Dual-Corpus Retrieval-Augmented Generation System for Women's Hormonal Health:

Component-Level Ablation Studies and Statistical Evaluation

Mikaela Connell

CSCI E-222 Foundations of Large Language Models
Harvard Extension School

May 2026

Video Presentation

Table of Contents

Abstract

Women with hormonal health conditions face systemic diagnostic delays that result in increased symptom severity, diminished quality of life, and adverse long-term reproductive outcomes (Li et al., 2025). Endometriosis, which affects approximately 10% of women of reproductive age, is associated with substantial delays between symptom onset and diagnosis, with global estimates averaging 6.8 years and country-level estimates ranging from 1.5 to 11.4 years (Gough et al., 2024). Polycystic ovary syndrome (PCOS), with an estimated prevalence of 10–13%, is similarly underrecognized (Teede et al., 2023; Gibson-Helm et al., 2017). These delays are shaped by several factors, including symptom normalization, diagnostic gender bias, limited specialist access, and the gap between the language patients use to describe symptoms and the language used in clinical evidence.

This project investigates whether retrieval-augmented generation (RAG) can produce evidence-grounded and accessible answers to women's hormonal health questions. I built a dual-corpus RAG system that pairs PubMedQA/PubMed abstract evidence with the National Library of Medicine's MedQuAD patient-facing question-answer dataset. The system uses a domain-specific bi-encoder, TimKond/S-PubMedBert-MedQuAD, with FAISS similarity search and compares three retrieval modes: clinical-only, patient-only, and dual-corpus retrieval. Retrieved context is passed to a 4-bit quantized BioMistral-7B generator, which produces a yes/no/maybe answer and a short explanation.

The system was evaluated on 111 women's-health questions filtered from PubMedQA. I measured generation accuracy using macro-F1, readability using Flesch-Kincaid grade level, and used qualitative review to assess retrieval relevance and groundedness. I also ran three ablation studies: cross-encoder reranking, generator substitution with Claude Haiku 4.5, and prompt variants designed to reduce excessive hedging. Bootstrap confidence intervals with 1,000 resamples were used to estimate uncertainty given the small evaluation set.

The results did not support the original dual-corpus hypothesis. Clinical-only RAG with BioMistral achieved the highest macro-F1 score at 0.344, outperforming the no-RAG BioMistral baseline at 0.222. Patient-only and dual-corpus retrieval improved readability somewhat but introduced a strong affirmative answer bias, while BioMistral's highest-scoring configuration

failed to generate any “maybe” predictions. Claude Haiku 4.5 showed the opposite pattern: it was more cautious and explicit about uncertainty, but overused “maybe” and performed worse on the PubMedQA benchmark. Overall, the results suggest that corpus choice matters more than corpus quantity. For patient-facing women’s health applications, future systems should separate answer determination from patient-friendly explanation: clinical evidence should be used to determine the answer, while a second stage should translate that evidence into accessible language.

1. Introduction and Problem Statement

1.1 Motivation

Women face persistent diagnostic delays for hormonal health conditions. Endometriosis, affecting approximately 10% of women, carries a 7-10 year delay between symptom onset and diagnosis (Karavadra et al., 2025; De Corte et al., 2025). Existing literature suggests that these delays are shaped by multiple factors, including symptom normalization, diagnostic gender bias, limited specialist access, and the difficulty patients face in having their symptoms recognized as clinically meaningful (Karavadra et al., 2025; Harrilal-Maharaj, 2025). Similar patterns appear in other hormonal conditions. Polycystic ovary syndrome (PCOS), which affects an estimated 10–13% of reproductive-age women, remains underrecognized despite established diagnostic criteria (Teede et al., 2023; Gibson-Helm et al., 2017). These delays can result in years of unmanaged symptoms, deferred fertility planning, and avoidable disease progression (Li et al., 2025).

I focus on one part of this problem: patients often describe symptoms in everyday language, while clinical evidence is written in technical language. Patients searching for explanations of their symptoms encounter medical literature written for trained practitioners; clinicians, in turn, may not always recognize lay descriptions of conditions whose presentations vary widely across individuals. This results in a documented health-literacy gap that can affect patient understanding and self-advocacy (Berkman et al., 2011). For women specifically, this translation problem may be compounded by gender bias in symptom interpretation, where women's reports of pain and other subjective symptoms are systematically underweighted compared with men's (Hoffmann & Tarzian, 2001).

Reen, a platform I am currently developing, supports continuous symptom and cycle tracking alongside evidence-grounded answers to user health questions, and is the inspiration behind this project. The core idea behind the platform is to give patients the ability to arrive at a clinical appointment, not with a list of vague complaints, but with their own tracked data alongside specific, medically-informed questions. This is an active form of self-advocacy that prior research has associated with shorter diagnostic timelines (Karavadra et al., 2025).

This project focuses specifically on the RAG component of that broader system: how to build and evaluate a retrieval-augmented question-answering pipeline that produces responses grounded in medical evidence and accessible to non-clinical readers. The system pairs two retrieval corpora: clinical evidence (PubMed abstracts) and patient-friendly explanations (the National Library of Medicine's MedQuAD question-answer dataset). The goal is to produce answers that are both grounded in medical evidence and accessible to non-clinical readers. The personal tracking and platform integration aspects of Reen are out of scope for this report and contain no proprietary code in this implementation.

1.2 Problem Statement

This project investigates whether retrieval-augmented generation can support evidence-based question answering for women's hormonal health. Given a natural language health question, the system retrieves relevant medical context from one or more corpora and generates a response with a yes/no/maybe label and a short explanation.

The project evaluates three retrieval configurations:

1. Clinical-only RAG, which retrieves from PubMedQA.
2. Patient-only RAG, which retrieves from MedQuAD.
3. Dual-corpus RAG, which retrieves from both PubMedQA and MedQuAD.

The main objective is to determine whether combining clinical and patient-facing retrieval improves answer quality compared with using either corpus alone. The system is evaluated using classification performance, retrieval quality, readability, groundedness, and qualitative examples of generated answers.

1.3 Project Scope and Contributions

The broader Reen platform includes personal health tracking and long-term product features that are outside the scope of this project. This report focuses only on the RAG-based question-answering component. The implementation uses public datasets, public model checkpoints or documented hosted models, and reproducible Python code. No proprietary Reen code or private user data is included.

