MenoEaze: An Empathetic Retrieval-Augmented Generation Companion for Menopause Management

Abhishek Suraj
Galgotia College of Engineering & Technology
abhisheksuraj221@gmail.com

Kashika Khurana
Galgotia College of Engineering & Technology
kashika.22gcebds020@galgotiacollege.edu

Harsh Jain
Galgotia College of Engineering & Technology
harsh.22gcebds053@galgotiacollege.edu

Harsh Mishra
Galgotia College of Engineering & Technology
harsh.22gcebds069@galgotiacollege.edu

Abstract* – *Menopause is a profound biological and psychological transition for women, and yet it is markedly under-served in the digital health sector, represented largely by limited, non-empathetic, and non-evidence-based digital support. Most existing solutions focus on passive data tracking without personalized empathetic guidance. It creates a clinical trust deficit and increases emotional isolation. This paper introduces MenoEaze, an AI-powered conversational avatar grounded in a state-of-the-art Retrieval-Augmented Generation architecture coupled with an advanced set of empathic design principles. The core functionality of the system is to provide highly individualized, evidence-based nutritional and emotional support for midlife women affected by menopausal symptoms. Anchoring responses in a curated knowledge base of authoritative medical sources, the RAG architecture suppresses the critical issues of hallucination and inconsistency of generalized LLMs. MenoEaze will be designed as a modular approach so that contextual and emotional awareness can be achieved to address the important ethical challenges regarding bias, interpretability, and the critical requirement for perceived authenticity in human-AI interaction. This will include rigorous multi- stage evaluation against the S.C.O.R.E. framework- Safety, Consensus, Objectivity, Reproducibility, and Explainability for the developed system to ensure clinical accuracy and trustworthiness superior to the generic chatbots. MenoEaze aims at transforming menopause care into a proactive, supportive emotionally intelligent experience from being a reactive, data-centric service.

Keywords—*Retrieval-Augmented Generation (RAG), Menopause, Empathetic AI, Reinforcement Learning from Human Feedback (RLHF), S.C.O.R.E. Framework.*

1. INTRODUCTION

Menopause generally occurs between the ages of 45 and 55 years and is characterized by a prolonged loss of estrogen that presents a variety of physiological and psychological symptoms, which include disturbed sleep, mood swings, weight fluctuations, and long-term risks to cardiovascular disease and osteoporosis alike. This is expected to involve more than 1.2 billion postmenopausal women by 2030-an area of crucial but underserved importance to global healthcare. The digital health revolution that took place worldwide, while highly effective in areas like fitness tracking, still lags significantly in providing tailored, reliable, and emotionally intelligent support for this stage in life. The current digital solutions serve primarily as passive symptom trackers and do not solve for the complex emotional and nutritional needs linked to hormonal changes. Generalized LLM-based chatbots, while widely available, fall short fundamentally when applied to this sensitive, high-stakes clinical domain, with particular underperformance in areas of clinical detail and factual grounding [1], [4], while evaluations of menopause-focused chatbots report deficiencies in empathetic performance and trustworthiness [11]. Assessments through the S.C.O.R.E. framework [10] have highlighted cases where generalized LLMs provided overly simple or clinically incongruent recommendations, lacked cited sources, and even demonstrated subtle biases based on hypothetical patient characteristics [23].

The MenoEaze project is designed to overcome these critical failures through the development of an AI-powered conversational companion that embeds computational rigor in empathetic intelligence [14]. In the center is the application of a Retrieval-Augmented Generation (RAG) system [5], [7], which anchors every recommendation into a verifiable, up-to-date knowledge base of medical guidelines [6], with a view to ensuring factual accuracy and clinical safety [15]. Further, MenoEaze integrates mechanisms for detecting the user's emotional state from her conversation to offer contextual and compassionate support [1], [3] in pursuit of a highly personalized and trustworthy resource that leaves the woman informed and empowered. This proposal delineates the rationale, architecture, and validation strategy to position MenoEaze as a transformational platform for the digital health of women in midlife.

