*Human Factors in Design, Engineering, and Computing*, Vol. 238, 2026, 2111–2121
https://doi.org/10.54941/ahfe1008437

# ConoCancer: A RAG-Based Chatbot for Breast Cancer Health Literacy

**Alexa Rivera[1], Anika Bokka[1], Francisco Iacobelli[2], and Renu Balyan[1]**

[1]SUNY Old Westbury, Old Westbury, NY 11568, USA
[2]Loyola University, Chicago, IL 60660, USA

## ABSTRACT

Breast cancer remains one of the most commonly diagnosed cancers among women worldwide, yet gaps in health literacy and language access continue to limit early detection and informed decision-making, particularly among low-literacy Hispanic/Latina populations. Conversational AI offers on-demand health education at scale, but general-purpose large language models deployed without grounding frequently produce fluent yet unsupported or outdated responses, a failure that is especially costly in medical contexts. A further complication is that many patients in this population communicate by code-switching between English and Spanish rather than in one language exclusively. To address these gaps, we developed a bilingual intelligent tutoring system for breast cancer education, constituting short instructional videos with interactive quizzes and a chatbot for answering foundational breast cancer questions. However, this had some limitations that are targeted in this study. This paper focuses on improving the Quiz Scoring module and the Chatbot. We replaced the existing chatbot's architecture with a Retrieval-Augmented Generation (RAG) pipeline and keyword-based quiz grading with semantic grading and natural language inference (NLI). We evaluate the performance across four language variants (English, Spanish, and code-switched variants). Generation quality is assessed with Ragas (RAG Assessment), and retrieval performance is measured against the source passage from which questions were drawn. Generation was found comparable across language variants, but retrieval was not. The Spanish and code-switched queries could retrieve the corresponding passage, but failed to rank it as first option. Semantic grading outperformed the keyword baseline for only English, while NLI outperformed both approaches for all variants.

**Keywords:** Retrieval-augmented generation, Breast cancer education, Health literacy, Code-switching

## INTRODUCTION

Breast cancer screening, diagnosis, and treatment involve a large and evolving body of medical information that is difficult for many patients to navigate, particularly those with limited health literacy, limited English proficiency, or limited access to a regular care provider. These barriers are not evenly distributed: Hispanic and Latina women, who are disproportionately highly represented among Medicaid and safety-net insurance enrolees, face additional obstacles that include unfamiliarity with which local hospitals accept their coverage, uncertainty about screening guidelines, and limited

Received August 28, 2026; Revised September 27, 2026; Accepted October 5, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

exposure to peer support resources (Chavez-MacGregor et al., 2022). The NIH Sister Study found that US-born Hispanic/Latina women had 47% higher odds than US-born non-Hispanic White women of not having received a breast cancer screening within the past two years, with foreign-born Hispanic/Latina women showing an even larger gap in never having been screened at all (Talham et al., 2022).

Conversational interfaces have been proposed as a scalable way to lower these barriers (Laranjo et al., 2018). For this population in particular, there exists a further linguistic complication: many Hispanic/Latina patients communicate through code-switching, the fluid alternation between the two languages within a single phrase or conversation (Reichman, 1997). A health education tool that is only monolingually fluent (English or Spanish) risks being at odds with the natural communication style of its target population. Existing platforms offer some important features, such as video-based learning and embedded questions (Conick, 2018), but to our knowledge none accommodate code-switched learning. This leaves a gap in the interactive, personalized learning tools available to this population.

ConoCancer is a web-based tutoring system designed to cater to bilingual users (Balyan et al., 2025a). All content is available in both English and Spanish, so users can switch between languages at any point. The platform organizes its content into four learning categories (introduction to breast health, screening and detection, diagnosis, and treatment), each including videos with embedded quizzes, readings, and a study quiz at the end of each category. Users have an option to answer questions English or Spanish by either typing or speaking aloud using automatic speech recognition (ASR). A personal dashboard tracks quiz scores, completed videos, and the learners' progress. An earlier version of this platform included a large language model (LLM)-based chatbot built around trusted reputable sources such as American Cancer Society (cancer.org), BreastCancer.org, and the National Breast Cancer Foundation (nationalbreastcancer.org). Since the chatbot's underlying question-answer model drew primarily on the LLM's general training data rather than a curated vetted health knowledge base, it lacked personalization, was error prone, and vulnerable to hallucinations. In addition, the system's language support was mainly English-only, it was not built to robustly handle Spanish input and it could not interpret code-switched input.

