



Navigating the Healthcare Maze through Natural Language: HAAN An AI-Based Health Tool

Tareq M F Abdullah AlHammoodi, Kaan Özer, Hulki Enes Uysal, Kutluhan Erol, İlker Korkmaz

Faculty of Engineering
Izmir University of Economics
Izmir, Türkiye

{tareq.alhammoodi, kaan.ozer, enes.uysal}@std.ieu.edu.tr, {kutluhan.erol, ilker.korkmaz}@ieu.edu.tr

Abstract—Many people have a hard time navigating the healthcare systems, especially young people with little previous experience, or the elderly, who might get lost in the maze of mobile and web applications. The proposed system aims to provide user-friendly guidance through a chatbot platform utilizing pretrained Large Language Models (LLMs) combined with a Retrieval-Augmented Generation (RAG) in order to deliver reliable answers with references to their direct sources.

Whether someone wants to know about a certain procedure, or learn about health insurance coverage, the system is designed to make the process easier and more understandable for everyone. This supportive approach not only empowers individuals to take better care of their health, but also helps reduce the burden on healthcare providers. Ultimately, the project seeks to close the information gap and offer a reliable source of assistance for people navigating everyday health concerns, particularly within the context of the Turkish healthcare system.

Keywords—digital health tool, retrieval augmented generation (RAG), large language models (LLMs), e-health chatbots.

I. INTRODUCTION

Although modern healthcare systems have made significant advances in diagnosis and treatment, the process of actually using their services can be daunting for many people, including the young and the elderly, nonnative speakers, or others with limited digital literacy. For example, someone might want to get an eye exam, but does not know where to start. Should they go to a hospital, or a clinic? Are they allowed to see a specialist directly, or do they need a referral from their primary care physician? Do they need an appointment? What costs are covered by their insurance versus out-of-pocket expenses? How do they track the ever changing rules and regulations?

In Türkiye, as in many places, these challenges are compounded by a decentralized healthcare environment involving public hospitals, private clinics, family physicians, and specialized care centers, each one of them is governed by different procedures, insurance policies and appointment systems. While the Turkish Ministry of Health provides digital services such as MHRS (Centralized Hospital Appointment System) [1], e-Nabız (Personal Health Record) [2], and the Health Ministry Communication Center (SABİM 184) [3], these platforms often assume a level of familiarity that many users do not possess. This information gap can lead to unnecessary emergency visits or confusion.

This paper introduces HAAN (Health-Aware AI Assistant for Needs), an AI-powered chatbot that utilizes a Retrieval-Augmented Generation (RAG) framework to provide guidance for everyday health-related questions. The system is built on a Large Language Model (LLM) and trained with reliable health documents, ministry guidelines, and hospital protocols. HAAN provides users with answers related to appointments, tests, referrals, insurance procedures, and more. It also incorporates a feedback mechanism where users can rate and review responses, improving the accuracy and maintainability.

By bridging the gap between individuals and the services they need, HAAN is a step toward a more inclusive, efficient, and supportive healthcare environment in alignment with several UN Sustainable Development Goals (SDG) [4]. It advances SDG 3 (Good Health and Well-being) by improving access to healthcare knowledge, supports SDG 10 (Reduced Inequalities) by empowering underserved communities, and promotes SDG 16 (Peace, Justice, and Strong Institutions) by improving transparency in healthcare interactions.

II. LITERATURE REVIEW

In recent years, the integration of Artificial Intelligence (AI) into healthcare has gained significant momentum in improving patient care and clinical decision-making, as well as in expanding access to medical services. While AI applications in healthcare have historically focused on diagnosis and data analytics, patient-facing chatbots now represent a growing area of innovation. WHO has recognized the role of AI chatbots in increasing health awareness and early intervention, especially in underserved or digitally inexperienced populations [10]. These tools are also being incorporated into chronic disease management, vaccination scheduling, and mental health support, expanding their value beyond initial triage. They are especially valuable in facilitating communication between patients and healthcare systems, reducing the burden on professionals, and increasing health literacy [5].

Despite widespread adoption of electronic health records, usability and infrastructure challenges remain [6]. Several studies suggest that AI chatbots with conversational interfaces and structured guidance, such as Ada Health and Babylon Health, can successfully help bridge this gap [7, 8]. These systems use natural language processing (NLP) for symptom

assessments and recommendations, though many rely on static knowledge bases with infrequent updates. To address this, our proposed chatbot leverages a RAG architecture, combining vector search and LLMs to provide contextually accurate, up-to-date responses. This design, as supported by Amugongo et al. [9], reduces hallucinations and improves user trust in clinical settings.

