IaraMed: A Women's Healthcare Chatbot for Portuguese Speakers

Fernanda B. Färber[1,2], Julia S. Dollis[1,2], Pedro S. F. B. Ribeiro[1,2], Iago A. Brito[1,2], Rafael T. Sousa[1,3], Arlindo R. Galvão Filho[1,2]

[1]Advanced Knowledge Center in Immersive Technologies (AKCIT)
Goiânia, Brazil

[2]Institute of Informatics (INF) – Federal University of Goias (UFG)
Goiânia, Brazil

[3]Federal University of Mato Grosso (UFMT)
Barra do Garças, Brazil

{fernandabufon,juliadollis,schindler,iagoalves}@discente.ufg.br

rafaelsousa@ufmt.br, arlindogalvao@ufg.br

Abstract. *The advent of large language models has significantly advanced natural language processing, revolutionizing numerous applications within the healthcare domain. Despite these advances, most existing research remains predominantly centered around English, resulting in notable disparities in medical AI accessibility for non-English speaking communities. To bridge this gap, we introduce a specialized Portuguese-language chatbot tailored explicitly to women's healthcare needs, addressing critical shortages in linguistic resources and culturally relevant data. Leveraging retrieval-augmented generation, our chatbot integrates accurate, evidence-based information directly into generated responses. Evaluations demonstrate that our approach markedly enhances reliability, highlighting the potential for tailored AI applications to significantly improve healthcare accessibility and outcomes for Portuguese-speaking women.*

1. Introduction

Recent advancements in artificial intelligence (AI) have been driven by the emergence of large language models (LLMs). Trained on vast amounts of text, these models have demonstrated exceptional capabilities in understanding and generating natural language. As a result, LLMs have been successfully applied to a wide range of tasks, including translation (Elshin et al. 2024), logical reasoning (Ye et al. 2025), and question answering (Robinson et al. 2022). In the healthcare domain, their potential is notable in a large set of areas such as patient support and diagnostic assistance (Tu et al. 2024).

However, despite these significant advances, research and development in AI and natural language processing (NLP) for healthcare remain largely focused on English (Jin et al. 2021). This English predominance has led to a scarcity of specialized resources, models, and datasets for low and mid-resource languages. Consequently, non-English speaking communities are often deprived of the state-of-the-art diagnostic tools, reliable medical information, and personalized health support that modern AI systems can offer (Chi et al. 2023).

Women's health encompasses a broad spectrum of unique needs, including pregnancy, reproductive health, and gynecological issues (Montenegro et al. 2022), many of which have been overlooked by existing AI solutions. To bridge this gap, our work focuses on developing a Portuguese LLM-based chatbot specifically designed for women's health. Due to the limited availability of Portuguese medical data, we construct a specialized corpus by scraping content from the reputable source Doctoralia[1], with an emphasis on topics relevant to women's healthcare. This corpus directly addresses the shortage of Portuguese-language medical data in this domain. By targeting these critical areas, our proposed system aims to provide accurate, context-aware medical information and support, ultimately improving health outcomes and well-being for Portuguese-speaking women.

To enhance our chatbot's capabilities beyond basic text generation and memorization, we integrate retrieval-augmented generation (RAG). This approach effectively combines a large language model with a knowledge retrieval mechanism, significantly reducing the risk of inaccurate responses by providing contextually relevant and citation-supported answers (Xiong et al. 2024). To build a robust and comprehensive knowledge base for our retriever system, we develop a database sourced from Portuguese-language medical literature and expert-reviewed web articles. This ensures access to high-quality, domain-specific knowledge, enhancing the chatbot's ability to deliver precise and reliable responses while addressing the broader challenge of data scarcity in AI-driven healthcare applications for Portuguese speakers.

The remainder of this paper is structured as follows: Section 2 surveys the relevant literature, after which Section 3 introduces our curated dataset. Section 4 then details the proposed methodology, and Section 5 presents the experimental setup and results. Finally, Section 6 discusses the study's limitations and outlines directions for future research.

2. Related Work

The recent uprising of LLMs has significantly advanced chatbot performance by producing responses that are both coherent and contextually nuanced. In contrast to earlier rule-based systems, which were constrained by static response templates and limited adaptability, modern LLMs deliver superior conversational fluency and heightened contextual sensitivity (Gemini Team et al. 2023). Consequently, this evolution has broadened the applicability of chatbot technology across diverse domains, including education (Chiu et al. 2024), entertainment (García-Méndez et al. 2021), and healthcare (Xu et al. 2021).