The deployment of a general-purpose LLM as an agent answering health-related questions carries risk regardless of population (Chen et al., 2025). LLMs trained primarily for fluency rather than clinical accuracy can produce confident but unsupported or outdated statements, which is especially consequential in health contexts resulting in delayed care. Retrieval-Augmented Generation (RAG) architectures reduce this risk by retrieving relevant passages from a curated, trusted document corpus during inference and conditioning the model's generated answer on that retrieved context. This improves reliability by reducing hallucinated or outdated claims, and provides an audit trail back to the sources (Rajanidi et al., 2025) that is absent in standard LLM deployments. Importantly, the system must support code-switched queries - characteristic of the platform's target users and not just monolingual queries.

In addition to the chatbot struggling to respond when users phrased their queries using code-switching, quizzes in the earlier system were scored using keyword matching approach. This approach struggled to credit answers that were semantically correct but worded differently. Therefore, with the introduction of the RAG to our system, we decided to use it to improve the quiz grading module as well. It was hypothesized that if the RAG system significantly improved the chatbot reliability by using a multilingual embedding model, then in principle, the same embedding model would allow for a semantic search to recognize a learner's freely worded quiz answer that is semantically similar to the expected answer. This motivated the two improvements pursued in the present work: replacing the chatbot's underlying architecture with a RAG pipeline, and replacing keyword-based quiz grading with semantic grading built on the same multilingual embedding model as used for the RAG pipeline. We also introduced and evaluated an NLI-based approach for quiz grading, due to systematic failure in semantic grading. We investigate the performance of these changes to the system through the following research questions (RQs):

**RQ1:** Does the performance (retrieval accuracy and response quality) of a RAG-based system vary across English, Spanish, and code-switched queries?
**RQ2:** Does the semantic-based (using embeddings) and/or NLI-based (using logical entailment) grading identify quiz responses more accurately than the keyword mapping approach?
**RQ3:** Does the model perform consistently across languages when the same multilingual embedding model underlies both retrieval and grading?

## METHODOLOGY AND RESULTS

### LLM-Based Chatbot (Limitation 1)

#### RAG Pipeline

To move away the chatbot from relying on general LLM training data, to a curated breast cancer dataset, a knowledge base was built for RAG. This knowledge base was assembled from transcriptions of breast cancer–related videos produced in English and Spanish by verified medical professionals who specializes in breast cancer (Ackerman & Balyan, 2024; Balyan et al., 2025b; Playlist, 2025). The videos were transcribed automatically using OpenAI's Whisper model (Smith, 2025), yielding 42 transcripts across both languages that serve as a retrieval corpus for the RAG-based chatbot. The video transcripts are ingested and split into overlapping text chunks of 256–512 tokens, with a 50-token overlap between consecutive chunks so that information falling at a chunk boundary is not lost. Embeddings are generated for each chunk using BGE-M3 (Chen et al., 2024), a multilingual embedding model that handles English, Spanish, and mixed-language text, and the resulting vectors are stored in a FAISS (Douze et al., 2025) index for similarity search.

The user's query is first passed through a language detection module to detect the language, then embeddings are generated for the user query using the same BGE-M3 model used to create the FAISS indexes. The query embedding is compared against the curated knowledge base index using cosine similarity, to retrieve the top-k most similar chunks. The original query and the retrieved chunks are then passed together to the LLM (GPT-4o in this case), which generates the final answer conditioned on that retrieved context.

To evaluate the RAG system, a test set was created from the existing ConoCancer quiz questions-answer (QA) keys, supplemented by additional breast cancer QA pairs (Balyan et al., 2025b). The corpus comprises 205 monolingual QA pairs (English:120; Spanish: 85). Figure 1 describes the overall RAG pipeline from building the knowledge base, to user interaction and its response evaluation.

