Domain-specific Embeddings for Question-Answering Systems: FAQs for Health Coaching

Andreas Martin[1], **Charuta Pande**[1], **Sandro Schwander**[1], **Ademola J. Ajuwon**[2], **Christoph Pimmer**[3]
[1]FHNW University of Applied Sciences and Arts Northwestern Switzerland, Intelligent Information Systems Research Group, Riggenbachstrasse 16, 4600, Olten, Switzerland
[2]Department of Health Promotion and Education, College of Medicine, University of Ibadan, Nigeria
[3]Swiss Tropical and Public Health Institute, Education and Training Department, Kreuzstrasse 2, 4123 Allschwil, Switzerland
(andreas.martin|charuta.pande|sandro.schwander)@fhnw.ch, ajajuwon@gmail.com, christoph.pimmer@swisstph.ch

Abstract

FAQs are widely used to respond to users' knowledge needs within knowledge domains. While LLM might be a promising way to address user questions, they are still prone to hallucinations, i.e., inaccurate or wrong responses, which, can, inter alia, lead to massive problems, including, but not limited to, ethical issues. As a part of the healthcare coach chatbot for young Nigerian HIV clients, the need to meet their information needs through FAQs is one of the main coaching requirements. In this paper, we explore if domain knowledge in HIV FAQs can be represented as text embeddings to retrieve similar questions matching user queries, thus improving the understanding of the chatbot and the satisfaction of the users. Specifically, we describe our approach to developing an FAQ chatbot for the domain of HIV. We used a predefined FAQ question-answer knowledge base in English and Pidgin co-created by HIV clients and experts from Nigeria and Switzerland. The results of the post-engagement survey show that the chatbot mostly understood the user's questions and could identify relevant matching questions and retrieve an appropriate response.

Introduction

Frequently Asked Questions, or FAQ, a tool to answer domain-specific queries with precise, accurate, and complete answers, is widely utilized in many domains. Conversational agents are a natural fit for FAQ implementation, saving users the hassle of searching FAQs on websites and the need for human support through emails or phone calls.

Humans can understand written text-based queries despite varied formulations and languages, and up to an extent, even when there are spelling and grammatical errors. Humans implicitly connect common sense and domain knowledge in various situations to resolve anomalies in text comprehension. To get a similar understanding in machines, a huge amount of training data is required. A standard approach to training conversational agents is to define "intents" and provide numerous examples of "user utterances" for those intents (de Lacerda and Aguiar 2019; Barus and Surijati 2022). This approach, although very effective for domain-specific conversational agents, is data-, time-, and resource-intensive.

In our research project to develop a healthcare coach chatbot to support young people living with HIV in Nigeria, the integration of FAQ on HIV is a core coaching functionality. Hence, it is crucial to interpret most queries and provide accurate answers. The FAQ should support English and Pidgin English. Although most users are conversant in English, it is not their first language. The challenge is to understand the user's intent behind a query, as it may contain incomplete questions, typos, specific local or regional terms, or out-of-scope questions.

Recently, studies show that LLM-based FAQ are quick to set up using domain-specific knowledge bases and are known to show good accuracy in responses (Huang and Chang 2022). However, LLMs are still prone to hallucinations (Rawte, Sheth, and Das 2023), i.e., inaccurate responses, which lead, among others, to ethical issues, which are particularly critical in the domain of global and public health. Additionally, due to data privacy constraints in our research domain, we have restrictions on using LLM-based techniques for HIV FAQ.

Taking the above-mentioned constraints into account, we explore if similarity-based retrieval from a pre-defined knowledge base of HIV FAQs can help in understanding user queries to a greater extent and reduce fallback scenarios, and what the role of LLMs could be in this process. We take the position that domain knowledge in the HIV FAQ can be represented as text embeddings to retrieve similar questions matching user queries, thus improving the understanding of the FAQ chatbot and the satisfaction of the users.

