent server storage. The system is explicitly designed as an educational decision-support tool; it displays a persistent disclaimer advising users to consult healthcare professionals and is not a substitute for professional medical diagnosis.

REFERENCES

[1] N. Santoro, C. Roeca, B. A. Peters, and G. Neal-Perry, "The Menopause Transition: Signs, Symptoms, and Management Options," *Journal of Clinical Endocrinology & Metabolism*, vol. 106, no. 1, pp. 1–15, 2021.

[2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2020.

[3] Z. Guo et al., "Evaluating Large Language Models on Medical Evidence Summarization," *npj Digital Medicine*, vol. 6, p. 158, 2023.

[4] A. Handy, M. Smith, and L. Johnson, "An Intelligent Chatbot for Menopause: Evaluation Using GPT-4 and Medical Guidelines," *arXiv preprint arXiv:2502.03579*, 2025. [Preprint; not yet peer reviewed.]

[5] Public Library of Science, "PLOS ONE," 2024. [Online]. Available: https://journals.plos.org. [Accessed: May 22, 2024].

[6] B. Zhu, R. Wang, and J. Wang, "EMERGE: Integrating RAG for Improved Multimodal EHR Predictive Modeling," *arXiv preprint arXiv:2410.04660*, 2024.

[7] Y. Liu, P. Wang, and Y. Chen, "Efficacy of Retrieval-Augmented Generation (RAG) Systems in Clinical Practice: A Systematic Review," *Journal of Medical AI*, vol. 1, no. 2, pp. 123–135, 2024.

[8] P. Xia et al., "MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models," *arXiv preprint arXiv:2410.13085*, 2024.

[9] A. K. Lahiri and Q. V. Hu, "AlzheimerRAG: Multimodal Retrieval Augmented Generation for PubMed Articles," *arXiv preprint arXiv:2412.16701*, 2023.

[10] J. Sohn, Y. Kim, and D. Lee, "RAG2: Retrieval-Augmented Generation with Rationales for Enhanced Medical QA," in *Proc. IEEE Int. Conf. on Healthcare Informatics (ICHI)*, 2023.

[11] Google DeepMind, "Gemini: A Family of Highly Capable Multimodal Models," *arXiv preprint arXiv:2312.11805*, 2023.

[12] Google DeepMind, "Gemini 2.0 Flash Model Card," *Google DeepMind Model Cards*, 2025. [Online]. Available: https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf

[13] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[14] X. Jiao et al., "TinyBERT: Distilling BERT for Natural Language Understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4163–4174.

[15] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei, "MiniLMv2: Multi-head Self-Attention Relation Distillation for Compressing Pretrained Transformers," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2140–2151.

[16] N. Reimers and I. Gurevych, "cross-encoder/stsb-distilroberta-base: Cross-Encoder for Semantic Textual Similarity," *Hugging Face Model Hub*, 2019. [Online]. Available: https://huggingface.co/cross-encoder/stsb-distilroberta-base

[17] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation," *arXiv preprint arXiv:2309.15217*, 2023.
[18] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation in Industry*, P. W. Jordan et al., Eds. Taylor & Francis, 1996, pp. 189–194.
[19] A. Bangor, P. Kortum, and J. Miller, "Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale," *Journal of Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.
[20] P. Chandak, K. Huang, and M. Zitnik, "Building a Knowledge Graph to Enable Precision Medicine," *Scientific Data*, vol. 10, p. 67, 2023.
[21] S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff, "C-Pack: Packaged Resources To Advance General Chinese Embedding," *arXiv preprint arXiv:2309.07597*, 2023.
[22] L. Wang et al., "Text Embeddings by Weakly-Supervised Contrastive Pre-training," *arXiv preprint arXiv:2212.03533*, 2022.
[23] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and Permuted Pre-training for Language Understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 16857–16867, 2020.
[24] W. Wang et al., "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 5776–5788, 2020.