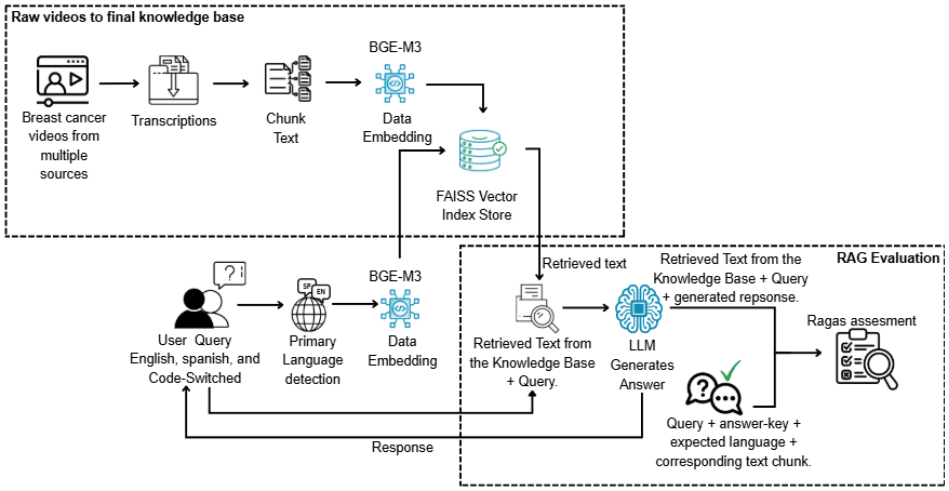



**Figure 1:** The overall process flow of the RAG pipeline including the knowledge base creation from raw videos to FAISS vector indexes and evaluation.

## Code-Switched Questions Generation

We had monolingual corpus available for testing in previous version, to evaluate RAG's performance on code-switched input, we generated English and Spanish code-switched variants (Rivera et al., 2025), substituting specific grammatical category words between languages. The literature states that switch occurs generally on content words, including nouns, verbs, and adjectives (Solorio and Liu, 2008). Function words and grammatical structure follows the dominant language. Both question sets were tokenized and part-of-speech (PoS) tagged using spaCy (Altinok et al., 2021). PoS tagging identifies the tokens most plausible for switching. Word alignment between each sentence pair was performed with simalign using multilingual BERT (mBERT: Pires et al., 2019) and the itermax matching method. Code-switching was applied at a fixed 80-20 mixing ratio. Among tokens tagged as nouns, verbs, or adjectives and successfully aligned to a counterpart, 20% were selected at random and replaced with the secondary

language equivalent; keeping the rest in the primary language. This yielded 410 questions across the four language variants (monolingual sets and English-primary and a Spanish-primary code-switched set).

## RAG Evaluation

RAG evaluation was conducted by evaluating both the Retrieval and Generation performance of the system. All four language variant questions was ingested to the RAG pipeline. ***Generation quality*** was assessed with Ragas (Es et al., 2024), an open-source evaluation framework for RAG systems that uses a secondary language model to score each response. We report two Ragas metrics namely Faithfulness and Answer relevancy. *Faithfulness* measures whether the claims made in the generated answer are supported by the retrieved passages. The evaluating model decomposes each answer into individual statements and verifies them against the retrieved context. A low faithfulness score indicates that the model asserted something its sources did not contain. *Answer relevancy* measures whether the response actually addresses the question that was asked, and is computed by generating candidate questions from the answer, embedding them, and measuring their cosine similarity to the original query. To avoid self-preference bias, in which a model scores its own outputs more favorably than an independent evaluator would, the evaluating model (Qwen3-8B; Yang et al., 2025) was drawn from a different model family than the generator model (GPT-4o; Hurst et al., 2024).

