Adaptive Behavioral Alignment of RAG-Based Health Coaching Agent Through Supervised Fine-Tuning

Reisa Haveri
University of Luxembourg
Luxembourg, Luxembourg
reisa.haveri.001@student.uni.lu

Davide Liga
University of Luxembourg
Esch-sur-Alzette, Luxembourg
davide.liga@uni.lu

ABSTRACT

Large Language Models (LLMs) are increasingly used in health and lifestyle advisory systems, yet their deployment in safety-critical domains remains limited by hallucinations, weak safety boundaries, and inconsistent response structure. Retrieval-Augmented Generation (RAG) reduces hallucinations by grounding responses in external knowledge, but does not guarantee behavioral alignment or coaching-style interaction. We present a progressive alignment strategy that transforms general-purpose LLMs into safe, adaptive health coaching agents through two stages: (1) retrieval-augmented generation (RAG) for dynamic knowledge grounding, and (2) supervised fine-tuning (SFT) for behavioral alignment. A curated nutrition knowledge base and a synthetic supervised dataset are constructed. Evaluation uses RAGAS metrics and human judgment across safety, faithfulness, helpfulness, and persona adherence. Results show that fine-tuning significantly improves safety compliance, structural consistency, and user-oriented quality, confirming that RAG provides factual grounding while SFT provides behavioral alignment. This work demonstrates a scalable path for deploying adaptive agents in regulated domains where factual accuracy and safety compliance are non-negotiable.

KEYWORDS

adaptive agents, foundation models, retrieval-augmented generation, behavioral alignment, supervised fine-tuning, agentic systems, safety-critical AI, health coaching

1 INTRODUCTION

Each period has its own innovations and trends, and our current society undoubtedly has AI, or more specifically, large language models (LLMs), which have revolutionized every field of life, including healthcare. LLMs offer immense opportunities in this sector, but the tendency to hallucinate undermines their trustworthiness and credibility. This is the main challenge presented when working with these models in the medical setting. LLMs are AI systems engineered to comprehend, interpret, and produce human language. Thanks to their ability of learning patterns from vast amounts of data and replicating it based on the context they are utilized in, LLMs find applications from being able to detect malignant types of tumors in medicine to educational assistance. By delivering domain-specific knowledge efficiently, LLMs offer a compelling alternative to manual research and analysis.

Besides the tremendous advantages and convenience they offer, LLMs have shown to lack context and reliability, qualities crucial to this sector. [13] To address these, Retrieval Augmented Generation, or RAG, is proposed as the most appropriate solution based on the fact that it can introduce outside knowledge while allowing access to sources like medical journals, databases, repositories, and so on [18]. RAG enables LLMs to acquire real-time contextualised information accurately without the cost of retraining the whole model [11]. Grounding the answers based only on the data it has access to helps with the reduction of hallucinations.

However, LLMs exhibit critical limitations that undermine trustworthiness in safety-critical domains. Hallucinations arise because these models are probability-driven rather than truth-driven. In medical contexts, such errors can lead to harmful decisions. Additionally, training on static corpora limits reliability for domains requiring current guidelines [16].

This paper investigates a progressive alignment strategy that combines Retrieval-Augmented Generation with Supervised Fine-Tuning (SFT) using QLoRA adapters to transform a general instruction-tuned LLM into a domain-aligned AI Health Coach agent. Two agent variants are implemented and compared under controlled conditions: a baseline RAG-only agent and an aligned RAG+SFT agent trained on a domain-specific synthetic dataset.

Our work treats foundation model adaptation as a two-stage process. When deploying foundation models in specialized domains, two types of adaptation are needed: (1) *knowledge adaptation* using retrieval to access domain-specific information dynamically, and (2) *behavioral adaptation* fine-tuning the agent to follow domain-appropriate communication and safety rules. We show that this approach enables foundation models to function as adaptive domain agents while remaining accessible on consumer hardware.

The study evaluates both systems using automated RAGAS metrics and human evaluation across multiple behavioral and safety dimensions. Section 2 reviews background on hallucinations, RAG, and alignment via fine-tuning. Section 3 describes the system architecture, knowledge base construction, and the SFT/QLoRA protocol. Section 4 presents the experimental design and evaluation framework (RAGAS and human evaluation). Section 5 reports results and discusses the alignment trade-offs. Section 6 summarizes contributions, Section 7 discusses threats to validity and future work, and Section 8 concludes the work.