2. LITERATURE REVIEW

The development of MenoEaze is based on the critical review in three interrelated domains: the need for and limitation of AI regarding women's health, the technical landscape of Retrieval-Augmented Generation architectures, and the complex challenge of developing authentic artificial empathy.

A. AI in Women's Health and the Trust Deficit

AI applications in women's health are rapidly moving beyond simple risk prediction, such as for cardiovascular disease or osteoporosis, toward integrated, holistic care [20]. Equitable and reliable AI support is still lacking during menopause [22]. Current digital solutions tend to be basic symptom trackers that often lack necessary clinical depth and personalized guidance [22].

Evaluations of general purpose Large Language Models responding to menopause queries underscore the critical need for a specialized system like MenoEaze:

- **Safety and Consensus:** General LLMs tended to have deficiencies, usually providing scant details about complex hormonal changes, and sometimes clinically inconsistent responses [12], not aligned with the current medical consensus [10].
- **Objectivity and Bias:** A comprehensive review unveiled subtle, yet important, biases. There were instances of insurance- and race-based discrimination. Insured users received more detailed information compared to the uninsured user simulations, which is of fundamental concern regarding equitable access to evidence-based guidance [23], [17].
- **Explainability:** These models often failed to provide credible source attribution or justify multi-component advice, leading to a profound clinical trust deficit [14], [13]. These findings confirm that a specialized and highly regulated approach is required so as not to worsen existing health disparities by using AI [17].

B. Need for RAG: Retrieval-Augmented Generation

RAG stands as the recognized state-of-the-art solution to counteract the core flaws of standalone LLMs: hallucination, lack of detail, and outdated information [5, 6]. RAG transforms a generalized LLM into a domain-specific expert by combining the model's fluency with the factual rigor of external knowledge.

a. **Core Function and Architecture:** RAG conditions a pre-trained LLM on verifiable context retrieved from an authoritative Knowledge Base KB [5, 9]. This is achieved by chunking source documents into pieces, converting them into vector embeddings [16], and indexing them to enable fast semantic retrieval based on the user's query [5].

b. **Mitigation of Hallucination:** In high-stakes environments such as medicine, the reduction of hallucination must be paramount. The following studies show that RAG dramatically increases clinical performance, recording up to an 18% increase in medical QA accuracy over prompt-based baselines and drastically reducing the factual inconsistency rate observed in base models [8], [9].

c. **Advanced RAG Architecture:** The MenoEaze architecture goes beyond Naive RAG [7]. The modular nature of the architecture thus allows the integration of specialized query rewriters, intent classifiers, and re-ranking modules, ensuring that only high-quality contexts relevant to the user's query reach the generative LLM [7, [15]. It is in this attention to the quality of retrieval that the process aims to maintain clinical safety and trustworthiness [13].

C. Empathy and Authenticity in Chatbot Design

Accuracy is solved by RAG, while for MenoEaze to function as a compassionate companion, it needs to solve the problem of perceived empathy and trust. [1]

1. **AI Empathy Gap:** Effective patient engagement is critically dependent on empathetic communication [1]. For chatbots, however, this is nuanced. Although LLMs can have similar or even superior scores compared to human clinicians in text-based empathy measures under blinded conditions [4], [19], this benefit is usually superficial. Crucially, the positive effect on trust disappears once users are informed they are interacting with an AI [2].
2. **Perceived Inauthenticity:** The direct and over-expressed emotional language used by an AI may appear inauthentic and may actively serve to undermine user trust by violating user expectations about the capabilities of a machine, according to the "Mind Perception Theory" [2].
3. **Behavioral Empathy Solution:** The most effective approach would be to manifest empathy through instrumental support or behavioral empathy [1]-such as asking clarifying questions, demonstrating intent to help, or providing structured, detailed, and relevant information [14], [2]. This then aligns the AI's "warmth" with its technical competence, hence perceived authenticity.
4. **Emotional Awareness Integration:** For this purpose, a hybrid/modular design will be required. This shall consist of an upstream analysis of non-lexical cues, such as sentiment analysis (NLP Intent Detection) on text and possibly on voice/pitch data [3], [15], to get an accurate reading of the emotional quotient of the user and adjust the tone of the RAG-grounded response accordingly [14].