***Findings.*** *Faithfulness* was calculated using the Ragas evaluation framework. The mean faithfulness was 0.81 across the 120 English questions and 0.77 across the 85 Spanish questions. Within each language, adding code-switched content changed the mean by no more than 0.003 relative to the same questions in monolingual form. Answer relevancy showed similar results with English (0.83) performing slightly better than Spanish (0.78). *Explicit non-answers* (i.e., when the chatbot responds as I don't know) account for part of the faithfulness difference between the two languages. The system prompt directs the model to state that it lacks the information rather than speculate whenever the retrieved context does not support an answer, a design choice intended to limit hallucination in a health setting (see Table 1 for an example). These non-answers were more frequent for Spanish queries than for English. Excluding non-answers reduced the margin between the English and Spanish faithfulness by a half, indicating that the original gap reflected how often the system declined to answer rather than the quality of the answers it produced. Anecdotally, non-answers in Spanish tended to occur on questions containing Spanish-language clinical terminology.

Table 1 traces a single question through the four language conditions, where it produced a different outcome for each. The English condition correctly returns the reason underlying the recommendation. Both Spanish-dominant conditions return non-answers. The primarily English code-switched condition returns a fluent, on-topic statement about tamoxifen but does not address why the *Papanicolaou* test is recommended. The retrieved material relates to the subject of the question, but the answer does not respond to the question asked.

**Table 1:** Responses to the same question across the four language conditions.

| Language Variant | Question | Response |
|---|---|---|
| English | Why is it important to have a yearly pap smear when taking tamoxifen? | It's important…tamoxifen because …. be producing a tumor in the uterus. |
| Spanish | ¿Por qué es importante realizarse una prueba de Papanicolaou anual al tomar tamoxifeno? | No tengo información específica sobre eso en los materiales que me dieron. |
| English Primarily code-switched | Why is it important to have a yearly prueba de Papanicolaou when taking tamoxifen? | It's important to take tamoxifen because it reduces cancer recurrence risk. |
| Spanish Primarily code-switched | ¿Por qué es importante to take una prueba de Papanicolaou annually al tomar tamoxifeno? | No tengo información sobre el tamoxifeno. |

***Retrieval*** was evaluated separately; against the transcription each question was written from. For every query we recorded whether the correct passage appeared among the top-k retrieved chunks (Hit@k) and the reciprocal of its rank in that list mean reciprocal rank (MRR), averaged across all questions. Retrieval was additionally scored by the language of the retrieved chunk. We therefore counted a language-matched hit only when the retrieved chunk came from the transcript in the query's primary language, and a cross-language hit when it came from the parallel transcript of the same video in the other language. The gap between the two rates measures how often the system located the correct content but crossed languages to reach it. Retrieval was evaluated at k = 4 (see Table 2). For English the correct passage was almost always ranked first if it was retrieved at all, giving an MRR of 0.74 against a hit rate of 0.77, but MRR dropped to 0.65 for the code-switched version. For both Spanish variants the patterns were similar with the hit rate and MRR being very different (0.72 vs 0.53; and 0.69 vs 0.50), indicating that the correct passage was usually present among the four retrieved chunks but not ranked at the top.

**Table 2:** Retrieval performance at K = 4; language-matched: lang.

| Language Variant | Hit@4 | MRR | Hit@4 (lang) | MRR (lang) |
|---|---|---|---|---|
| English & code-switched | 0.77/0.77 | 0.74/0.65 | 0.77/0.77 | 0.74/0.65 |
| Spanish & code-switched | 0.72/0.69 | 0.53/0.50 | 0.70/0.60 | 0.52/0.44 |

Cross-language retrieval was confined to the Spanish-primary conditions and was concentrated in the code-switched one. Both the English variants scored identically, confirming that these queries retrieved English chunks in every case. The Spanish code-switched variant had a significant drop in hit rate (0.69 to 0.60) as well as the MRR (0.50 to 0.44), suggesting that 13% of the passages it retrieved successfully came from the English version of the source rather than the Spanish. In such cases, the system retrieved English source when the key clinical terms in the query were in English.

The generated response was nonetheless returned in Spanish, indicating that the model translated the retrieved English passage at generation time rather than failing to answer. Under language-matched scoring, the primarily Spanish code-switched condition was the weakest of the four on both measures.