Related Work

Traditional approaches to developing FAQ chatbot involve using chatbot development language AIML (Ranoliya, Raghuwanshi, and Singh 2017), chatbot development frameworks and platforms like Rasa (de Lacerda and Aguiar 2019) and Dialogflow (Barus and Surijati 2022), and other techniques like rule and pattern matching (Sethi 2020). However, since the launch of ChatGPT, development of FAQ chatbots or Question-Answering systems using LLMs has been on the rise.

For developing FAQ chatbots using LLMs, different approaches, such as prompt engineering, fine-tuning, and retrieval augmentation, have been explored (Huang and Chang 2022; Zhu et al. 2023). Researchers have also experimented

with the use of LLM-only approaches. Seth et al. (2023) evaluated ChatGPT4's potential to answer queries related to a medical topic (breast augmentation) without applying any of the LLM-related approaches mentioned. Their findings show that the responses were coherent and without using medical jargon, but they were superficial, sometimes inaccurate, and inconsistent. Nordgren and E Svensson (2023) have experimented with different starting prompts for psychotherapy and found that an extensive prompt may not result in a better performance. However, providing domain knowledge as zero- or few-shot learning through prompt engineering has shown improved results for medical question-answering (Wang et al. 2023).

A common technique used in question-answering using LLMs is "Retrieval Augmented Generation (RAG)" which uses data from input documents or text as an additional context to generate responses (Lewis et al. 2020). RAG has been used in works like the FAQ chatbot to answer university-related questions (Cherumanal et al. 2024), open-domain question-answering (Siriwardhana et al. 2023) and question-answering in the blockchain domain (Mansurova, Nugumanova, and Makhambetova 2023). Ren et al. (2023) observe that the performance of question-answering systems using RAG depends on the quality of knowledge contained in the supporting documents.

Another approach to FAQ chatbot using LLMs includes semantic search on question embeddings (Pandya and Holia 2023; Huang et al. 2023; Medeiros et al. 2023). Arz von Straussenburg (2023) propose a hybrid model to combine traditional and LLM-based approaches.

HIV FAQ Development

In this section, we describe our approach to developing an HIV FAQ chatbot.

HIV FAQ Knowledge Base

The knowledge base for the HIV FAQ consisted of curated questions on HIV and subtopics around it. The process involved a comprehensive co-creation approach. The first subset of the questions was developed by a group of 'champions', 23 representatives of the user group, who came up with questions about HIV that they would like to have answered. Another subset was then added by experienced HIV counselors, who added questions typically asked by their clients. We also used ChatGPT to generate additional questions. Answers were generated by the project's HIV experts. For the final set of questions, ChatGPT was also used to generate not just questions but also answers, which were then revised by the Nigerian health experts and adapted to the specific context. This process resulted in the development of more than 400 different questions with corresponding answers on 15 topics on HIV, such as "HIV Basics", "Medication", "Relationships" etc., including question variants in both English and Pidgin English. This knowledge engineering exercise was considered necessary for our domain of supporting HIV in Nigeria for several reasons:

1. We cannot rely entirely on LLM-based approaches to generate FAQ responses due to their limitations on providing factual and accurate information, as explained before
2. We cannot leverage LLMs directly due to privacy and ethical considerations related to our application domain
3. LLMs are limited in their ability to work with questions in Nigerian Pidgin English
4. The questions and responses need to be tailored to the specific social, economic, and cultural situation of our target group in Nigeria. The questions curated by our experts consolidate and augment the topics available in FAQs on different websites in Nigeria[1] [2]

Experiments on Embeddings Models

In the dynamic and evolving field of natural language processing, the selection of an appropriate embedding model is critical for the successful implementation of a chatbot's frequently asked questions (FAQ) solution. Our approach necessitated the adoption of open-source embedding models that we could host on our own servers, catering specifically to stringent privacy and security requirements. Given the specific requirements of our domain and the tasks involved, we opted for the Hugging Face sentence transformers library. The library's models were evaluated based on several key metrics: the number of downloads, ranking on the SBERT index [3], position on the MTEB Leaderboard [4] with a focus on overall and retrieval tasks (Muennighoff et al. 2022), model size, and speed.