This project contributes:

1. A working dual-corpus RAG pipeline combining PubMed and MedQuAD retrieval over a domain-specific bi-encoder.
2. A four-layer evaluation framework that covers retrieval relevance, generation accuracy, and readability.
3. Three independent ablation studies isolating the effects of retrieval improvements (cross-encoder reranking), generator strength (Claude Haiku 4.5 vs BioMistral-7B), and prompt design (grounding-focused variants).
4. A final system selection based on comparison across 17 experimental conditions, including baseline RAG, no-RAG controls, reranking, generator substitution, and prompt variants.
5. A statistical treatment via bootstrap confidence intervals to support inference on the small evaluation set (N=111).

2. Data Description and Preprocessing

2.1 Data Sources

This project uses publicly available medical datasets accessed through Hugging Face Datasets. The evaluation set was built from PubMedQA (https://huggingface.co/datasets/qiaojin/PubMedQA) (Jin et al., 2019), a biomedical question-answering benchmark containing questions derived from PubMed abstracts with yes/no/maybe answer labels. Specifically, the pqa_labeled split, which contains 1,000 expert-labeled questions, was used as the source for the filtered evaluation set. The pqa_unlabeled split, which contains 61,249 PubMed abstract-based examples, was used as the clinical retrieval corpus.

PubMedQA serves two roles in this project. First, it provides the labeled benchmark used to evaluate whether the system can correctly generate yes/no/maybe answers. Second, it provides a clinical evidence corpus for retrieval. This makes it well suited for testing whether retrieved biomedical literature can improve answer generation for women's hormonal health questions.

The patient-facing retrieval corpus was built from MedQuAD, a collection of medical question-answer pairs derived from authoritative U.S. government and medical information sources, including MedlinePlus, the National Cancer Institute, and the Genetic and Rare Diseases Information Center (https://huggingface.co/datasets/AnonymousSub/MedQuAD_47441_Question_Answer_Pairs) (Ben Abacha & Demner-Fushman, 2019). The specific Hugging Face variant used in this project, AnonymousSub/MedQuAD_47441_Question_Answer_Pairs, contains 47,441 question-answer pairs before cleaning. In contrast to PubMedQA, which is based on biomedical abstracts, MedQuAD is written in a more patient-facing question-and-answer format. This makes it useful for testing whether accessible medical explanations can complement clinical evidence in a RAG pipeline.

Together, these datasets support the central comparison in this project: clinical-only retrieval, patient-only retrieval, and dual-corpus retrieval. PubMedQA represents the clinical evidence source, while MedQuAD represents the patient-facing explanation source.

2.2 Evaluation Set Construction

The evaluation set was created by filtering the PubMedQA pqa_labeled split for women's health and hormonal health questions. The filtering process used a keyword list related to reproductive and hormonal health, including terms such as PCOS, polycystic ovarian syndrome, endometriosis, menstruation, menopause, fertility, infertility, ovulation, ovarian, uterine, pregnancy, mammography, and related concepts. The final evaluation set contained 111 labeled questions, with 50 yes, 49 no, and 12 maybe labels. This distribution is important because the "maybe" class is much smaller than the yes and no classes. For that reason, macro-F1 was used as the primary accuracy metric instead of accuracy alone. Macro-F1 gives equal weight to each class and better exposes whether the model fails on minority labels.

2.3 Retrieval Corpora Preprocessing

The clinical retrieval corpus was built from the PubMedQA pqa_unlabeled split. Each example contains a PubMed ID, a question, a structured abstract context, and a long answer. The patient-facing retrieval corpus was built from MedQuAD. Each MedQuAD record contains a medical question and answer. During preprocessing, records with missing or empty questions or answers were removed. The original MedQuAD dataset contained 47,441 entries. After removing null or empty records, the final patient-facing corpus contained 16,407 question-answer pairs. This means 31,034 entries, or 65.4% of the original dataset, were dropped during cleaning. MedQuAD questions were short and patient-facing, while PubMedQA documents were abstract-style clinical evidence.

2.4 Embedding and Index Construction

Both corpora were embedded using TimKond/S-PubMedBert-MedQuAD, a sentence-transformer model (Reimers & Gurevych, 2019) based on PubMedBERT (Gu et al., 2021) and fine-tuned for medical question-answering similarity. Embeddings were L2-normalized before indexing, and FAISS IndexFlatIP (Johnson et al., 2019) was used for similarity search. Because the embeddings were normalized, inner product similarity is equivalent to cosine similarity.

3. Models and Methods

3.1 RAG Architecture Overview

The system follows a retrieval-augmented generation architecture. Given a user question, the system first embeds the question using the same biomedical sentence encoder used to embed the corpora. It then retrieves relevant documents from one or both FAISS indexes. The retrieved documents are inserted into a prompt, and the prompt is passed to BioMistral-7B to generate an answer.

The pipeline has four main stages:

1. Embed the user question
2. Retrieve relevant documents from the selected corpus or corpora
3. Construct a prompt using the retrieved context
4. Generate a yes/no/maybe answer and explanation

3.2 Retriever

The retriever used the TimKond/S-PubMedBert-MedQuAD sentence-transformer model. This model was selected because the project required matching medical questions to relevant biomedical and patient-facing medical text. Since the model is domain-specific, it was expected to perform better on medical terminology than a general purpose sentence embedding model.

One important limitation is that the encoder was fine-tuned on MedQuAD. This may bias similarity scores toward patient-style question-answer pairs. To address this, the project did not rely only on FAISS cosine similarity as a retrieval-quality metric. Instead, retrieval relevance was also evaluated using an LLM-as-judge workflow.

3.3 Generator

The primary generator was BioMistral-7B, a biomedical-domain model fine-tuned from Mistral 7B (Jiang et al., 2023; Labrak et al., 2024). Because the project was implemented in Google Colab, the model was loaded using 4-bit NF4 quantization with bitsandbytes to reduce GPU memory requirements.

The model was prompted to produce a structured answer beginning with one of three labels: yes, no, or maybe. This format was chosen because PubMedQA uses yes/no/maybe labels, which

allows generation to be evaluated as a classification task while still preserving natural-language explanations.

3.4 Prompting Strategy

The prompt instructed BioMistral to answer using the retrieved context and to begin with a yes/no/maybe label. The generated response was then parsed by extracting the first clear yes/no/maybe token from the model output. The use of a structured label was important for evaluation, but it also exposed one of the system's weaknesses: the model often avoided the "maybe" label entirely, even when the gold answer was maybe.