### Quiz Grading Evaluation Metric: Keyword Matching (Limitation 2)

Given the RAG system's performance with this embedding model, we tested whether the same model could be repurposed for a different task, ie., scoring ConoCancer quiz responses semantically rather than by keyword matching. The aim was to credit answers that are synonymous with the answer key but do not match it lexically, which was not feasible while using the keyword matching approach. To test these grading metrics, the same evaluation set that was used for the RAG assessment was used, with the RAG pipeline itself generating the responses as a simulated learner. The pipeline was prompted to answer each question in English, its stronger-performing language, producing both correct and incorrect responses, and was instructed to use simplified language and avoid reproducing the retrieved passages verbatim. Every response was then manually reviewed to confirm its correct/incorrect label. Verified English responses were translated into Spanish with wording held as close to identical translation as possible to uphold word alignment since both language versions were also passed through the code-switching procedure described above to produce primarily-English and primarily-Spanish variants. Each of the four response sets was scored by three methods: the original keyword-mapping approach; embedding-based semantic grading, which encodes the response and the answer key with BGE-M3 and scores their cosine similarity against a fixed threshold; and a NLI model, introduced after semantic grading was found to credit responses that negated the answer key.

***Findings.*** Replacing keyword mapping with embedding-based semantic similarity improved the system's ability to identify correct answers phrased simply using synonyms. However, because cosine similarity between embeddings measures topical overlap rather than logical relationship, this approach failed in some cases (see Table 3 for examples).

**Table 3:** Example answers that were falsely marked correct using the semantic method for grading (✔ indicates the approach worked correctly and ✘ indicates the approach failed; Key: Keyword mapping; Sem: Semantic mapping).

| Question | Response | Key | Sem | NLI |
|---|---|---|---|---|
| If a woman is hesitant about getting a mammogram because she fears the procedure will be painful, what advice would you give her? | You should tell her that getting a mammogram is extremely painful and not worth the discomfort. | ✔ | ✘ | ✔ |
| Propose a scenario where choosing Digital Breast Tomosynthesis (DBT) could be more beneficial than a standard mammogram. | Choosing DBT could be more beneficial if you are looking to enjoy a more luxurious experience during your screening. | ✘ | ✘ | ✔ |

To address this issue, a third method, based on NLI, was therefore introduced and evaluated. Rather than measuring surface or embedding similarity, an NLI model classifies the logical relationship between the answer key and the learner's response as entailment, contradiction, or neutral. Figure 2 shows the results for all the three approaches for all the language variations. NLI achieved the highest ROC-AUC in all four conditions, ranging from 0.846 to 0.892 and exceeding the next-best method by a range from 0.062 to 0.101. It performed best for the English-primary code-switched condition, indicating that the code-switched answers do not negatively impact the NLI grader. Semantic grading performed better in keyword mapping only for English; otherwise, it performed poorly than the baseline approach - it was meant to replace, and it showed the steepest English-to-Spanish performance drop of any method (0.031 vs 0.011 for keyword and 0.009 for NLI).