Alongside these advancements, retrieval-augmented generation has emerged as a pivotal enhancement for LLMs in addressing knowledge-intensive tasks, particularly within the healthcare domain. By dynamically retrieving relevant, domain-specific information from external sources, RAG not only mitigates hallucinations and reduces the risks of misinformation but also ensures that responses are both contextually precise and up-to-date (Lewis et al. 2020). This integration has been transformative, as evidenced by its successful applications across various medical specialties such as nephrology (Miao et al. 2024), infectious diseases (Kirubakaran et al. 2024) and ophthalmology (Passinato et al. 2024). Collectively, these advancements highlight RAG's potential to

[1]https://www.doctoralia.com.br/

significantly enhance the specificity and reliability of chatbot responses, thereby making them exceptionally well-suited for the complex demands of medical applications.

However, in the context of women's healthcare, few studies have explored this approach, revealing a significant gap in the literature. While studies relying on rule-based methods demonstrate the feasibility of chatbots in supporting pregnant women (Montenegro et al. 2022; Puspitasari et al. 2022), their dependence on Hardcoded logic limits adaptability to complex, context-dependent inquiries. Similarly, traditional retrieval and search-based systems, such as those examined in (Tawfik et al. 2023), have shown promise in enhancing patient engagement. Nevertheless, these approaches often lack the dynamic contextual reasoning and nuanced language comprehension that RAG-enhanced LLMs provide, which are essential for delivering more accurate, personalized, and contextually relevant responses in women's health applications.

For Portuguese language, the scarcity of research is even more pronounced. To the best of our knowledge, only (Montenegro et al. 2022; dos Santos Junior et al. 2021) has explored this topic, with a limited scope due to its reliance on rule-based interactions. Findings from (Montenegro et al. 2022) reveal that users found the women's healthcare chatbot to be both educational and a valuable complement to medical consultations. These insights underscore the pressing need for further studies that address the linguistic and cultural particularities of the Brazilian context, as well as the distinct healthcare needs of Brazilian women.

Given these gaps, this study aims to advance the field by developing a RAG-enhanced LLM specifically designed for women's healthcare in Portuguese. By incorporating retrieval-based techniques, our approach enhances the accuracy and reliability of AI-driven medical question-answering while addressing the linguistic and contextual challenges of non-English healthcare applications. This contribution expands the scope of AI-driven healthcare solutions while promoting greater inclusivity and accessibility for Portuguese-speaking communities.

3. Medical Data

The performance of machine learning models is critically dependent on both the quantity and quality of the data on which they are trained (Gong et al. 2023). In the healthcare domain, particularly for text-based AI models, the scarcity of high-quality Portuguese medical data poses a formidable challenge. Although datasets such as BRATECA (Consoli et al. 2022) and SemClimBR (Oliveira et al. 2022) are available to researchers, they lack the comprehensive detail needed to capture the nuances of patient-medical interaction. Furthermore, the scarcity of dedicated datasets on women's healthcare poses a major obstacle to improve the truthfulness on the LLMs answer, relying solely on general pre-training corpora with no possibilities to domain-specific fine-tuning or using in-context learning techniques.

3.1. Question Answering Dataset

Through an extensive web scraping initiative on Doctoralia, a trusted platform in Brazil where users submit queries and receive answers exclusively from verified healthcare professionals, we developed a dedicated Brazilian women's healthcare question-answering dataset. This strategy enabled us to obtain a robust collection of question-answer pairs that