4. Implementation Details

4.1 Environment and Dependencies

The project was implemented in Python 3 and run on Google Colab with a T4 GPU. The main libraries used were:

- **PyTorch**: base deep learning framework
- **Transformers**: model loading and tokenization for BioMistral
- **Bitsandbytes**: 4-bit NF4 quantization to fit BioMistral-7B on the T4
- **Accelerate**: device placement for the quantized model
- **Sentence-transformers**: bi-encoder for retrieval (TimKond/S-PubMedBert-MedQuAD) and cross-encoder for reranking (ms-marco-MiniLM-L-6-v2)
- **Faiss-cpu**: exact similarity search over normalized embeddings
- **Datasets**: Hugging Face dataset loading for PubMedQA and MedQuAD
- **Anthropic**: Claude Haiku 4.5 generation and LLM-as-judge calls
- **Textstat**: Flesch-Kincaid grade level calculation
- **Scikit-learn**: macro-F1 and confusion matrix computation
- **Pandas, numpy, matplotlib**: data handling, statistical computation, and figures

4.2 Setup Instructions

The project was developed across three main notebooks:

1. **Notebook1_Data_and_Index.ipynb**: loads PubMedQA and MedQuAD from Hugging Face, filters the labeled split for women's hormonal health questions, builds the clinical and patient corpora, embeds both with the bi-encoder, and saves FAISS indexes and embeddings to a Drive cache. This notebook runs once and takes about 15–20 minutes on a T4.
2. **Notebook2_Experiments.ipynb**: loads the cache from Notebook 1 and runs all generation experiments: the baseline RAG sweep across three retrieval modes, the no-RAG baselines for BioMistral and Claude, cross-encoder reranking, Claude generator substitution, and the three Claude prompt variants. It also runs the LLM-as-judge layers for retrieval relevance and groundedness. All outputs are saved as JSON files. Total

runtime is approximately 2.5 to 3 hours, but each generation loop checkpoints every 20 iterations so the notebook can be resumed if Colab disconnects.

3. **Notebook3_Analysis_and_Report.ipynb**: loads all result JSONs, parses predictions, computes macro-F1 and readability for each condition, builds the master comparison table, computes bootstrap confidence intervals, and generates the figures used in this report. This notebook does not require a GPU or API access and runs in about three minutes.

The notebooks use a shared Drive cache directory so that artifacts (embeddings, FAISS indexes, evaluation set, and experiment results) are computed once and reused across subsequent runs.

4.3 Reproducing the Results

To reproduce the results:

1. Open each notebook in Google Colab with a T4 GPU runtime.
2. Mount Google Drive when prompted in the first cell of each notebook.
3. Add your ANTHROPIC_API_KEY to Colab Secrets and grant notebook access (required for Notebook 2 only).
4. Run Notebook1_Data_and_Index.ipynb to build the cache.
5. Notebook2_Experiments.ipynb - If the runtime disconnects, simply re-run the cells; each generation loop will resume from its last checkpoint.
6. Run Notebook3_Analysis_and_Report.ipynb to regenerate every table and figure from the saved results.

All experimental conditions in this report can be reproduced from the cached results without re-running the BioMistral or Claude generations.

5. Experiments and Results

5.1 Evaluation Framework

Medical RAG can fail in several ways, so the system was evaluated across four layers. For example, the model can retrieve the wrong context, generate the wrong yes/no/maybe label, produce an answer that is not supported by the retrieved sources, or write an answer that is too technical for a non-clinical reader.

The four evaluation layers were:

1. **Retrieval relevance**: measured by an LLM-as-judge protocol that scores each retrieved document for topical alignment with the question on a 0–2 scale (0 = unrelated, 1 = related, 2 = directly relevant). This layer measures whether the retrieved documents were actually relevant to the question.
2. **Generation accuracy**: measured by macro-F1 on the yes/no/maybe labels. Macro-F1 was chosen over accuracy because of the class imbalance in the evaluation set, and because it weights minority-class performance equally with majority classes. The final evaluation set contained 50 yes questions, 49 no questions, and only 12 maybe questions. Macro-F1 gives equal weight to all three labels, so it makes it easier to see when the model is failing on the smaller maybe class. This layer measures whether the generated label matched the PubMedQA gold label.
3. **Groundedness**: measured by an LLM-as-judge protocol that scores each generated response on a 1-5 scale for the degree to which the response is supported by the retrieved context. This layer detects hallucination and whether the answer was supported by the retrieved context independent of label accuracy.
4. **Readability**: measured by Flesch-Kincaid grade level computed on the generated explanation text (Kincaid et al., 1975). This layer addresses whether the system produces output accessible to non-clinical readers.

The full baseline experiment produced 333 generations: 111 evaluation questions across three retrieval modes (clinical-only, patient-only, and dual-corpus).

5.2 Baseline RAG Results

The first experiment compared the three main RAG configurations. The original hypothesis was that dual-corpus retrieval would perform best because it combines clinical evidence from PubMedQA with patient-facing explanations from MedQuAD.

The hypothesis was not supported by the results: Clinical-only RAG performed best on macro-F1, with a score of 0.344. Patient-only RAG had the lowest macro-F1 at 0.260, while dual-corpus RAG was only slightly higher at 0.271. Patient-only RAG predicted “yes” for 101 out of 111 questions. Dual-corpus RAG behaved similarly, predicting “yes” for 99 out of 111 questions. This shows a strong yes-bias when MedQuAD context was included. MedQuAD is written as patient-facing educational content, so it often explains what a condition is, what symptoms are associated with it, and what patients should know.

Clinical-only RAG predicted “yes” 69 times and “no” 42 times, but still never predicted “maybe.” That demonstrated how the baseline RAG system was not good at expressing uncertainty, even when the evidence was incomplete or mixed. For a medical QA system, that is a significant limitation.