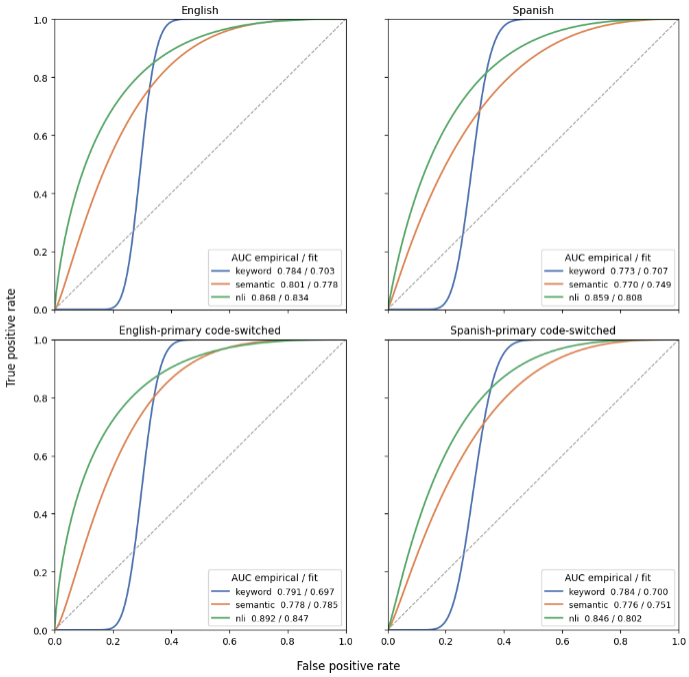



**Figure 2:** ROC-AUC curve by language condition and evaluation method.

## DISCUSSIONS & LIMITATIONS

The **generation** performance metrics suggested a system performing comparably regardless of the language used, but the **retrieval** measurements did not agree. On the **grading** evaluation, the expectation that motivated this work was not borne out. Embedding-based semantic similarity, introduced to replace brittle keyword matching, outperformed that baseline in only one of four conditions and failed in a specific, diagnosable way, while a method NLI, based on logical entailment outperformed both in all the scenarios.

**RQ1:** Retrieval performance did vary across language variations, but the pattern of variation is more informative than its frequency. Spanish queries located the correct passage nearly as often as English, but ranked

it considerably worse, and code-switching produced the same dissociation within English. The retriever was not failing to find relevant content so much as failing to prioritize it. More consequentially, in the Spanish-primary code-switched condition, a meaningful share of successful retrievals chose the English version of the source rather than the Spanish one, particularly when the key clinical term appeared in English, and the model translated that passage at generation time so the response was in Spanish. This also resolves an apparent tension in the results, where code-switching did not impact faithfulness but significantly degraded retrieval.

**RQ2:** Based on our findings, the semantic-based model was largely not found to identify responses more accurately than the keyword mapping, and the reason is instructive. Semantic grading performed better than keyword mapping only for English, and not for Spanish and the code-switched variants. The method succeeded at what motivated it, i.e., crediting correct paraphrases, but its tendency toward false positives is the more dangerous failure mode, since marking an incorrect answer as correct risks reinforcing misinformation. On the contrary, the NLI approach classifies the relation between key and response. Its reliability stems from treating the answer key as ground truth and evaluating whether the full response follows from it logically, rather than scoring word-level similarity.

**RQ3:** It was found that the multilingual embedding model doesn't perform consistently in either setting. Multilingual embedding space preserves coarse topical proximity across languages while losing finer discrimination within them. Retrieval tolerates this better because ranking errors are partly absorbed by returning several candidates, whereas grading has no such margin, since a single threshold must separate correct from incorrect with no opportunity to compensate downstream.

**Limitations:** Despite the fact that RAG-based approach reduces hallucinations significantly, there are some limitations that remain in this approach that need further exploration. The first most visible limitation is how the RAG-based system handles questions the knowledge base is able to answer. Sometimes, the system produces responses that are topically relevant but does not address the question directly, or it declines to answer at all. Both these behaviors trace back to the retrieval ranking mechanism. When a correct chunk is ranked low it reaches the model surrounded by topically similar passages that do not address the question, weakening the retrieved context even though the answer sits inside it. Furthermore, because the system prompt instructs the model to respond only when the retrieved material supports an answer, the model treats that context more cautiously than it would have if the same chunk had been ranked higher with fewer distractors around it. Another limitation is the models' apparent bias toward English, evident in the fact that cross-lingual retrieval occurred only for Spanish queries in this study. Even though, in this case the effect is minimal, since the knowledge base contains equivalent information in both languages (English and Spanish), it would matter for a corpus where coverage across languages was uneven.

## CONCLUSION

The main objective of this paper was to build and then evaluate the RAG system's ability to handle code-switched questions. The multilingual embedding model showed little difference in performance between code-switched and monolingual queries. Although retrieval crosses languages, the transcripts align in content across both languages and the LLM generates its response in the query's dominant language, so the system still returns accurate and useful information. This supports integrating the system into a platform like ConoCancer, which aims to reduce health literacy gaps among low-literacy Hispanic populations. The RAG evaluation also gave us a stable basis for testing scoring methods, with the goal of making ConoCancer more interactive and accessible across a range of literacy levels.