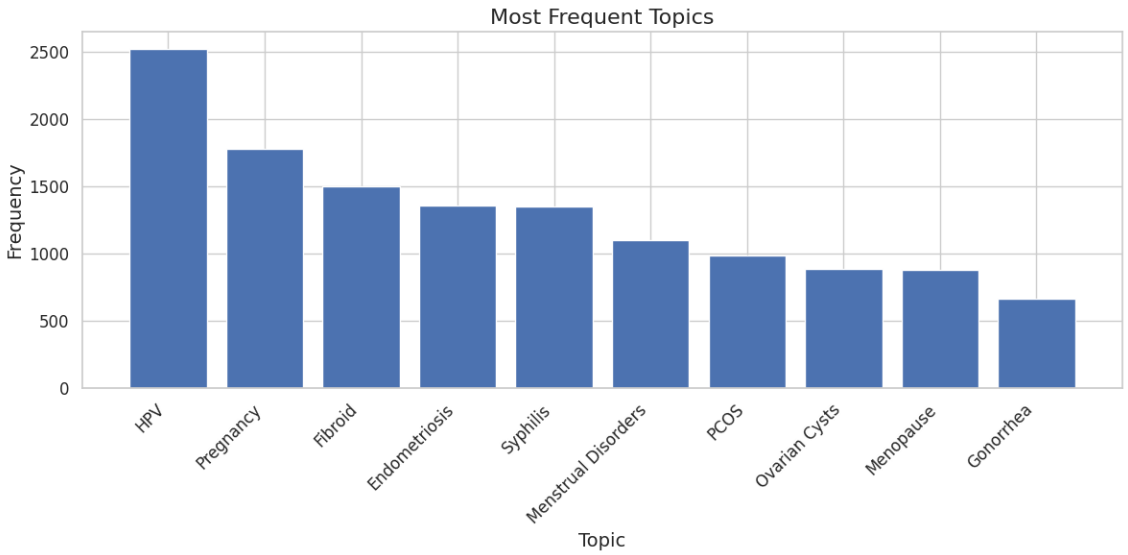

Figure 1. Most frequency addressed topics in users questions

encapsulate the prevalent healthcare concerns among Brazilian women. Consequently, our curated dataset offers a richer, more focused resource that addresses the unique challenges of Portuguese inquires, paving the way for the development of more precise and context-aware NLP solutions in women's healthcare.

In our initial data scraping phase, we gathered over 2.3 million publicly available samples across diverse healthcare areas, each characterized by eight distinct features: specialty, subspecialty, user's medical inquiry, doctor's answer, name, area of specialization, rating and location. To enhance the dataset's relevance for our analysis, we selectively retained only the most critical elements: the complete text of the user's medical inquiry, the validated answer provided by a healthcare professional, and the professional's area of specialization. This targeted approach not only streamlined the dataset but also ensured that subsequent analyses focused on the most impactful and informative components.

To construct a specialized corpus of question-answer pairs focused on women's healthcare, we employed a rigorous filtering process based on healthcare professionals' areas of specialization. Specifically, 19,538 samples were extracted where the specialization was identified as either gynecology or obstetrics, thereby ensuring that our dataset remained strictly relevant to women's health. This targeted selection offers a robust foundation for the development and evaluation of domain-specific NLP solutions. Figure 1 illustrates the most frequently addressed topics.

To refine our dataset and bolster its reliability, we implemented a comprehensive cleaning process that systematically discarded irrelevant, incomplete, and duplicate samples. This step was essential for eliminating noise that could otherwise undermine our model's performance. Additionally, we applied a data enhancement process by leveraging a LLM to correct grammar and formatting for all samples from our crawled dataset. This step was particularly important because, although the crawled data was accurate and reliable, it contained numerous grammatical and writing errors, as it was scraped from web platforms where users freely submit questions.

Finally, to enhance the system's ability to differentiate between relevant and irrelevant queries to women's healthcare, we incorporated 3,920 random samples from five other medical specializations (surgery, internal medicine, dentistry, pediatrics, psychology), achieving a total of 19,600 samples. These negative samples enabled the creation of a robust dataset for fine-tuning a specialization classifier able to detect if a given query is or is not related to the scope of women's healthcare. Further details are provided in Section 4.1.

3.2. Knowledge Base

LLMs frequently generate plausible yet factually incorrect responses, a phenomenon commonly referred to as hallucination (Zhang et al. 2023). Such inaccuracies arise because these models rely primarily on statistical patterns learned from massive textual datasets, rather than explicit verification against authoritative sources. In sensitive domains like healthcare, where precision and factual correctness are imperative, hallucinations pose significant risks to patient safety and clinical decision-making, potentially leading to misinformation and compromised patient trust.