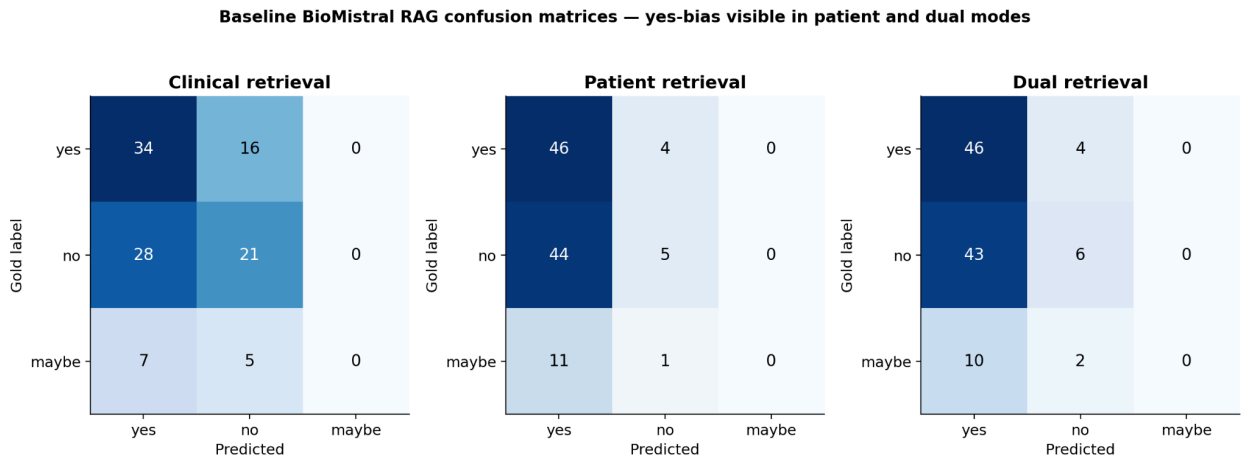

Figure 1. Confusion matrices for the three baseline RAG configurations. The bottom row of each matrix is empty: BioMistral never predicts "maybe" across any retrieval mode, despite 12 of 111 gold labels being "maybe".

The readability results partially supported the patient-facing retrieval idea. Patient-only RAG had the lowest Flesch-Kincaid grade level at 13.22, compared with 14.75 for clinical-only RAG and 14.51 for dual-corpus RAG. However, all three modes were still written at a college reading level, so patient retrieval alone was not enough to make the answers truly patient-accessible.

5.3 No-RAG Baseline

To test whether retrieval was helping at all, I also ran a no-RAG BioMistral baseline. In this setting, BioMistral answered the same 111 questions without any retrieved context.

No-RAG BioMistral achieved a macro-F1 of 0.222. Clinical RAG with BioMistral achieved 0.344. The bootstrap confidence intervals did not overlap: no-RAG BioMistral had a 95% CI of [0.184, 0.262], while clinical RAG had a 95% CI of [0.279, 0.403].

This suggests that retrieval did provide measurable value for BioMistral. Even though the dual-corpus hypothesis did not hold, the clinical RAG setup still improved over the generator alone.

5.4 Ablation Studies

After the baseline experiment, I ran three ablation studies to better understand which part of the system was limiting performance. Each ablation changed one component while keeping the rest of the pipeline the same.

The three ablations were:

1. **Cross-encoder reranking**, to test whether better retrieval ordering improved answers.
2. **Claude generator substitution**, to test whether a stronger instruction-following generator improved answers.
3. **Claude prompt variants**, to test whether prompt wording could reduce over-hedging.

5.4.1 Cross-Encoder Reranking (Phase 4A)

The first ablation tested whether retrieval quality was the main issue. The baseline retriever used a bi-encoder and FAISS similarity search. In the reranking experiment, the system first retrieved

a larger set of top-20 candidate documents and then used a cross-encoder, cross-encoder/ms-marco-MiniLM-L-6-v2, to rerank those candidates. The final top-5 documents were passed to BioMistral.

Reranking did not improve macro-F1. The clinical mode decreased slightly from 0.344 to 0.318, patient mode stayed the same at 0.260, and dual mode decreased from 0.271 to 0.247. However, these differences were not statistically meaningful because the bootstrap confidence intervals overlapped.

The takeaway is not that reranking clearly made the system worse. The more accurate conclusion is that this reranking setup did not produce a reliable improvement.

One possible reason is that the cross-encoder was trained on MS MARCO, a general-domain web search dataset, not biomedical QA. It may have improved topical matching in some cases, but that did not translate into better yes/no/maybe answers. A biomedical cross-encoder might perform differently.

5.4.2 Claude Generator Substitution (Phase 4B)

The second ablation tested whether the generator was the main bottleneck. I replaced BioMistral-7B with Claude Haiku 4.5 while keeping the same retrieval modes and overall prompt structure. Claude experiments were run through the Anthropic API using the Claude Haiku 4.5 model available at the time of experimentation.

This resulted in Claude performing much worse on macro-F1, with all three retrieval modes scoring 0.066. The reason was clear: Claude predicted "maybe" for almost every question. When I manually reviewed several Claude outputs, Claude was actually using the retrieved context and often explaining that the evidence did not directly answer the exact question. Claude's caution may even be safer than BioMistral's tendency to force yes/no answers. But for the PubMedQA benchmark, where the output must match a yes/no/maybe gold label, Claude’s caution was heavily penalized.

This ablation showed that a stronger general-purpose model does not automatically produce better benchmark performance. Claude was more cautious and more explicit about evidence

gaps, but that behavior reduced macro-F1.