Crowdsourced evaluation has also emerged as a valuable method for refining AI chatbot performance. Amershi et al. [11] emphasize the importance of crowdsourced evaluation in improving interactive AI systems, especially in real-world, user-centered applications. This approach invites users to rate chatbot responses based on usefulness or relevance, and optionally explain any issues they faced. These ratings are stored and analyzed to prioritize the most helpful answers in future conversations. As a result, each interaction contributes to improving the system's performance over time, creating a continuous feedback loop between users and AI.

Ultimately, AI chatbots based on RAG can advance digital health equity and offer scalable, cost-effective communication tools as the healthcare sector continues to digitize.

III. METHODOLOGY

This section presents the overall approach used to design and develop the system. The methodology focused on selecting effective AI techniques, building a relevant knowledge base, implementing a modular architecture, and validating the system through testing and user feedback.

After exploring various methods for building AI systems that deliver accurate and relevant responses, we chose RAG, as it integrates document retrieval with text generation, enabling AI to leverage a curated knowledge base. By grounding responses in retrieved documents, we aimed to reduce the risk of hallucinations and improve the factual reliability.

A. Data Collection and Knowledge Base Creation

In order to create a reliable knowledge base, we manually gathered hundreds of documents from trusted sources like the Ministry of Health, insurers, and hospitals and organized them into smaller, structured sections optimized for retrieval. To enable semantic search, we used an embedding model to convert each text chunk into a high-dimensional vector that captures its meaning as shown in Fig. 1. This allows the system to find relevant content based on semantics rather than just keywords, using models like nomic-embed-text or OpenAI's text-embedding-ada-002.

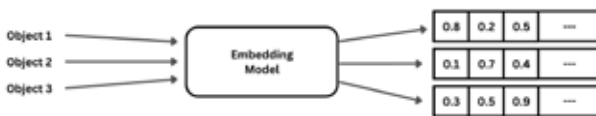


Fig. 1: Embedding representation.

These embeddings were then stored in a Vector Database (VDB), allowing efficient similarity search during the retrieval

phase. As represented in Fig. 2, when a user asks a question, the system uses the embedding of the question to find the most semantically similar document chunks from the VDB.

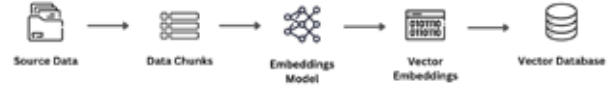


Fig. 2: Knowledge base preparation.

This is done by using a mathematical approach known as cosine similarity to compare the stored embeddings. Cosine similarity measures the cosine of the angle between two vectors in a multi-dimensional space, which reflects how similar they are in terms of direction rather than magnitude [12]. A cosine similarity score close to 1 indicates that the vectors are pointing in nearly the same direction, suggesting a high degree of similarity between the corresponding data points as shown in Fig. 3.

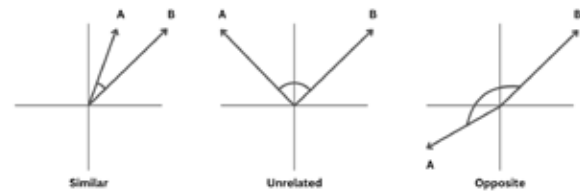


Fig. 3: Cosine similarity representation.

This makes cosine similarity particularly effective for comparing high-dimensional data like text embeddings, where the actual values may differ but the overall meaning or context is preserved in their directional alignment. This approach enables the chatbot to understand user queries even when they are phrased differently from the original documents.

B. System Architecture

The system was designed to offer a smooth and secure user experience. The front-end, built with React.js and styled using Tailwind CSS [13], provides an interactive chat interface where users can ask questions, see real-time updates, and leave feedback. It communicates with the backend securely using Axios. On the backend, Node.js with Express.js handles user authentication (via Google OAuth), session management, and chat storage using MongoDB [14]. A separate AI service, developed with Python and FastAPI, manages message processing and response generation by integrating tools like Ollama [15], Chroma DB [16], and LangChain [17]. The architecture follows a modular structure to ensure clarity, maintainability, and scalability across both the application and AI services.

The chatbot's core AI was built using Ollama, an open-source platform for running LLMs locally. Ollama was chosen for its support of retrieval-augmented workflows, efficient hardware usage, and easy model customization. We tested several mid-size models (e.g., Mistral 7B, LLaMA 3.2 derivatives)

and selected one optimized for conversational performance and low hallucination.

The answer generation process uses a RAG architecture with two main stages: retrieval and generation as shown in Fig. 4. In the document retrieval phase, the system first embeds the user's question into a dense vector using an embedding model. This vector is then used to perform a semantic search against a VDB to retrieve the top-k most relevant document chunks. These retrieved chunks are concatenated and formatted to build the context needed for answering.