To address the pressing need for domain-specific resources in Portuguese women's healthcare, we constructed a comprehensive knowledge base from reputable medical sources. This repository includes 80 specialist-reviewed articles from *Tua Saúde*[2], along with nearly 150 gynecology and obstetrics books, totaling 2.5 million words. By spanning an extensive range of medical conditions, symptoms, treatments, and medications, our knowledge base provides a robust, up-to-date foundation for reliable and contextually nuanced responses in women's healthcare.

By incorporating this curated knowledge base into a retrieval-augmented generation framework, our chatbot can effectively cross-verify and contextualize its responses, thereby substantially reducing the risk of hallucinations. This strategy not only ensures greater factual accuracy but also delivers contextually relevant and trustworthy medical guidance, enhancing both the reliability and practical applicability of the model's answers.

4. Methodology

Our system architecture begins by assessing the relevance of each user inquiry to women's healthcare. For queries deemed off-topic, the LLM is prompted not to generate a response but rather to clarify the chatbot's limited scope and recommend an appropriate specialist. When the inquiry is relevant, the system retrieves supplementary information from a meticulously curated knowledge base. This retrieved content is then integrated with the original query and processed by the LLM through a retrieval-augmented generation approach, ensuring responses that are both contextually rich and accurate. By seamlessly combining classification, retrieval, and generation modules, our framework delivers precise and personalized medical guidance in women's healthcare. Figure 2 illustrates our system architecture.

4.1. Specialization Classifier

To restrict the model's responses strictly to women's health and mitigate risks of misinformation or hallucinations from out-of-domain queries, we developed a specialization

[2] https://www.tuasaude.com

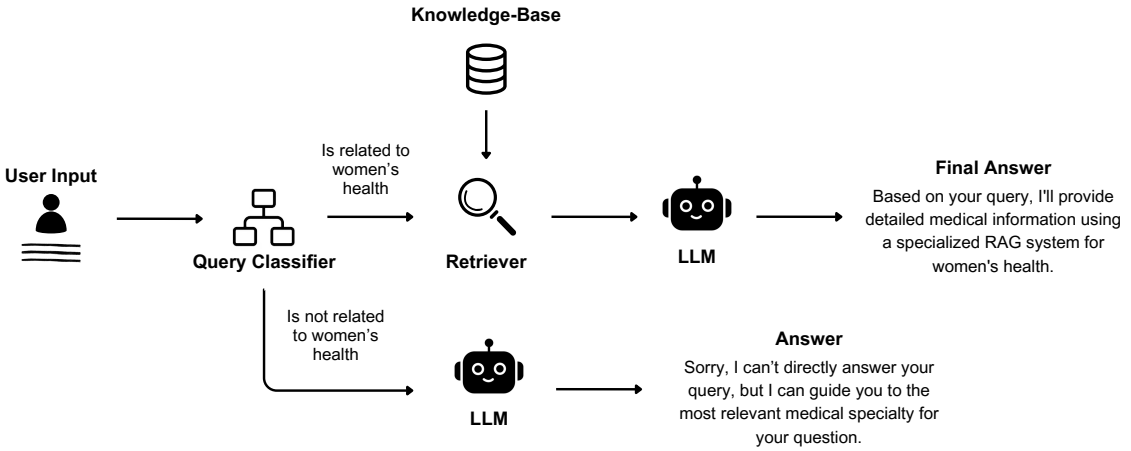

Figure 2. Proposed methodology for retrieval-augmented response generation.

classifier. This binary classifier differentiates inputs as either relevant or irrelevant to women's healthcare. Leveraging an encoder-only architecture for its ability to convert text into semantically rich spatial vectors, we integrated a multi-layer perceptron as the classification head to accurately perform this distinction.

4.2. Information Retrieval

Using the knowledge base described in Section 3.2, we employed a semantic search approach to retrieve relevant documents for each user query. Both queries and documents were encoded using E5-multilingual (Wang et al. 2022), a model optimized for multilingual retrieval tasks. For efficient search, we utilized Facebook AI Similarity Search (FAISS) (Douze et al. 2024) as the vector database, selecting the top five documents based on cosine similarity. This method ensures that the LLM generates responses grounded in comprehensive, accurate, and up-to-date information on women's healthcare.