Three models were shortlisted for in-depth analysis (as of August 2023):

1. all-mpnet-base-v2: This model stood out for being at the top in model downloads and on the SBERT index, also ranking 20th on the MTEB general leaderboard and 22nd for the retrieval task.
2. all-MiniLM-L12-v2: It matched the top model in downloads and SBERT ranking, with a slightly lower position on the MTEB leaderboard, ranking 24th in general and 23rd in retrieval.
3. multi-qa-mpnet-base-dot-v1: Tuned specifically for semantic search, it was a top performer on SBERT and in model downloads, though it was not ranked on the MTEB

Our dataset consisted of an FAQ question set derived from queries generated by our users. The initial questions were enhanced with variations incorporating domain-specific synonyms, e.g., ("HIV"-"H", "ARV drug" - "sweet", "ARV medication") and then rephrased with the assistance of GPT-3.5-turbo with a "temperature" 0 as we wanted to keep the randomness in the rephrased questions to a minimum. This resulted in a comprehensive dataset of 1552 question-answer pairs.

[1] https://ihvnigeria.org/faq/

[2] https://www.aun.edu.ng/index.php/campus-life/health-center/health-tips/facts-on-hiv-aids-and-tips-on-prevention

[3] https://www.sbert.net/docs/pretrained_models.html\#sentence-embedding-models/

[4] https://huggingface.co/spaces/mteb/leaderboard

Model	80/20 split	English	Pidgin	Average	Variance
all-mpnet-base-v2	91.0%	85.70%	91.80%	89.5%	0.00073
all-MiniLM-L12-v2	85.2%	83.70%	81.60%	83.5%	0.00022
multi-qa-mpnet-base-dot-v1	95.10%	81.60%	93.90%	90.2%	0.00372

Table 1: Performance comparison of three main models as % of correctly retrieved answers with k=1

To assess the performance of these models, we employed three distinct approaches:

1. An 80/20 train/test split, a standard method for validating machine learning models.
2. Evaluation against 50 new questions generated by GPT-3.5-turbo in English, designed to test the models' generalization capabilities.
3. Analysis based on the translation of the same 50 questions into Pidgin English, aligning with the local vernacular of our project to test linguistic adaptability.

Performance outcomes as displayed in Table 1 were binary, categorized as either correct (successfully retrieving the correct answer with k=1) or incorrect. This stringent evaluation criterion aimed to simulate realistic conditions under which the chatbot would operate and ensure the reliability of the selected model in a live environment. We used the FAISS vector store provided through LangChain [5].

Experiment Results

The evaluation of the embedding models for our health coaching chatbot has produced enlightening insights into the robustness and reliability of these systems in practical applications. The performance analysis of the three models revealed stable and relatively comparable outcomes (see Table 1), the all-MiniLM-L12-v2 lagging slightly behind the other two. We opted for the all-mpnet-base-v2 taking into account the variance as a measure of robustness across the three approaches.

Our analysis demonstrated an impressive average correct retrieval rate of 89.5%. Despite this success, the real-world application of this technology necessitates a system capable of handling the unpredictability of user queries. To enhance the adaptability of our chatbot to incorrect retrievals and inquiries outside the scope of our dataset, we implemented an extended retrieval strategy. We employed a retrieval with k=10, returning a sequence of confirmation questions (e.g. "Did I understand that correctly, that you wanted to know the purpose and usage of ARV drugs?") ranked from the highest to the lowest relevance, while discarding duplicates. This strategy allowed users to confirm the chatbot's understanding of their initial query.

Furthermore, we introduced a scoring threshold, rejecting documents with a score lower than 0.2. This measure was taken to effectively exclude responses to questions beyond the domain of our dataset (e.g. "how are you?"), as well as other unforeseen or out-of-context user inputs. The observation was made that lower confidence scores could be reliably disregarded due to their high predictive certainty of irrelevance. This refinement was facilitated by the utilization of the similarity search function with scoring capability provided by the FAISS vector store. Upon registering a document score that fell below the established threshold, an automated fallback response was activated and delivered to the user. This response communicated that the query was not understood, and it prompted the user either to reformulate their question or to ensure that their subsequent queries pertained to the specific domain of HIV-related subjects. This function has proven critical in ensuring that the chatbot's responses remain within the relevant context and preserve the quality of user interaction. A comprehensive representation of the final implementation of our FAQ system is depicted in Figure 1.