3. METHODOLOGY

The MenoEaze methodology adopts a rigorous, three-part framework designed to overcome the core technical (accuracy/hallucination) and non-technical (empathy/trust) challenges identified in existing literature [10], [13].

A. System Architecture: Modular RAG

The core architecture will rely on a Modular RAG design [7] to

allow for maximum customization, high transparency, and targeted integration of specialized components for empathy and privacy.

- **Frontend (React/Next.js):** Light, intuitive user interface, optimized for various age groups and nonspecialist users, including integration with symptom tracking [22].
- **Backend (Node.js/Python):** Handles API communication and mediates the interaction with the RAG engine and with a persistent database, either PostgreSQL or Firebase, to store user data.
- **Knowledge BaseKB & Indexing:** Aggregated repository of authoritative medical sources - guidelines, peer-reviewed journals [8] - structured through vector embeddings [16], to support efficient semantic search [5].
- **NLP EngineIntent Detection:** This upstream module processes user queries to classify the intent of users, for example, nutritional query, symptom description, or emotional support query [14], and key entities such as mood or sleep issues.
- **Retrieval System:** It conducts a vector-based search in the KB [5] with the refined query/intent for retrieving the best and contextually appropriate documents.
- **Generation System:** A strong LLM-e.g., fine-tuned open-source model such as Mistral-7B-is specifically conditioned on retrieved, verified clinical context to create a response [7], while minimizing hallucination and explicitly citing sources for transparency [15], [13].

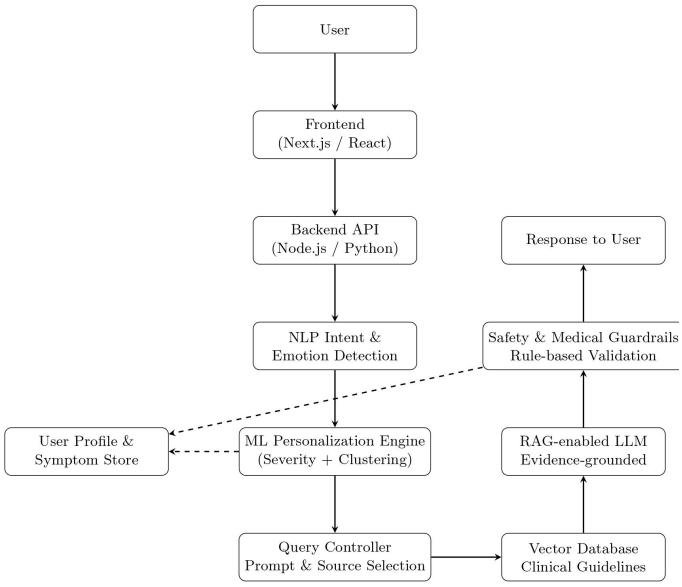

Figure 1: Modular AI system architecture of MenoEaze integrating intent and emotion detection, personalization, and RAG over clinical knowledge sources.

B. The Alignment Strategy: PPO for Safety and Empathy

To do so with a response style aligned with the twin goals of clinical safety and perceived empathy [1], it is envisioned to employ a PPO-based RLHF strategy [19], [18]:

a. **PPO-Policy:** The fine-tuned, RAG-enabled LLM is the policy model that gets updated using gradient steps with respect to the reward signal [19].

b. **PPO-Reward Model (RM):** A separate model will be trained to act as the "Empathy Critic," which predicts a scalar reward for a response based on human-annotated feedback, focusing on:
a. Safety/Correctness: Conformity to retrieved clinical evidence**.**
b. Behavioral Empathy: Convey concern and a wish to help without excessive disclosure of pseudo-emotional states [2].

c. Ethical Check-and-Balance: PPO enables MenoEaze to impose a penalty on the KL-distance during optimization, such that the model's new empathetic behavior does not cause it to drift too far from its clinically-focused original
policy, therefore maintaining stability and avoiding over-optimization of the model [19].