4.3. Retrieval Augumented Generation

Our RAG combines information extracted from our curated knowledge base with the generative strengths of the large language model. In this phase, we utilize Gemini 2.0 Flash (Gemini Team et al. 2023) as generator, merging the user's original query with context refined through a document filtering process to generate the final response. This integrated strategy not only tailors outputs to the query but also ensures that the generated content is grounded in verified, external data.

By incorporating medically validated information during inference, RAG significantly enhances both the contextual relevance and factual accuracy of its outputs. This approach effectively mitigates the risk of hallucinations while ensuring the high reliability standards required for medical applications in women's healthcare.

4.4. Evaluation Methods

After retrieving relevant documents and consolidating context through RAG, our specialized Portuguese healthcare LLM produces a final response that is both accurate and contextually appropriate. To systematically evaluate these outputs, we employed an LLM-as-judge framework, a well-established methodology for automated quality assessment

in NLP literature comparable to human quality (Xiong et al. 2024). Specifically, in this work we utilize Gemini 2.0 Pro (Gemini Team et al. 2023) as evaluator, analyzing both the the quality of the RAG process and the quality of the final response.

The quality of retrieved knowledge was evaluated by assigning relevance scores ranging from 1 (irrelevant) to 3 (highly relevant), reflecting the degree of alignment between the retrieved content and user queries. Additionally, the quality of the generated answers was assessed by direct comparison with validated medical responses, allowing precise measurement of medical accuracy and appropriateness. This evaluation strategy ensures that the final chatbot outputs consistently adheres to professional medical standards, providing trustworthy guidance tailored to user needs.

5. Experiments and Results

This section describes the experimental setup, evaluation metrics, and results from testing both individual components and the complete system. For the specialiazation classifier, we made experiments with three encoder-only models: BERTimbau (Souza et al. 2020), XLM-RoBERTa (Conneau et al. 2020), and mBERT (Devlin et al. 2019). BERTimbau is a portuguese-specific encoder-only model, while mBERT and XLM-RoBERTa are trained on a diverse corpus spanning multiple languages. All tests were conducted using the the base versions of these models.

All models were trained using the same dataset and configurations. The dataset comprises 39,056 real-world patients questions, divided into 80% training and 10% for test and validation sets. It is balanced between related or unrelated to women's health as described in Section 3.1. Training and evaluation were performed with a batch size of 32. In the training, we employed AdamW as the optimizer, applying a linear decay of 5e-08 each step to the learning rate, which started at 5e-5. To evaluate classification performance, we computed standard binary classification metrics, including accuracy, F1-score, ROC-AUC, and Matthews correlation coefficient (MCC). A comparative analysis of the results for each model is present on Table 1.

Model	Accuracy	F1-Score	MCC	AUC-ROC
BERTimbau (Souza et al. 2020)	**0.946**	**0.946**	**0.893**	**0.984**
XLM-RoBERTa (Conneau et al. 2020)	0.943	0.943	0.886	0.983
mBERT (Devlin et al. 2019)	0.945	0.945	0.891	0.983

Table 1. Results of the encoder-only models after fine-tune.

As the results indicate, all models performed well across all evaluated metrics, with an F1-score exceeding 0.94 in all cases. This underscores the quality of our dataset, as all models achieved strong performance despite the complexity of the task. Among the tested models, BERTimbau-base outperformed XLM-RoBERTa and mBERT, achieving the highest scores across all metrics. This advantage is likely due to BERTimbau's further pretraining on a significantly larger volume of Portuguese data, whereas XLM-RoBERTa and mBERT adopt a multilingual approach. These results confirm that our Specialization Classifier is well-equipped to handle user queries effectively, ensuring that the system consistently provides relevant responses.

While our specialization classifier ensures domain relevance, the quality of responses also depends on the volume of retrieved context. To optimize retrieval, we split our corpus into 17,539 documents and varied the number of retrieved documents (k). Although higher k values supply richer context, they also raise computational costs and risk token truncation. Evaluations on 2,000 query-context pairs with LLM-as-Judge (see Figure 3) revealed that despite increasing k improves relevance, the effect diminishes beyond k = 5, suggesting that additional documents contribute minimal. Consequently, we selected k = 5 as the optimal balance between contextual completeness and computational efficiency.