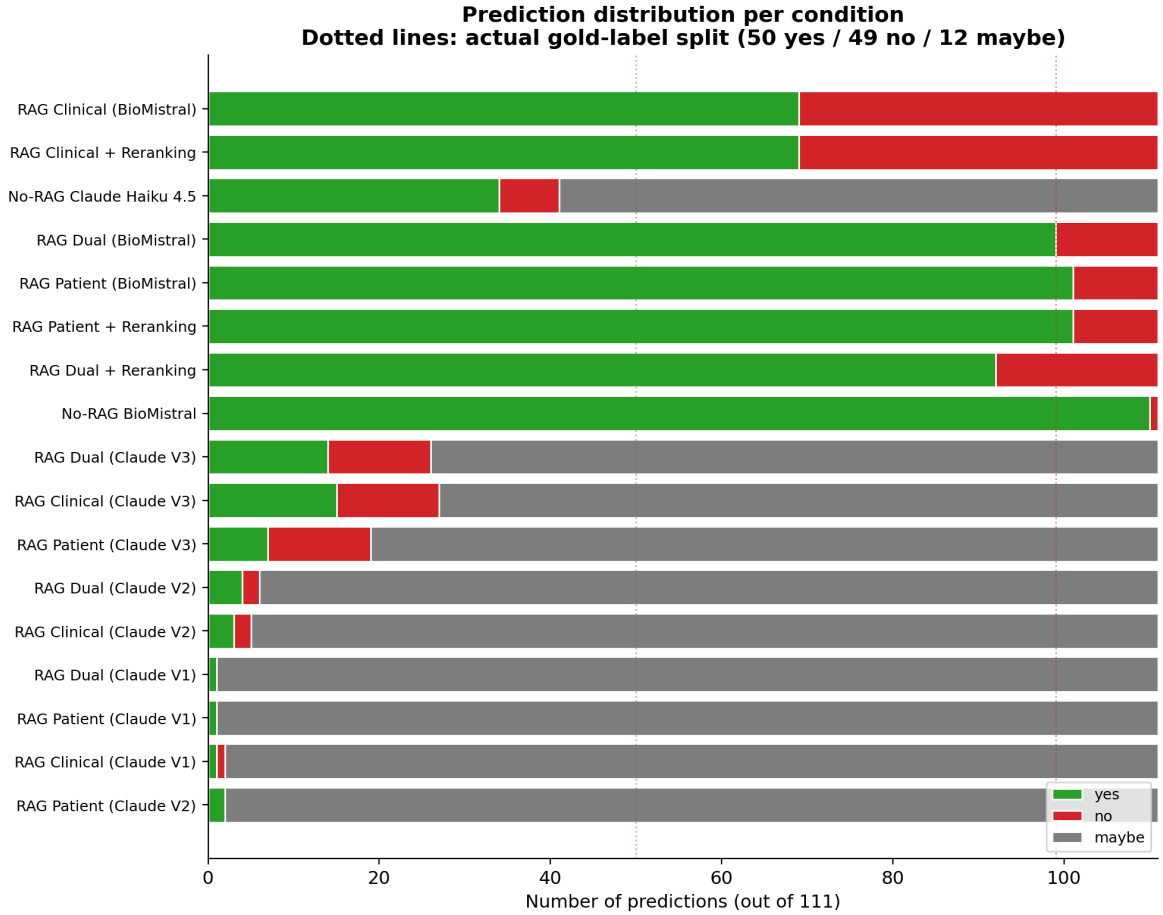

Figure 2. Prediction distributions across baseline (BioMistral) and Claude V1 conditions, with the gold label distribution at top for reference. Swapping in Claude collapses nearly all predictions to "maybe" across all three retrieval modes.

5.4.3 Claude Prompt Design Variants (Phase 4C)

The third ablation tested whether prompt design could reduce Claude's overuse of "maybe." I tested three Claude prompt versions:

- **V1 (original)** - instructed Claude to answer with yes, no, or maybe based on the retrieved evidence, with "maybe" presented as the appropriate response when evidence was partial or uncertain. This was the prompt used in Phase 4B.

- **V2 (reworded hedge)** - softened the “maybe” instruction by removing language that explicitly invited hedging on partial evidence, while preserving the option of “maybe” for genuinely unclear cases.
- **V3 (explicit anti-hedge)** - instructed Claude that the retrieved evidence would rarely match the question's exact wording or scope, that its task was to synthesize related findings into a directional answer, and that “maybe” should be reserved for cases where the retrieved sources were completely unrelated or contained genuinely contradictory findings. Excessive hedging was framed as an incorrect interpretation of the instructions.

Prompt design helped. V3 improved Claude’s macro-F1 across all three retrieval modes. The best Claude result was dual-corpus RAG with the V3 prompt at 0.187.

However, Claude still predicted “maybe” most of the time, even with the anti-hedge prompt. In the V3 Claude configuration, it still produced 85 “maybe” predictions out of 111. This suggests that prompt design can shift Claude’s behavior, but it cannot fully remove its tendency to be cautious when retrieved medical evidence is incomplete. The main takeaway is that prompt design had a strong impact, but it was not enough to make Claude outperform BioMistral on macro-F1.

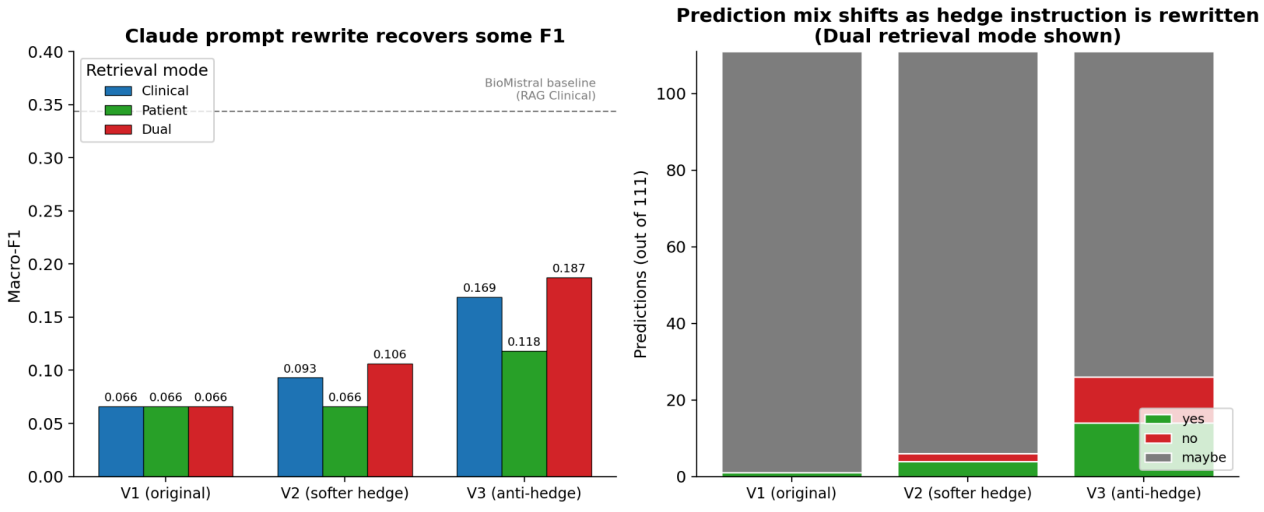

Figure 3. Macro-F1 (left) and prediction distribution (right) across the three Claude prompt variants. V3 anti-hedge framing recovers a portion of the F1 lost in V1 by shifting predictions away from "maybe", but does not reach the BioMistral baseline (dashed line).

5.5 Final System Selection

Across all 17 tested conditions, the highest macro-F1 came from the original clinical-only RAG system with BioMistral. If the goal is to maximize PubMedQA macro-F1, BioMistral clinical RAG is the best configuration. But if the goal is patient-facing medical support, the Claude V3 behavior is notable since it is more cautious.

For this project, I report clinical RAG with BioMistral as the final selected system because it performed best on the formal benchmark.

5.6 Bootstrap Confidence Intervals

To address the small evaluation set size (N=111), 95% confidence intervals on macro-F1 were computed for each condition via bootstrap resampling (Efron, 1979). For each condition, I resampled the 111 evaluation questions with replacement 1,000 times and recalculated macro-F1. The 2.5th and 97.5th percentiles were used as the 95% confidence interval.

The bootstrap analysis supports five main conclusions. First, retrieval helps BioMistral. RAG Clinical reaches macro-F1 of 0.344 with a 95% CI of [0.279, 0.403], while no-RAG BioMistral reaches only 0.222 with a CI of [0.184, 0.262]. Second, cross-encoder reranking did not meaningfully improve performance over baseline retrieval. The reranked clinical condition (0.318, CI [0.251, 0.374]) overlaps substantially with the baseline clinical condition (0.344, CI [0.279, 0.403]), and patient and dual modes show the same pattern. The point-estimate decrease of about 0.026 shows that reranking does not provide meaningful improvement, although it does not substantially hurt performance.