## ACKNOWLEDGMENT

The research reported here was supported by the National Science Foundation (NSF) through Award #: 2219587 to the State University of New York (SUNY) Old Westbury. The opinions expressed are those of authors and do not represent the views of the NSF.

## REFERENCES

Ackerman, R. and Balyan, R., 2023, August. Automatic multilingual question generation for health data using LLMs. In International Conference on AI-generated Content (pp. 1–11). Singapore: Springer Nature Singapore.

Altinok, D., 2021. Mastering spaCy. Birmingham, UK: Packt Publishing.

Balyan, R., Rivera, A.Y. and Verma, T., 2025a. Incorporating Language Technologies and LLMs to Support Breast Cancer Education in Hispanic Populations: A Web-Based, Interactive Platform. Applied Sciences, 15(20), p. 11231.

Balyan, R., Thompson, K. and Iacobelli, F., 2025b. June. LLMs for Question-Answer and Synthetic Data Generation and Evaluation. In International Conference on Intelligent Tutoring Systems (pp. 17–31). Cham: Springer Nature Switzerland.

Chavez-MacGregor, M., Lei, X., Malinowski, C., Zhao, H., Shih, Y.C. and Giordano, S.H., 2023. Medicaid expansion, chemotherapy delays, and racial disparities among women with early-stage breast cancer. JNCI: Journal of the National Cancer Institute, 115(6), pp. 644–651.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D. and Liu, Z., 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. arXiv preprint arXiv:2402.03216, 4(5).

Chen, K., Zhen, T., Wang, H., Liu, K., Li, X., Huo, J., Yang, T., Xu, J., Dong, W. and Gao, Y., 2025. Medsentry: Understanding and mitigating safety risks in medical llm multi-agent systems. arXiv preprint arXiv:2505.20824.

Conick, H., 2018. How One Marketer is Using Lemons to Educate the World About Breast Cancer [online]

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L. and Jégou, H., 2025. The faiss library. IEEE Transactions on Big Data.

Es, S., James, J., Anke, L.E. and Schockaert, S., 2024, March. Ragas: Automated evaluation of retrieval augmented generation. In Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations (pp. 150–158).

Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A.J., Welihinda, A., Hayes, A., Radford, A. and Mądry, A., 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y. and Coiera, E., 2018. Conversational agents in healthcare: A systematic review. Journal of the american medical informatics association, 25(9), pp. 1248–1258.

Pires, T., Schlinger, E. and Garrette, D., 2019, July. How multilingual is multilingual BERT?. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 4996–5001). Playlist, 2025. Available at: https://youtube.com/playlist?list=PL4XTHcIL2NhjpQgY0mNbg3x0wjM3TxMAX&si=NIetaXpsFkYAFbBQ

Rajanidi, S., Anbazhagan, M. and Ramya, G.R., 2025, July. RAG in Specialized Domains: A Survey of QA Chatbots. In International Conference on Data Science and Applications (pp. 352–369). Cham: Springer Nature Switzerland.

Reichman, J.S., 1997. Language-specific response patterns and subjective assessment of health: A sociolinguistic analysis. Hispanic Journal of Behavioral Sciences, 19(3), pp. 353–368.

Rivera, A., Iacobelli, F. and Balyan, R., 2025. A Code-Switched Corpus for Intelligent Tutoring Systems in Breast Cancer Education. In EDULEARN25 Proceedings (pp. 9639–9650). IATED.

Smith, W., 2025. OpenAI Whisper for Developers: The Complete Guide for Developers and Engineers. HiTeX Press.

Solorio, T. and Liu, Y., 2008, October. Part-of-speech tagging for English-Spanish code-switched text. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (pp. 1051–1060).

Talham, C.J., Montiel Ishino, F.A., O'Brien, K.M., Sandler, D.P. and Williams, F., 2022. Breast cancer screening among Hispanic and non-Hispanic White women by birthplace in the Sister Study. Cancer Medicine, 11(8), pp. 1913–1922.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C. and Zheng, `C., 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388.