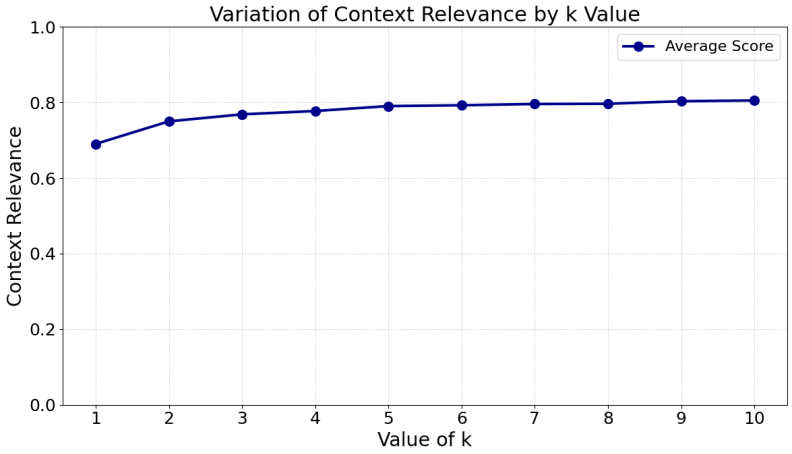

Figure 3. Comparison of Retrieved Document Count (value of k) and Context Relevance.

We employed Gemini 2.0 Flash (Gemini Team et al. 2023) to compare each RAG-enhanced LLM response against a medically validated gold standard using 2,000 random samples. Each output was labeled as "correct", "partially correct", "partially wrong", or "wrong", with partial correctness treated as "correct". As shown in Table 2, retrieving up to five document chunks steadily improves correctness, after which performance degrades. This trend aligns with (Liu et al. 2024), suggesting that excessive contextual information can overwhelm the model and undermine accuracy.

Chunks Retrieved (k)	Correctness (%)
1	93.00
3	94.25
5	**95.35**
7	94.85

Table 2. Results of final answer as the number of chunks varies

As result, Figure 4 presents a representative interaction showcasing our final proposed chatbot system. In this example, The user inquires about the possibility of pregnancy post-menopause, and the chatbot provides a medically grounded, response using specialized knowledge retrieved. This example underscores how our approach integrates domain-specific classification, retrieval-augmented generation, and vetted medical content to deliver contextually accurate and trustworthy information, illustrating both the

feasibility and promise of AI-driven medical assistance to women's healthcare needs in Portuguese.

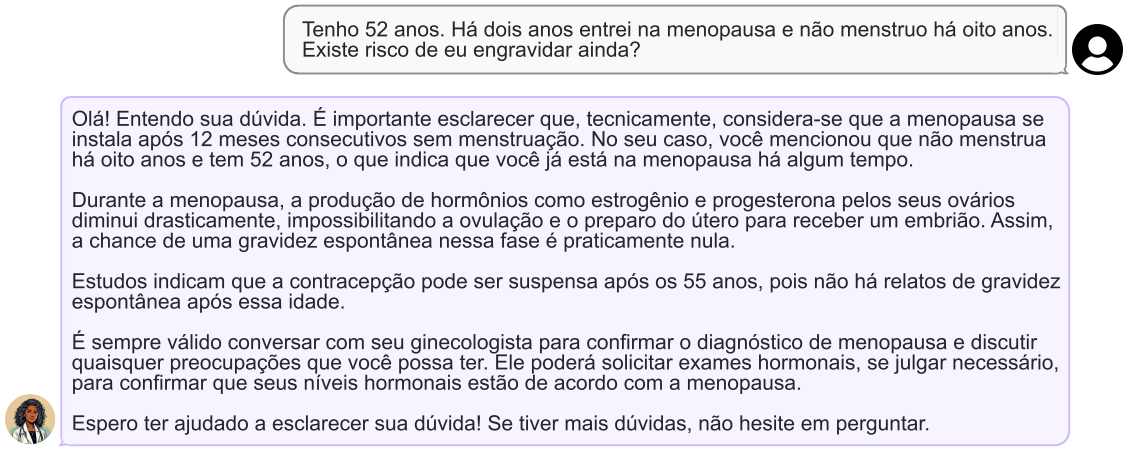

Figure 4. Interaction between user the proposed system answering question about menopause-related query.