C. Bias Mitigation and Privacy

MenoEaze will address issues of fairness and ethical considerations throughout its development proactively :

I. **Privacy-by-Design:** Base principles include end-to-end encryption and differential privacy techniques to handle and aggregate sensitive user data.

II. **Bias Auditing:** Testing algorithmic fairness will be included in finding and mitigating race and insurance-based biases during RLHF training to make high-quality advice [23] equally accessible [17].

III. **Diverse Feedback:** The RLHF process will make use of a diverse set of human evaluators to ensure variance in the values captured within the reward model, avoiding inadvertently promoting single-perspective biases observed in homogeneous feedback groups.

4. EXPERIMENTAL SETUP AND RESULTS

As this proposal outlines, the following section presents the planned experimental design and anticipated outcomes required to assess the MenoEaze platform's alignment with established safety and empathy standards [13].

A. Experimental Setup

The study will employ a mixed-methods design that integrates computational benchmarking with structured human-in-the-loop evaluations. This approach allows for a direct comparison between MenoEaze and generalized large language model baselines [13], [23].

- **Baseline Models:**

The evaluation will incorporate widely used

LLMs (e.g., GPT-4, Meta AI) alongside standard RAG configurations to establish reference points for system performance [9], [13].

- **Datasets:**

  A curated dataset encompassing a broad range of menopause-related queries. Spanning nutritional, emotional, and clinical topics; will serve as the foundation for testing [23], [20].

- **Metrics:**

  The platform will be assessed through a multi-layered evaluation framework designed to capture both quantitative and qualitative indicators of system performance [10].

B. **Evaluation Framework and Metrics**

MenoEaze's performance will be evaluated through a structured, three-stage process designed to measure clinical reliability, user trust, and overall conversational quality.

1. **Clinical Accuracy & Safety (Computational Assessment)**

   - **S.C.O.R.E. Framework:** The system will be evaluated using the S.C.O.R.E criteria: Safety, Consensus, Objectivity, Reproducibility and Explainability to ensure compliance with established clinical and computational standards [10].

   - **Automated Metrics:** A set of automated and reference-based measures will be applied to validate factual correctness and determine the effectiveness of the platform's retrieval-augmented generation (RAG) pipeline.

     a. **FactScore / Faithfulness:** These metrics will quantify hallucination rates by verifying the model's generated responses against the retrieved knowledge base documents, ensuring alignment with source-grounded evidence [7].

     b. **Semantic Similarity / ROUGE:** These measures will assess the system's coherence and its alignment with expert- crafted medical responses, capturing both linguistic and conceptual fidelity [11].

2. **Empathy & Trust (Human-in-the-Loop Evaluation)**

   - **Expert Review:** Licensed medical professionals will evaluate a selected subset of system outputs to assess their clinical soundness, contextual appropriateness, and adherence to safety guidelines [13].

   - **User Validation (Perceived Empathy):** End-users will rate the chatbot's empathetic communication using validated measures such as adapted versions of the CARE or PEI scales—rather than subjective, unstructured impressions [1], [3], [19].

   A key objective of this phase is to determine whether the system's Behavioral Empathy strategy yields higher user-perceived authenticity compared to models that rely on explicit or overly affective expressions of empathy [2].

C. **Expected Results**

The MenoEaze system is projected to outperform comparative models across several key domains central to safety, empathy, and usability.

1) **Clinical Reliability:**

MenoEaze is expected to achieve consistently higher performance on the S.C.O.R.E. criteria [10], supported by its retrieval-augmented generation architecture. This grounding is expected to reduce the incidence of serious factual errors or clinically unsafe hallucinations to near zero, reinforcing the platform's reliability in delivering evidence-based guidance [8].

2) **Empathy and Trust:**

Through reinforcement learning from human feedback (RLHF), the system is designed to produce responses that users perceive as more authentic, trustworthy, and genuinely supportive [2], [14]. Its emphasis on behavioral empathy—demonstrating understanding through clarity, guidance, and responsiveness aims to outperform models that rely primarily on overt emotional expressions [1].

D. Expected Results

The MenoEaze system is projected to outperform comparative models across several key domains central to safety, empathy, and usability.