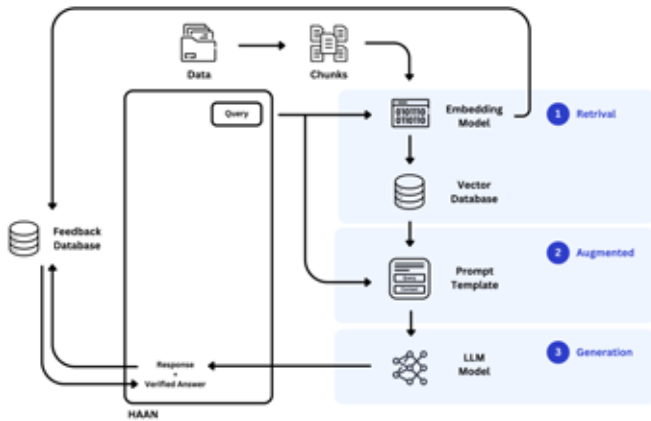


Fig. 4: RAG data flow diagram.

In the contextual generation phase, this context is passed to the Ollama model. A structured prompt is created that includes a system message, the retrieved documents, and the original user's question. This prompt is sent to the Ollama API for inference, where the model is configured with a low temperature (e.g., 0.2, meaning minimal randomness), which generates an answer based strictly on the provided content and reduces deviations from the source material. Fig. 5 depicts an example of the system prompt.

```
System Message: "You are a helpful assistant that only answers based on the following reference materials. If information is not available, respond with 'I'm sorry, but I don't have enough knowledge to answer this question.'"
```

```
Reference Materials:
```

```
[Data Chunk 1]
```

```
[Data Chunk 2]
```

```
[Data Chunk 3]
```

```
User Question:
```

```
[User's Input]
```

Fig. 5: A sample prompt structure.

This approach helps maintain answer reliability and minimize hallucination (i.e., the model making up facts). By grounding the AI's outputs in official documents, we make sure that users receive contextually appropriate advice tailored to their needs.

HAAN was designed with a focus on user experience. The design philosophy was rooted in simplicity and clarity. It is important to provide a system that is easy to interact with users from different backgrounds. Therefore, the interface

was deliberately kept focused on the following essentials: ask, get guidance, and act as seen in Fig. 6. It supports responsive design across various devices, aligning with the goal of simplifying access to healthcare services.



Fig. 6: HAAN user interface.

C. User Feedback and Analysis Mechanism

Recognizing that AI performance cannot be fully accurate, we implemented a user feedback system. After each response, users can indicate if it was helpful. Negative feedback prompts a brief explanation (e.g., "Incomplete," "Not relevant"). Satisfactory responses along with their query, with user consent, are stored as a QUESTION object in the database as a semantic embedding vector. Embeddings allow us to compare meaning rather than just text, so similar questions with different wording will have close vectors. Using cosine similarity, new queries are matched against past ones. If similarity exceeds a threshold value (e.g., greater than 0.90), the system retrieves past answers that are positively rated, anonymized and stripped of any user-identifiable information, then displays them alongside a freshly generated AI response. All questions and responses are stored whether they are voted or not, using a unified schema. This supports analysis of common issues, content improvement, and better future responses without needing to retrain the model. The feedback loop enables a self-improving system driven by real user input.

IV. DISCUSSION

The HAAN system demonstrated that RAG has the potential to assist users in complex healthcare tasks. By using embedding-based semantic search over a curated medical knowledge base, the chatbot consistently provided contextually relevant and factually grounded answers, and showed flexibility in handling different phrasings of similar questions in our preliminary tests.

HAAN combines information retrieval and generative AI to reduce hallucination by grounding answers in real sources, a critical factor in healthcare applications. The use of vector-based embeddings and cosine similarity effectively bridged the gap between user queries and source content, corroborating existing literature on the role of vector databases in NLP.

By offering accurate and understandable health information in a conversational format, HAAN has the potential to make healthcare more accessible. Its adaptable nature also makes it useful for institutions managing communication during public health events. The built-in feedback system promotes trust and ensures continuous improvement based on real user input.

Although the system performed well in early scenarios, certain limitations remain. Some legal and bureaucratic topics are highly dynamic and require constant updates, which place a burden on maintaining the knowledge base. Furthermore, while the feedback mechanism is powerful, it relies on users being proactive in rating responses, which may not always occur in reality. In addition, users might shy away from giving consent to have their questions and responses utilized by the system. To address these challenges integration with government databases or APIs in real time, enhancing anonymization and scrubbing techniques could prove useful. Other improvements could include multilingual support and adaptive summarization for complex documents. The system should also be tested more extensively in cases involving rare legal situations or scenarios with incomplete or ambiguous user input to see its performance in diverse real-world contexts.