6. Limitations and Future Work

Although our methodology demonstrates promising results, several limitations remain. First, our current dataset primarily comprises question-answer pairs, lacking the interactive dialogue structure necessary for modeling nuanced patient-provider conversations. Consequently, the chatbot's capacity for fluid and contextually dynamic interactions may be limited. Addressing this will require the future inclusion of conversational data that better captures natural dialogue flow, responsiveness, and the subtle complexities typical in medical interactions.

Moreover, despite the effectiveness of automated evaluation methods such as LLM-based assessment, these models still suffers from problems such as bias and hallucinations, making human judgment indispensable due to critical ethical, clinical, and interpretability considerations. Nevertheless, human evaluation methodologies pose significant practical challenges, including ethical considerations, participant recruitment, and considerable resource investments. Future research should thus aim to balance scalable automated assessments with rigorous human evaluations, ensuring both efficiency and reliability in assessing clinical alignment, accuracy, and safety.

Lastly, our current methodology does not explicitly incorporate empathy or personality modeling into the chatbot's responses, potentially limiting user engagement and satisfaction. Future research should explore methods for embedding empathetic response strategies and distinct personality traits into LLM-generated outputs. Integrating these elements could enhance user trust, improve emotional resonance, and foster more meaningful interactions, ultimately leading to greater acceptance and adoption among users.

7. Conclusion

The development of a specialized, Portuguese-language chatbot for women's healthcare represents a crucial step toward equitable AI-driven medical assistance, especially for

non-English speaking populations. Our work directly addresses the scarcity of high-quality, domain-specific datasets and models tailored to the linguistic and cultural nuances of Brazilian healthcare. By systematically curating and processing data from reputable medical platforms, we have created a comprehensive, richly annotated dataset dedicated exclusively to women's healthcare, thereby significantly advancing NLP research capabilities in Portuguese.

Furthermore, through the integration of state-of-the-art retrieval-augmented generation and domain-specific classification, our proposed framework ensures the delivery of precise, contextually informed, and personalized responses. This approach not only enhances the accuracy and reliability of medical guidance but also underscores the potential for specialized AI systems to positively influence women's health outcomes. Moving forward, this research lays the groundwork for further exploration into nuanced healthcare domains, fostering greater inclusivity, accessibility, and personalization within digital health interventions.

Acknowledgments

This work has been fully funded by the project Research and Development of Algorithms for Construction of Digital Human Technological Components supported by the Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

References

[Chi et al. 2023] Chi, Z., Huang, H., Liu, L., Bai, Y., Gao, X., and Mao, X.-L. (2023). Can pretrained english language models benefit non-english nlp systems in low-resource scenarios? IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32:1061–1074.

[Chiu et al. 2024] Chiu, T. K., Moorhouse, B. L., Chai, C. S., and Ismailov, M. (2024). Teacher support and student motivation to learn with artificial intelligence (ai) based chatbot. Interactive Learning Environments, 32(7):3240–3256.

[Conneau et al. 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

[Consoli et al. 2022] Consoli, B., dos Santos, H. D. P., Ulbrich, A. H. D. P. S., Vieira, R., and Bordini, R. H. (2022). BRATECA (Brazilian tertiary care dataset): a clinical information dataset for the Portuguese language. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5609–5616, Marseille, France. European Language Resources Association.

[Devlin et al. 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, Proceedings of the 2019 Conference of

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[dos Santos Junior et al. 2021] dos Santos Junior, J. B., Gomes, J., da Silva Dias, J., de Souza, L. N. O., Zanotti, A. C. N., Dias, R. P., and de Carvalho, Â. B. (2021). Uma proposta de chatbot para apoio a gestantes no contexto do sistema de saúde brasileiro. Revista Ibérica de Sistemas e Tecnologias de Informação, (E42):344–352.

[Douze et al. 2024] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. (2024). The faiss library. arXiv preprint arXiv:2401.08281.

[Elshin et al. 2024] Elshin, D., Karpachev, N., Gruzdev, B., Golovanov, I., Ivanov, G., Antonov, A., Skachkov, N., Latypova, E., Layner, V., Enikeeva, E., Popov, D., Chekashev, A., Negodin, V., Frantsuzova, V., Chernyshev, A., and Denisov, K. (2024). From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, Proceedings of the Ninth Conference on Machine Translation, pages 247–252, Miami, Florida, USA. Association for Computational Linguistics.