Regarding prompt design, the results in the Phase 4C prompt redesign proved to be helpful for Claude. In the dual-retrieval setting, Claude V1 scored 0.066 with a 95% CI of [0.034, 0.098], while Claude V3 scored 0.187 with a 95% CI of [0.110, 0.261]. The non-overlapping intervals suggest that prompt wording meaningfully changed Claude’s behavior. Despite prompt design changes, no Claude configuration matched the BioMistral baseline. Even the strongest Claude variant (RAG Dual with V3, 0.187, CI [0.110, 0.261]) has a confidence interval entirely below the BioMistral baseline's lower bound of 0.279. The closest comparison was between RAG

Clinical and BioMistral (0.344, CI [0.279, 0.403]) and no-RAG Claude Haiku 4.5 (0.292, CI [0.207, 0.373]). On macro-F1 alone, BioMistral performed best.

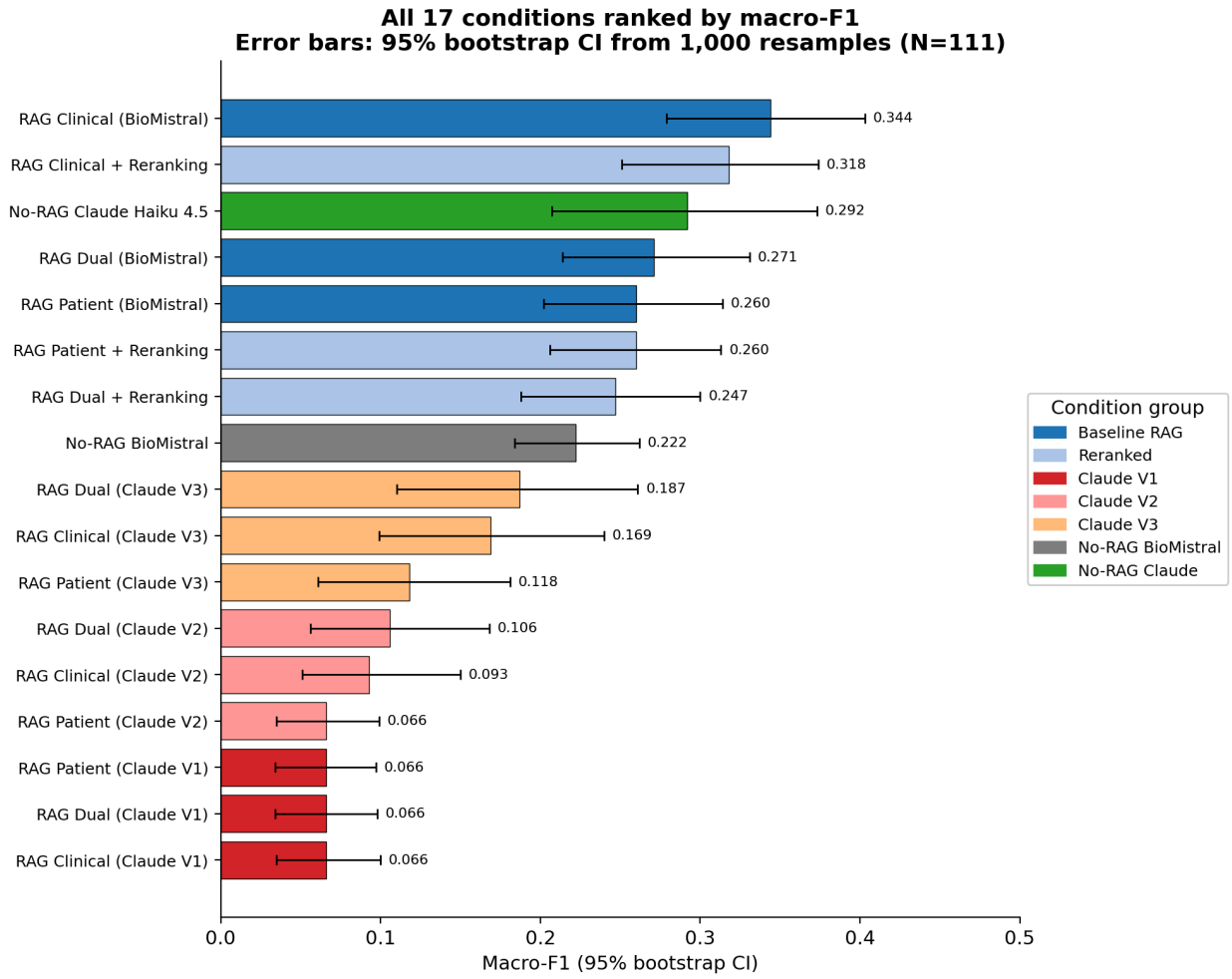

Figure 4. Macro-F1 with 95% bootstrap confidence intervals for all 17 experimental conditions. RAG Clinical with BioMistral has the highest point estimate but its interval substantially overlaps with no-RAG Claude Haiku 4.5.