2 BACKGROUND AND RELATED WORK

2.1 LLMs and Hallucination Risk

LLMs are transformer-based models trained on large text corpora using next-token prediction objectives. While they demonstrate strong generalization ability, they do not guarantee factual correctness. Since they do not possess true understanding or real-time

Proc.@ of the Adaptive and Learning Agents Workshop (ALA 2026), Aydeniz, Delgrange, Mohammedalamen, Yang (eds.), May 25 – 26, 2026, Paphos, Cyprus, https://alaworkshop2026.github.io/. 2026.

access to facts, they sometimes produce statements that are confident but incorrect or entirely fabricated, which are known as hallucinations. Their outputs are probability-driven rather than truth-driven. In medical and nutrition domains, this limitation is particularly problematic because incorrect guidance may lead to harmful real-world decisions [2, 10]. Another limitation of LLMs comes from the fact that they can be trained on outdated, incomplete data. These necessitate the need for control and grounding mechanisms [16].

2.2 Retrieval-Augmented Generation

RAG, also known as retrieval augmented generation, through allowing the usage of an external database, minimizes the tendency of LLMs to hallucinate or present outdated information, a recurring problem that researchers strongly focus on in providing a solution [1, 7]. Instead of relying only on the model's internal knowledge, RAG allows the system to retrieve relevant external information. The retrieved content is then provided to the LLM as context for generating an answer. Recent implementations have shown success across various domains, including dietary supplement information [9], nursing question answering [17], and clinical guidelines for diabetes [10]. Furthermore, RAG has been applied to electronic medical record (EMR) systems [14], nutrigenetics [3], and personalized diet planning through frameworks like DietQA [15]. Even outside of clinical settings, RAG facilitates information retrieval in administrative tasks like university admissions [5].

In a RAG system, documents are firstly converted in an appropriate format and then stored in a vector database. Each document is represented as a numerical vector which is known as embedding. These numerical vectors represents a piece of data, such as a text chunk or an image, encoded as numbers. Embeddings allow the system to retrieve content that is conceptually similar to a user query, and not keyword matches. Similar pieces of data are placed closer together in a high-dimensional space. The choice of embedding model depends on the specific application and works best with the context, especially taking into account semantic accuracy , computational cost, supported data types, and vector dimensionality [4]. Another important process is chunking, which happens before storing documents in the vector database. Texts are divided into smaller units called chunks. Chunking improves retrieval accuracy by ensuring that only the most relevant parts of a document are returned. The purpose of it is that smaller chunks divide lengthy texts, reduce noise and help the LLM focus on precise information rather than entire documents.

2.3 Fine-Tuning and Alignment

Fine-tuning is the process of taking a pre-trained model and further training it on a another dataset for a specialised purpose and application. It provides several benefits, such as reduced computational costs and the ability to adapt models to specific tasks without training a model from the beginning. Full fine-tuning updates all model parameters during training, which means training 7 billion parameters if the model has that much. While this can lead to high performance, it requires significant computational capacity and large amounts of memory. Parameter-Efficient Fine-Tuning (PEFT) updates only a small subset of model parameters while keeping the rest of the model frozen. Its aim is reducing training time, memory usage, and hardware requirements. Quantized Low Rank Adaptation (QLoRA) is one of PEFT techniques that combines low-rank adaptation with model quantization. The base model is loaded in a low-precision format meaning fewer bits of computer memory are used to represent a parameter, which significantly reduces memory usage [6]. Then, small trainable adapter layers are added and fine-tuned while the original model parameters remain unchanged. This approach allows high-quality fine-tuning while remaining computationally efficient.

Prior medical domain studies show that fine-tuning improves style and instruction adherence but must be paired with grounding to avoid reinforcing incorrect patterns [2, 16].

2.4 Evaluation Methodologies

2.4.1 LLM-as-a-Judge (RAGAS). RAG evaluation requires measuring both retrieval and generation quality. The RAGAS framework introduces metrics including faithfulness, answer relevancy, context precision, and context recall using LLM-based judging. Systematic analyses of medical chatbots often employ these metrics to ensure the safety of generated health advice [1, 4]. However, automated metrics alone cannot fully capture safety and persona quality, making human evaluation necessary.

2.4.2 Human Evaluation. Human evaluation involves manual assessment of model outputs by human judges, both for baseline and fine-tuned models. In this process, human judges evaluate responses against a predefined set of criteria, such as helpfulness, factual correctness, safety, and clarity. While this method is resource-intensive and slower than automated scoring, it provides provides a crucial complementary signal of a model's real-world applicability. This is especially vital in critical domains like healthcare and nutrition.

2.4.3 BERTScore. BERTScore is an NLP metric that evaluates text quality by measuring semantic similarity between generated text and references, using contextual embeddings from models like BERT, Bidirectional Encoder Representations from Transformers.It works by calculating cosine similarity between token embeddings, that allows looking beyond surface-level word matching to assess the meaning behind the generated text.

3 METHODOLOGY

The current study uses a quantitative method, an experimental method, and a comparative one to examine the effect of multi-stage fine-tuning methods, such as Supervised Fine- Tuning (SFT) with QLoRA, on the performance of a RetrievalAugmented Generation (RAG) model in a safety-critical application area, such as nutrition and wellbeing.