3) Clinical Reliability:

MenoEaze is expected to achieve consistently higher performance on the S.C.O.R.E. criteria [10], supported by its retrieval-augmented generation architecture. This grounding is expected to reduce the incidence of serious factual errors or clinically unsafe hallucinations to near zero, reinforcing the platform's reliability in delivering evidence-based guidance [8].

4) Empathy and Trust:

Through reinforcement learning from human feedback (RLHF), the system is designed to produce responses that users perceive as more authentic, trustworthy, and genuinely supportive [2], [14]. Its emphasis on behavioral empathy—demonstrating understanding through cl

5) Efficiency:

Finally, the system is expected to balance its safety and empathy objectives with strong operational performance. By optimizing latency and computational efficiency, MenoEaze should maintain the responsiveness required for a seamless conversational experience [16].

5. DISCUSSION

The MenoEaze initiative directly targets the most significant shortcomings in current AI applications within women's health, as identified across recent literature [20], [22]. A core design challenge lies in balancing two essential but competing priorities: ensuring high factual accuracy—driven primarily by the system's RAG engine [5]—and cultivating an emotionally supportive, trustworthy conversational experience through empathetic design and RLHF-based alignment [1], [19].

The adoption of a Modular RAG architecture [7] is essential for any system intended for clinical use. Generic LLMs, despite their broad capabilities, continue to pose unacceptable risks due to their propensity for generating inaccurate or unverified information—a critical safety vulnerability in medical contexts [8], [13]. By requiring that all responses be grounded in authoritative and traceable source material [6], MenoEaze establishes a safety foundation that conventional chatbots cannot reliably provide. This design also strengthens Explainability by enabling transparent source attribution, which is central to building trust among both users and healthcare professionals [10], [14].

Equally important is the system's emphasis on cultivating authentic artificial empathy [2]. Research consistently shows that while AI can simulate emotional tone, excessive or overtly affective expressions can reduce credibility by appearing disingenuous, ultimately undermining user trust [2]. MenoEaze mitigates this risk by aligning its behavioral style—via PPO optimization and fine-grained human feedback [19]—toward practical support, clarity, and helpful intent [1]. This strategy ensures that empathetic communication enhances, rather than detracts from, the system's reliability and perceived sincerity [14].

Lastly, the project's commitment to a multidimensional evaluation framework, grounded in the S.C.O.R.E. criteria [10], elevates the standards for clinical AI validation. By integrating rigorous quantitative safety metrics with assessments of Objectivity and Bias [17], the evaluation process foregrounds equity from the outset. This proactive approach helps counteract systemic biases commonly found in large, uncurated training corpora [18], thereby supporting safer, more inclusive AI-driven care.

6. CONCLUSION AND FUTURE WORK

The MenoEaze project introduces a novel Modular RAG–based companion system [7], aligned through a targeted PPO-RLHF training process [19], to provide accurate, personalized, and empathetic support for women experiencing menopause [20], [23]. The design explicitly addresses two long-standing challenges in digital health: reducing clinical inaccuracies through robust RAG grounding [8] and strengthening emotional trust through adaptive, behaviorally informed empathetic modeling [1], [2]. If successful, the system is expected to deliver a highly dependable digital health resource that surpasses existing tools across core metrics of safety, factual accuracy, and user engagement [10], [13].

Future work includes:

- **Multimodal Integration:**

  Expanding the current RAG pipeline to incorporate non-textual signals—such as vocal features including pitch and tone—to generate a more nuanced understanding of users' emotional states. This enhancement aims to elevate the quality and authenticity of empathetic responses beyond purely lexical cues [3], [15].

- **Continual Learning:**

  Implementing dynamic retrieval and continual learning mechanisms [7] to ensure the knowledge base remains up-to-date and that the model's policy adapts in real time to new clinical research and user feedback. This approach supports sustained accuracy and long-term system relevance [6].

- **Long-term Clinical Validation:**

  Planning a comprehensive randomized controlled trial (RCT) involving real users to evaluate long-term impacts on clinical outcomes, adherence to health guidance, and the durability of user trust and engagement over time [13], [12].