[García-Méndez et al. 2021] García-Méndez, S., De Arriba-Perez, F., González-Castaño, F. J., Regueiro-Janeiro, J. A., and Gil-Castiñeira, F. (2021). Entertainment chatbot for the digital inclusion of elderly people without abstraction capabilities. IEEE Access, 9:75878–75891.

[Gemini Team et al. 2023] Gemini Team, R. A., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

[Gong et al. 2023] Gong, Y., Liu, G., Xue, Y., Li, R., and Meng, L. (2023). A survey on dataset quality in machine learning. Information and Software Technology, 162:107268.

[Jin et al. 2021] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., and Szolovits, P. (2021). What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14).

[Kirubakaran et al. 2024] Kirubakaran, S., Kathrine, J. W., E, G. M. K., J, M. R., Singh A, R. G., and E, Y. (2024). A rag-based medical assistant especially for infectious diseases. In 2024 International Conference on Inventive Computation Technologies (ICICT), pages 1128–1133.

[Lewis et al. 2020] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, Advances in Neural Information Processing Systems, volume 33, pages 9459–9474. Curran Associates, Inc.

[Liu et al. 2024] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173.

[Miao et al. 2024] Miao, J., Thongprayoon, C., Suppadungsuk, S., Garcia Valencia, O. A., and Cheungpasitporn, W. (2024). Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. Medicina, 60(3):445.

[Montenegro et al. 2022] Montenegro, J. L. Z., da Costa, C. A., and Janssen, L. P. (2022). Evaluating the use of chatbot during pregnancy: A usability study. Healthcare Analytics, 2:100072.

[Oliveira et al. 2022] Oliveira, L. E. S. e., Peters, A. C., Da Silva, A. M. P., Gebeluca, C. P., Gumiel, Y. B., Cintho, L. M. M., Carvalho, D. R., Al Hasan, S., and Moro, C. M. C. (2022). Semclinbr-a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. Journal of Biomedical Semantics, 13(1):13.

[Passinato et al. 2024] Passinato, E., Rios, W., and Filho, A. G. (2024). Integração de modelos de linguagem e rag na criação de chatbots oftalmológicos. In Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde, pages 354–365, Porto Alegre, RS, Brasil. SBC.

[Puspitasari et al. 2022] Puspitasari, I. W., Rinawan, F. R., Purnama, W. G., Susiarno, H., and Susanti, A. I. (2022). Development of a chatbot for pregnant women on a posyandu application in indonesia: From qualitative approach to decision tree method. In Informatics, volume 9, page 88. MDPI.

[Robinson et al. 2022] Robinson, J., Rytting, C. M., and Wingate, D. (2022). Leveraging large language models for multiple choice question answering. arXiv preprint arXiv:2210.12353.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: Pretrained bert models for brazilian portuguese. In Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I, page 403–417, Berlin, Heidelberg. Springer-Verlag.

[Tawfik et al. 2023] Tawfik, E., Ghallab, E., and Moustafa, A. (2023). A nurse versus a chatbot–the effect of an empowerment program on chemotherapy-related side effects and the self-care behaviors of women living with breast cancer: a randomized controlled trial. BMC nursing, 22(1):102.

[Tu et al. 2024] Tu, T., Palepu, A., Schaekermann, M., Saab, K., Freyberg, J., Tanno, R., Wang, A., Li, B., Amin, M., Tomasev, N., et al. (2024). Towards conversational diagnostic ai. arXiv preprint arXiv:2401.05654.

[Wang et al. 2022] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., and Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533.

[Xiong et al. 2024] Xiong, G., Jin, Q., Lu, Z., and Zhang, A. (2024). Benchmarking retrieval-augmented generation for medicine. In Ku, L.-W., Martins, A., and Srikumar, V., editors, Findings of the Association for Computational Linguistics: ACL 2024, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

[Xu et al. 2021] Xu, L., Sanders, L., Li, K., Chow, J. C., et al. (2021). Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR cancer, 7(4):e27850.

[Ye et al. 2025] Ye, Y., Huang, Z., Xiao, Y., Chern, E., Xia, S., and Liu, P. (2025). Limo: Less is more for reasoning. arXiv preprint arXiv:2502.03387.

[Zhang et al. 2023] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2023). Siren's song in the ai ocean: A survey on hallucination in large language models. arXiv preprint arXiv:2309.01219, 2(5).