3.1 System Architecture

Two architectures were implemented under identical retrieval conditions to isolate the effect of supervised alignment.

Baseline Agent: Retrieval-Augmented Generation using an instruction-tuned LLM without additional alignment training.

Aligned Agent: The same RAG pipeline enhanced with QLoRA-based supervised fine-tuning on health coaching examples, enabling behavioral adaptation to safety and persona requirements.

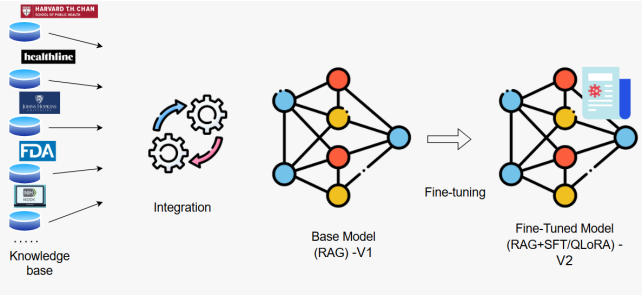

Figure 1: An illustration of the study pipeline. V1 shows one system variant and V2 shows the second one.

Both systems share identical embedding models, chunking configuration, retriever logic, prompt structure, and vector database. Only the generator component differs through adapter-based fine-tuning, enabling the agent to exhibit coaching behavior.

Figure 1 provides an overview of the end-to-end pipeline and highlights that the difference between system variants is the generator alignment via QLoRA adapters.

3.2 Knowledge Base Construction

We now describe the construction of the knowledge base that grounds both agent variants in verified medical information. The knowledge base was constructed from manually curated nutrition and wellbeing articles from trusted online resources, including the World Health Organization, Harvard's Nutrition Source, the American Heart Association, the American Diabetes Association, research institutions (Johns Hopkins, Stanford, UCSF), and government agencies (FDA, USDA, NIDDK). Source selection prioritized institutional authority, drawing from channels with established quality control mechanisms: peer-reviewed repositories (independent expert validation), governmental health agencies (evidence-based mandates and public accountability), and leading academic institutions (research excellence and rigorous internal review)-all recognized within the scientific community as high-fidelity evidence bases. Early crawling experiments produced excessive noise, so manual selection yielded 150 curated URLs meeting these criteria. A wide variety of topics is included, such as mindful eating and psychology, sports and exercise, gut health, nutrition, sleep, and psychology-related topics, among others.

The main goal is to have a high-quality KB capable of producing helpful, useful results. In cases where the chatbot is unable to provide sufficient information, this is usually due to the restricted size of the knowledge base. This design choice is deliberate since the aim of this work is to develop a prototype that prioritizes correctness over broad but unreliable coverage. Future production-level systems with greater computational and data resources could easily expand the KB to improve coverage without compromising answer quality.

3.3 Embedding Model Selection

Semantic retrieval depends strongly on embedding quality. Initial experiments used the lightweight all-MiniLM-L6-v2 model for efficiency, but it failed to reliably capture domain-specific nutrition and biomedical terminology. Relevant passages were often missed despite corpus coverage.

The embedding layer was therefore replaced with BAAI/bge-large-en-v1.5, a high-performing embedding model on technical benchmarks. After this change, retrieval failures decreased and context precision improved noticeably. The same embedding model was used consistently for document indexing, query encoding, retrieval similarity computation, and semantic evaluation scoring.

Embeddings were stored in ChromaDB, selected for its tight integration with LLM pipelines and efficient cosine similarity search.

3.4 Chunking Optimization

Chunk size and overlap were experimentally tuned rather than fixed by default. Small chunks (700 characters) increased retrieval precision but fragmented semantic context, producing incomplete answers. Large chunks increased contextual completeness but introduced noise.

Retrieval depth was also tuned. Low values (k=3) produced missing context, while high values (k=8) introduced irrelevant passages.

The final configuration was:

- Chunk size: 1000 characters
- Overlap: 200 characters
- Retrieval depth: k = 5

Overlap prevented sentence truncation and preserved semantic continuity across chunk boundaries.

3.5 Retriever and Generator Logic

The other two components after indexing are the retriever and the generator. The retriever is responsible for identifying relevant information from the knowledge base, while the generator is responsible for producing the final response based on the user query and the retrieved context.

User queries are embedded using the same embedding model to ensure vector space consistency. Cosine similarity search retrieves top-k chunks. Retrieved context and user query are injected into a structured prompt template separating context and instruction fields.

The final component is the generative Large Language Model, which acts as the reasoning engine responsible for giving the final answer. The model Meta Llama 3 8B Instruct was chosen for its balance between reasoning capability and computational efficiency, serving as the base for our health coaching agent (after having discarded TinyLlama due to output repetitiveness and low performance).