FAQ Evaluation

Following the initial selection and adaptation of the embedding model, we proceeded to validate the effectiveness of our FAQ solution through practical application. This validation process involved the "champions", who interacted with the chatbot, thereby generating valuable data through conversation transcripts and post-interaction evaluations.

Each user was tasked with inputting three distinct questions to the chatbot. Post-engagement, they were required to provide feedback on two critical aspects: first, whether the chatbot understood their question (Yes/No), and second, the degree of satisfaction with the response (Yes/Somehow/No). In instances where satisfaction was not achieved, users were prompted to articulate the reasons and suggest potential enhancements. This step was crucial in identifying specific areas for refinement.

We conducted two evaluations with the champions with different scopes of questions, adding up to six questions each, or 126 in total. The overall understanding and satisfaction rates are displayed in Table 2.

A deeper analysis revealed that nearly half of the negative responses related to 'understanding' came from only five users, suggesting that their questions may have been out of scope or were ambiguously phrased, as the chatbot consistently failed to respond correctly to these individuals.

Incorporation of Domain Knowledge in HIV FAQ

The knowledge in the FAQs is contained in the domain-specific question-answers created by our experts. This domain knowledge is represented as text embeddings and used to find the top most similar questions that match the user's query. As explained before, there is a considerable possibility that users misspell or type in partial queries. Also, it is quite usual for users to enter a generalized query that spans

[5]https://python.langchain.com/docs/integrations/vectorstores/faiss

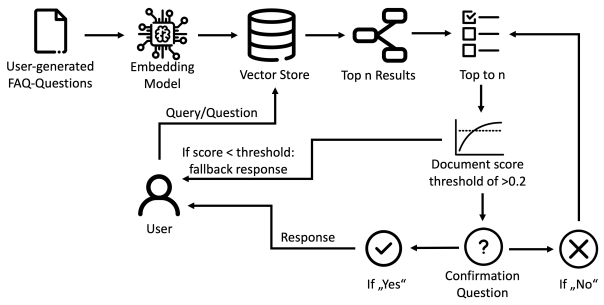

Figure 1: Overview of final implementation of FAQ system

multiple precise questions in the FAQ base. Creating embeddings out of queries and performing a similarity search on the FAQ base improves the chances of finding one or more questions that are likely to fit a user's intent. The domain knowledge in the FAQ response remains unchanged as we do not paraphrase the responses but only retrieve them by identifying appropriate questions. This helps in keeping the responses accurate as long as the right question fitting the user's intent is identified. However, we assume paraphrasing responses and creating a conversation around them can improve the naturalness and user experience - we take this up as future work.

Ethical Considerations

As mentioned earlier, the main reason for not using LLMs directly for our FAQ approach was ethical and privacy concerns. The questions on HIV are personal and sensitive. Techniques like RAG allow to easily include domain knowledge and generate seemingly accurate responses. However, there are two risks in using LLMs for answering HIV FAQ: 1. if the models are not hosted on own infrastructure, the sensitive information in the user queries is at security risk; and 2. the responses generated by the models may seem accurate, but due to the sensitive nature of the HIV domain, there is an utmost necessity for these responses to be verified by domain experts for accuracy. We address these risks by curating our own set of HIV FAQ questions and representing domain knowledge as embeddings. The accuracy and domain knowledge of the responses remain unaffected, as only an expert-defined answer is retrieved from the existing repository of FAQs.