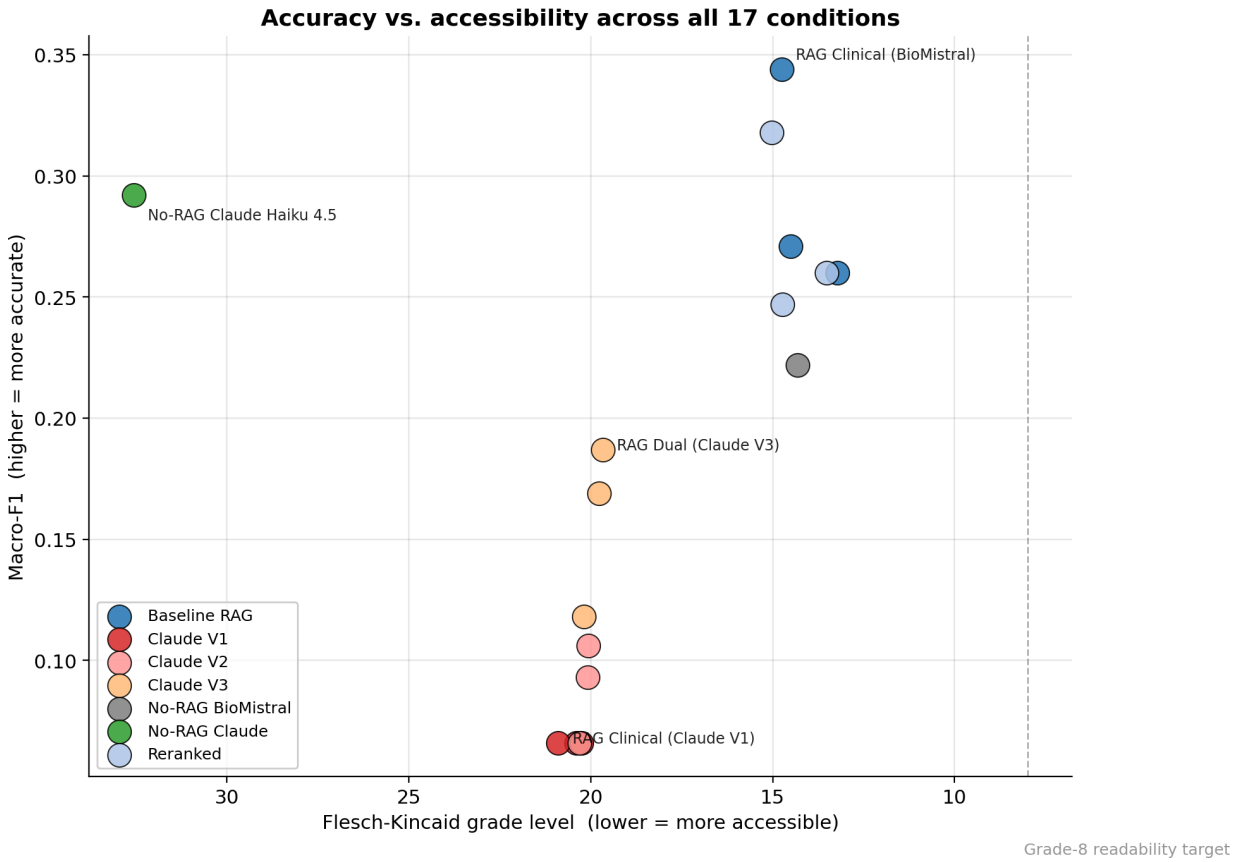

Figure 5. Accuracy and readability trade-off across all 17 conditions. No configuration reaches the grade-8 readability target while maintaining macro-F1 above the BioMistral baseline.

6. Discussion

6.1 What Worked Well

Several parts of this project worked well, including using multiple evaluation layers. In addition to evaluating the accuracy of the yes/no/maybe label, I chose to evaluate the system on several categories, and separated retrieval relevance, generation accuracy, groundedness, and readability, to make it easier to identify where the system was succeeding and failing.

The ablation study design helped make the results more useful than a single baseline comparison would have been. I tested three possible sources of weakness in the system: retrieval quality, generator voice, and prompt design. Cross-encoder reranking tested whether improving the ordering of retrieved documents would improve the answer quality. Generator substitution tested whether a stronger instruction-following model would improve responses. The different prompts tested whether Claude could reduce the amount of hedging. Although these changes did not improve the final macro-F1 score, it was useful to explain the behavior of the system.

Additionally, the Claude prompt experiments showed that prompt design had an impact on results. Claude V1 produced almost all “maybe” answers, but the V3 prompt made Claude more willing to give directional yes/no answers. Macro-F1 increased from 0.066 in the original Claude prompt to 0.187 in the best Claude V3 dual-corpus condition. Although this did not outperform the BioMistral baseline, it showed that prompt wording can meaningfully change model behavior.

Lastly, bootstrap confidence intervals were useful because the evaluation set was small. Some differences looked important from the point estimates alone, but the confidence intervals showed that they were not reliable at this sample size. This was especially important for interpreting the reranking results, where the small decreases in macro-F1 should not be treated as meaningful evidence that reranking made the system worse.

6.2 What Did Not Work

The biggest result that did not work as expected was the dual-corpus retrieval setup. My original hypothesis was that combining clinical evidence from PubMedQA with patient-facing explanations from MedQuAD would improve both accuracy and accessibility. That did not

happen. Clinical-only retrieval performed best on macro-F1, while patient-only retrieval was more readable but less accurate. Dual-corpus retrieval did not give the best of both worlds.

One reason appears to be the style of the MedQuAD corpus. MedQuAD is written as patient education content. It often explains what a condition is, what symptoms are associated with it, and what patients should know. That makes it useful for patient understanding, but in this benchmark it seemed to push the generator toward “yes” answers.

Cross-encoder reranking also did not improve the final results. The reranker sometimes appeared to improve the topical relevance of the retrieved documents, but that improvement did not translate into better yes/no/maybe answers.

6.3 Lessons Learned

The biggest lesson is that more context is not automatically better. Adding MedQuAD to PubMedQA did not improve the system. In fact, the patient-facing context appeared to change the model’s answer style in a way that hurt performance on this benchmark. This suggests that corpus alignment matters more than simply adding more documents.

In addition, prompt design proved to be a critical point of the system, and Claude’s behavior changed substantially across the V1, V2, and V3 prompts. This shows that when comparing generators, the prompt can strongly affect the result.

6.4 Future Extensions

There are several ways I would improve this project in future work. To continue working on this, I would test a biomedical cross-encoder reranker. The reranker used in this project was trained on general-domain web search data. A reranker trained on biomedical relevance pairs might be better at selecting answer-bearing clinical evidence. Next, I would fine-tune or adapt the bi-encoder on PubMedQA-style contrastive pairs. The current embedding model was fine-tuned on MedQuAD, which is structurally different from PubMedQA abstracts. Training the retriever on examples that better match the clinical corpus could improve retrieval quality. Lastly, I would expand the evaluation set. The final evaluation set had 111 questions, and only 12 of them were labeled “maybe.” A larger and more balanced evaluation set would make it easier to draw stronger conclusions, especially about uncertainty handling.

7. Limitations and Responsible Use

7.1 Sample Size and Statistical Power

The evaluation set contained 111 questions, with 50 “yes” labels, 49 “no” labels, and 12 “maybe” labels. This is enough for an exploratory class project, but it is still a small sample. The “maybe” class is especially limited, so results related to “maybe” recall should be interpreted carefully.

To address this, I used bootstrap confidence intervals with 1,000 resamples. The confidence intervals helped show which differences were meaningful and which were likely due to sampling variation. For example, the reranking results looked slightly worse than the baseline in point estimates, but the confidence intervals overlapped, so I would not claim that reranking truly hurt performance.