Despite these limitations, HAAN lays a meaningful foundation for accessible, AI-driven health-care support, and with future improvements, it has the potential to become an even more reliable and inclusive tool for those who need it the most.

V. CONCLUSION

In this project, we developed HAAN, a web-based chatbot designed to help users navigate the healthcare system by providing accurate, document-backed answers using a combination of LLMs and RAG. Built on a structured knowledge base from trusted sources and featuring a user-friendly interface with semantic search and feedback capabilities, HAAN demonstrated the viability of delivering relevant, grounded responses. While the system as it is currently implemented has several aforementioned limitations including language support and occasional gaps in completeness, future improvements and further testing can address those. Future improvements should focus on expanding content coverage, adding multilingual and accessibility features, and strengthening data protection.

Scalability and security remains another critical area of improvement. Overall, HAAN highlights how AI can improve healthcare access when designed with reliability, usability, and real-world needs in mind.

REFERENCES

- [1] T.C. Sağlık Bakanlığı, “Merkezi Hekim Randevu Sistemi (MHRS),” <https://www.mhrs.gov.tr>, (accessed on July 25, 2025).
- [2] T.C. Sağlık Bakanlığı, “E-Nabız Kişisel Sağlık Sistemi,” <https://enabiz.gov.tr>, (accessed on July 25, 2025).
- [3] Sağlık Bakanlığı İletişim Merkezi (SABİM), <https://sabim.gov.tr>, (accessed on August 5, 2025).
- [4] United Nations Sustainable Development Goals, <https://sdgs.un.org/goals>, (accessed on May 10, 2025).
- [5] L. Laranjo, A.G. Dunn, H.L. Tong, A.B. Kocaballi, J. Chen, R. Bashir, D. Surian, B. Gallego, F. Magrabi, and A.Y.S. Lau, “Conversational agents in healthcare: a systematic review,” *Journal of the American Medical Informatics Association*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018, doi: 10.1093/jamia/ocy072.
- [6] İ. Köse, J. Rayner, Ş. Birinci, M.M. Ulgu, İ. Yılmaz, and Ş. Güner, HIMSS Analytics Team, and MoH Team, “Adoption rates of electronic health records in Turkish hospitals and the relation with hospital sizes,” *BMC Health Services Research*, vol. 20, article no. 967, 2020, doi:10.1186/s12913-020-05767-5.
- [7] S.M. Jungmann, T. Klan, S. Kuhn, and F. Jungmann, “Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: Comparative case study with lay and expert users,” *JMIR Formative Research*, vol. 4, no. 4, e13863, Oct. 2019.
- [8] S. Razzaki, A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, M. Taliercio, M. Butt, A. Majeed, A. DoRosario, M. Mahoney, and S. Johri, “A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis,” *arXiv*, Jun. 2018, <https://arxiv.org/abs/1806.10698>, (accessed on May 10, 2025).
- [9] L.M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, “Retrieval augmented generation for large language models in healthcare: A systematic review,” *Nature Medicine*, 2025, Article ID 12157099.
- [10] World Health Organization, “Ethics and governance of artificial intelligence for health,” WHO, Geneva, 2021, <https://www.who.int/publications/i/item/9789240029200>, (accessed on August 5, 2025).
- [11] S. Amershi, M. Cakmak, W.B. Knox, and T. Kulesza, “Power to the people: The role of humans in interactive machine learning,” *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014. doi:10.1609/aimag.v35i4.2513.
- [12] “Cosine similarity,” ScienceDirect Topics, Elsevier, <https://www.sciencedirect.com/topics/computer-science/cosine-similarity>, (accessed on May 10, 2025).
- [13] Meta Platforms, Inc., “React – A JavaScript library for building user interfaces,” <https://reactjs.org>, (accessed on August 5, 2025).
- [14] MongoDB Inc., “MongoDB: The developer data platform,” <https://www.mongodb.com>, (accessed on August 5, 2025).
- [15] Ollama Inc., “Ollama: Run LLMs locally,” <https://ollama.com>, (accessed on August 5, 2025).
- [16] Chroma, “Chroma: The AI-native open-source vector database,” <https://www.trychroma.com>, (accessed on August 5, 2025).
- [17] LangChain, “LangChain: Building applications with LLMs through composability,” <https://www.langchain.com>, (accessed on August 5, 2025).