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge- Intensive NLP Tasks," Advances in Neural Information Processing Systems, vol. 33, 2020.

[2] N. Lambert, T. K. Gilbert, and T. Zick, "Entangled preferences: The history and risks of reinforcement learning and human feedback," arXiv preprint arXiv:2310.13595, 2023.

[3] N. Lambert et al., "Reinforcement Learning from Human Feedback: A short introduction to RLHF and post-training focused on language models," arXiv preprint arXiv:2404.12501v3, 2025.

[4] R. Yang et al., "Retrieval-Augmented Generation for Generative Artificial Intelligence in Medicine," arXiv preprint arXiv:2406.12449, 2024.

[5] T. F. Tan et al., "A Proposed S.C.O.R.E. Evaluation Framework for Large Language Models—Safety, Consensus, Objectivity, Reproducibility and Explainability," arXiv preprint arXiv:2407.07666v1, 2024.

[6] F. Neha, D. Bhati, and D. K. Shukla, "Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review," AI, 2025.

[7] L. M. Amugongo et al., "Retrieval augmented generation for large language models in healthcare: A systematic review," PLOS Digital Health, 2025.

[8] F. Gkrozou et al., "Mobile Health (mHealth) apps focused on menopause: Are they any good?" Post Reproductive Health, 2019.

[9] A. Bora and H. Cuayáhuitl, "Systematic Analysis of Retrieval- Augmented Generation-Based LLMs for Medical Chatbot Applications," Machine Learning and Knowledge Extraction, 2024.

[10] G. A. R. Panjwani et al., "Artificial Intelligence in Postmenopausal Health: From Risk Prediction to Holistic Care," Journal of Clinical Medicine, 2025.

[11] R. Deva et al., "A Mixed-Methods Evaluation of LLM-Based Chatbots for Menopause," arXiv preprint arXiv:2502.03579v1, 2025.

[12] V. M. Vargas-Hernandez, "Artificial Intelligence in Menopause Management," Mathews Journal of Gynecology & Obstetrics, 2025.

[13] O. Kohandel Gargari and G. Habibi, "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," Digital Health, 2025.

[14] R. Sanjeewa et al., "Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review," JMIR Mental Health, 2024.

[15] R. Sanjeewa et al., "Perception of Empathy in Mental Health Care Through Voice-Based Conversational Agent Prototypes: Experimental Study," JMIR Formative Research, 2025.

[16] L. Seitz, "Artificial empathy in healthcare chatbots: Does it feel authentic?" Computers in Human Behavior: Artificial Humans, 2024.

[17] P. Soubhagyalakshmi et al., "A Healthcare Chatbot Powered by Retrieval-Augmented Generation (RAG)," International Journal of Creative Research Thoughts (IJCRT), 2025.

[18] F. Busch et al., "Evaluation of a Retrieval-Augmented Generation-Powered Chatbot for Pre-CT Informed Consent: A Prospective Comparative Study," Journal of Imaging Informatics in Medicine, 2025.

[19] A. Howcroft et al., "AI chatbots versus human healthcare professionals: a systematic review and meta-analysis of empathy in patient care," British Medical Bulletin, 2025.

[20] K. Benkirane, J. Kay, and M. Perez-Ortiz, "How Can We Diagnose and Treat Bias in Large Language Models for Clinical Decision-Making?" arXiv preprint arXiv:2410.16574v1, 2024.

[21] K. González Barman, S. Lohse, and H. W. de Regt, "Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives?" Philosophy & Technology, 2025.

[22] M. Bhimani et al., "Real-World Evaluation of Large Language Models in Healthcare (RWE-LLM): A New Realm of AI Safety & Validation," medRxiv preprint, doi:10.1101/2025.03.17.25324157, 2025.

[23] V. Gumma, M. Jain, A. Raghunath, and S. Sitaram, "HEALTH-PARIKSHA: Assessing RAG Models for Health Chatbots in Real-World Multilingual Settings," arXiv preprint arXiv:2410.13671v1, 2024.