Understanding		
Yes	95	75%
No	31	25%
Satisfaction		
Yes	86	68%
Somehow	16	13%
No	24	19%

Table 2: Results of Post-Engagement Survey

Conclusion and Future Work

In this work, we describe our approach to developing an FAQ chatbot for the HIV domain. We used the FAQ question-answer knowledge base in English and Pidgin developed by HIV experts in Nigeria. We carried out experiments to identify an appropriate embeddings model for our approach. We try to find the most similar questions from the FAQ base that resemble the user's query. Our preliminary user survey results show that the chatbot mostly understood the user's questions and could identify relevant matching questions and retrieve an appropriate response. To improve the validity of results, we plan evaluations with a larger group of participants. Additionally, as a future work, we plan to explore how responses can be presented in a conversational style including some paraphrasing.

Acknowledgments

This work has been funded by the Swiss National Science Foundation (SNSF) under grant IZSTZ0_202602 within the Swiss Programme for International Research by Scientific Investigation Teams (SPIRIT).

References

Arz von Straussenburg, A. F. 2023. Towards Hybrid Architectures: Integrating Large Language Models in Informative Chatbots.

Barus, S. P.; and Surijati, E. 2022. Chatbot with dialogflow for FAQ services in Matana university library. *International Journal of Informatics and Computation*, 3(2): 62–70.

Cherumanal, S. P.; Tian, L.; Abushaqra, F. M.; de Paula, A. F. M.; Ji, K.; Hettiachchi, D.; Trippas, J. R.; Ali, H.; Scholer, F.; and Spina, D. 2024. Walert: Putting Conversational Search Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot. *arXiv preprint arXiv:2401.07216*.

de Lacerda, A. R.; and Aguiar, C. S. 2019. FLOSS FAQ chatbot project reuse: how to allow nonexperts to develop a chatbot. In *Proceedings of the 15th International Symposium on Open Collaboration*, 1–8.

Huang, D.; Wei, Z.; Yue, A.; Zhao, X.; Chen, Z.; Li, R.; Jiang, K.; Chang, B.; Zhang, Q.; Zhang, S.; et al. 2023. DSQA-LLM: Domain-Specific Intelligent Question Answering Based on Large Language Model. In *International Conference on AI-generated Content*, 170–180. Springer.

Huang, J.; and Chang, K. C.-C. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

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

Mansurova, A.; Nugumanova, A.; and Makhambetova, Z. 2023. Development of a question answering chatbot for blockchain domain. *Scientific Journal of Astana IT University*, 27–40.

Medeiros, T.; Medeiros, M.; Azevedo, M.; Silva, M.; Silva, I.; and Costa, D. G. 2023. Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manuals. *Vehicles*, 5(4): 1384–1399.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2022. MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316*.

Nordgren, I.; and E Svensson, G. 2023. Prompt engineering and its usability to improve modern psychology chatbots.

Pandya, K.; and Holia, M. 2023. Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations. *arXiv preprint arXiv:2310.05421*.

Ranoliya, B. R.; Raghuwanshi, N.; and Singh, S. 2017. Chatbot for university related FAQs. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1525–1530. IEEE.

Rawte, V.; Sheth, A.; and Das, A. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Ren, R.; Wang, Y.; Qu, Y.; Zhao, W. X.; Liu, J.; Tian, H.; Wu, H.; Wen, J.-R.; and Wang, H. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Seth, I.; Cox, A.; Xie, Y.; Bulloch, G.; Hunter-Smith, D. J.; Rozen, W. M.; and Ross, R. J. 2023. Evaluating chatbot efficacy for answering frequently asked questions in plastic surgery: a ChatGPT case study focused on breast augmentation. *Aesthetic Surgery Journal*, 43(10): 1126–1135.

Sethi, F. 2020. FAQ (Frequently Asked Questions) ChatBot for Conversation. *International Journal of Computer Sciences and Engineering*, 8(10).

Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; and Nanayakkara, S. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11: 1–17.

Wang, J.; Shi, E.; Yu, S.; Wu, Z.; Ma, C.; Dai, H.; Yang, Q.; Kang, Y.; Wu, J.; Hu, H.; et al. 2023. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.

Zhu, Y.; Yuan, H.; Wang, S.; Liu, J.; Liu, W.; Deng, C.; Dou, Z.; and Wen, J.-R. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.