7.2 Bias Considerations

There are several sources of bias in this project. First, MedQuAD is mostly educational patient-facing content. It is helpful for explaining medical concepts, but it may not contain enough examples of negation, uncertainty, or conflicting evidence. This likely contributed to the yes-bias in patient-only and dual-corpus retrieval. Second, PubMedQA is based on PubMed abstracts, which inherit the biases of published biomedical literature. Some conditions are studied much more than others, and positive findings may be more likely to appear in the literature than negative findings. This affects what evidence is available for retrieval. Third, the women’s-health evaluation set was created using keyword filtering. That means it may miss questions that are relevant but use different language. This is especially important for patient-facing applications, where users may describe symptoms in informal or non-clinical terms. Fourth, the embedding model was fine-tuned on MedQuAD. This may make it better at matching patient-facing question-answer pairs than abstract-style clinical evidence. That difference may have affected retrieval behavior.

7.3 Hallucination and Groundedness Risks

The system still has hallucination risk. BioMistral often generated confident answers, and in qualitative review some answers did not appear strongly tied to the retrieved context. This is a major concern for medical QA because a fluent answer can sound trustworthy even when it is not

well supported. Claude behaved differently. It was more likely to acknowledge when evidence was incomplete, but it also overused the “maybe” label under the original prompt.

7.4 Medical Disclaimer and Intended Use

This project is an academic technology demonstration. It is not a medical device and has not been validated for clinical use. It should not be used for diagnosis, treatment decisions, or as a substitute for a licensed healthcare professional.

The intended use case is educational support and question preparation. A system like this could help a user better understand medical information and prepare more specific questions for a clinician, but it should not make medical decisions.

7.5 LLM-as-Judge Limitations

LLM-as-judge evaluation can be useful, but it has limitations. LLM judges can have their own biases, and if the same model family is used as both generator and judge, there is a risk of self-preference. Future work should validate LLM-as-judge results against human reviewers, especially clinicians or medically trained annotators.

In this project, LLM-as-judge evaluation helped identify patterns in retrieval relevance and groundedness, but it should not be treated as a substitute for expert medical review.

8. Conclusion

This project built and evaluated a dual-corpus RAG system for women's hormonal health question answering. The system used PubMedQA as a clinical evidence source and MedQuAD as a patient-facing explanation source. It combined a domain-specific bi-encoder, FAISS similarity search, and LLM generation to answer women's-health questions with a yes/no/maybe label and a short explanation.

The original hypothesis was that dual-corpus retrieval would improve both accuracy and accessibility. The results did not support that hypothesis. Clinical-only RAG with BioMistral achieved the highest macro-F1 score, while patient-only retrieval produced the most readable answers. Dual-corpus retrieval did not outperform the best single-corpus configurations.

The ablation studies showed that the system's behavior depended heavily on the generator and prompt. Cross-encoder reranking did not produce a statistically meaningful improvement. Claude Haiku 4.5 produced more cautious answers but overused maybe, which reduced macro-F1. Prompt design improved Claude's performance, but not enough to match the BioMistral clinical RAG baseline. Overall, the results suggest that retrieval quality depends less on corpus size and more on whether the corpus matches the decision being evaluated.

9. References

Ben Abacha, A., & Demner-Fushman, D. (2019). A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1), 511. https://doi.org/10.1186/s12859-019-3119-4

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. (2011). Low health literacy and health outcomes: An updated systematic review. *Annals of Internal Medicine*, 155(2), 97–107. https://doi.org/10.7326/0003-4819-155-2-201107190-00005

De Corte, P., Klinghardt, M., von Stockum, S., & Heinemann, K. (2025). Time to diagnose endometriosis: Current status, challenges and regional characteristics—A systematic literature review. *BJOG: An International Journal of Obstetrics & Gynaecology*, 132(2), 118–130. https://doi.org/10.1111/1471-0528.17973

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Gibson-Helm, M., Teede, H., Dunaif, A., & Dokras, A. (2017). Delayed diagnosis and a lack of information associated with dissatisfaction in women with polycystic ovary syndrome. *The Journal of Clinical Endocrinology & Metabolism*, 102(2), 604–612. https://doi.org/10.1210/jc.2016-2963

Gough, K., Brand, A., Manderson, L., Healey, M., Cheng, Y., Zarnowiecki, D., & Girling, J. E. (2024). Understanding diagnostic delay for endometriosis: A scoping review using the social-ecological framework. *Health Care for Women International*, 45(11–12), 1342–1366. https://doi.org/10.1080/07399332.2024.2413056

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), 1–23. https://doi.org/10.1145/3458754

Harrilal-Maharaj, K. (2025). Gender bias and diagnostic delays in young women: A narrative review. *Cureus*. https://doi.org/10.7759/cureus.100004

Hoffmann, D. E., & Tarzian, A. J. (2001). The girl who cried pain: A bias against women in the treatment of pain. *The Journal of Law, Medicine & Ethics*, 29(1), 13–27. https://doi.org/10.1111/j.1748-720X.2001.tb00037.x

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). *Mistral 7B*. arXiv. https://arxiv.org/abs/2310.06825

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). PubMedQA: A dataset for biomedical research question answering. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 2567–2577). https://arxiv.org/abs/1909.06146

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3), 535–547. https://arxiv.org/abs/1702.08734

Karavadra, B., Thorpe, G., Morris, E., & Semlyen, J. (2025). Exploring delay to diagnosis of endometriosis, a healthcare professional perspective. *BMC Health Services Research*, 25(1), 1483. https://doi.org/10.1186/s12913-025-13536-5

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report 8-75). Naval Technical Training Command.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P. A., Rouvier, M., & Dufour, R. (2024). *BioMistral: A collection of open-source pretrained large language models for medical domains*. arXiv. https://arxiv.org/abs/2402.10373

Li, W., Feng, H., & Ye, Q. (2025). Factors contributing to the delayed diagnosis of endometriosis—A systematic review and meta-analysis. *Frontiers in Medicine*, 12, 1576490. https://doi.org/10.3389/fmed.2025.1576490

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 3982–3992). https://arxiv.org/abs/1908.10084

Teede, H. J., Tay, C. T., Laven, J. J., Dokras, A., Moran, L. J., Piltonen, T. T., Costello, M. F., Boivin, J., Redman, L. M., Boyle, J. A., Norman, R. J., Mousa, A., & Joham, A. E. (2023). Recommendations from the 2023 international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *The Journal of Clinical Endocrinology & Metabolism*, 108(10), 2447–2469. https://doi.org/10.1210/clinem/dgad463

Appendix A: Video Presentation

Video URL: *https://www.youtube.com/watch?v=s1ZtTDrFvHY&t=175s*