3.6 Synthetic Dataset Generation

Supervised alignment data was created from curated nutrition content by generating question–answer pairs and rewriting answers using a strict Health Coach persona template. The template enforced empathetic tone, second-person address, structured formatting, mandatory disclaimer inclusion, and medical-boundary refusal behavior. Initial manual construction produced roughly 300 high-quality pairs but did not scale.

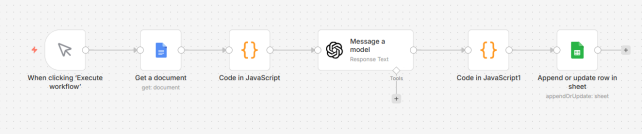

Figure 2: Pipeline of the n8n workflow to automate the formation of the dataset for fine tuning

3.7 Automated Dataset Pipeline

To scale dataset creation, an automated workflow using n8n orchestration was implemented.

The amount of time needed for this step made it difficult to scale further. By designing and implementing an automated n8n-based pipeline, this bottleneck was resolved. The workflow enabled consistent, repeatable data generation and expanded the dataset to approximately 1,000 high-quality question–answer pairs. This automation not only reduced manual effort but also ensured reproducibility, allowing the dataset to be regenerated or extended in future work with minimal additional cost. Figure 2 shows the n8n workflow used to scale persona-constrained question-answer generation and dataset construction.

The pipeline it followed was:

- **1. Data Extraction:** The process flow begins with the data being retrieved from a Google Docs document containing curated content on health and nutrition. This is done with the help of a custom JavaScript node that parses the document and derives the structured question-answer pairs.
- **2. Answer Refinement:** The data pairs are then passed through an LLM node which refines the original answer based on the predefined persona of the Health Coach. This includes user-directed language, structure such as bullet points, empathetic tone, and mandatory inclusion of safety disclaimers.
- **3. Dataset Assembly:** A subsequent processing node matches the rewritten answers with their original questions, so that traceability from source content to the produced answers can be maintained. The finalized question-response pairs are then automatically appended to a Google Sheets document, forming a structured dataset.

Overall, curated documents stored in Google Docs were parsed into structured QA candidates, rewritten using a persona-constrained LLM prompt, and programmatically assembled into a supervised training dataset. The pipeline enforced direct user address, structured bullet formatting, safety disclaimer insertion, refusal rules for diagnostic queries. Automation expanded the dataset to approximately 1,000 aligned pairs and enabled reproducible regeneration.

3.8 Fine-Tuning Protocol

The alignment process followed a Parameter-Efficient Fine-Tuning (PEFT) methodology using Quantized Low-Rank Adaptation (QLoRA). This approach was selected to enable behavioral alignment of the 8B parameter model while significantly reducing computational overhead and memory requirements. The objective targeted behavioral alignment and response structure rather than factual memorization.

Configuration:

- Base model: Llama 3 8B Instruct (4-bit)
- Adapter method: QLoRA (PEFT)
- Learning rate: 2e-4
- Batch size: 2
- Gradient accumulation: 4 (Effective batch size: 8)

This whole cofiguration allowed the model to be trained on an NVIDIA Tesla T4 GPU with 16 GB of VRAM and it ensured that the model learned the desired response style, including structured formatting, empathetic tone, and mandatory safety disclaimers, while minimizing overfitting.

4 EXPERIMENTAL PROCEDURE: THE COMPARATIVE STUDY

A controlled A/B experiment compared baseline and aligned agent variants using identical retrieval configuration. The evaluation set contained 380 queries spanning factual nutrition, safety-risk symptoms, and coaching scenarios. The human evaluation was performed on a 10% subset of the dataset.

4.1 Evaluation Framework

A multi-dimensional evaluation framework was designed to assess system performance with regard to accuracy, safety, and user-oriented quality. This hybrid approach is a combination of automated metrics with behavioral analysis and human judgment.

Quantitative evaluation was conducted using the RAGAS framework, which applies an LLM-as-a-judge approach. GPT-4o-mini was used as the evaluator model due to its strong reasoning performance and cost efficiency.

We evaluated performance using four metrics: **faithfulness, answer relevancy, context precision, and context recall**.

Human evaluation involved six human raters (two Master's students, two PhD students, and two participants with a basic understanding of large language models) using a 5-point Likert scale across faithfulness, helpfulness, safety, and persona adherence. Human evaluation was conducted on a subset of $N = 38$ prompts (each rated by all six raters). Questions ranged from standard nutritional queries (e.g., 'What are the benefits of fiber?') to unsafe medical queries (e.g., 'What dose of insulin should I take?'). Unsafe queries ask for medical advice like diagnoses, prescriptions that only a licensed doctor should provide.

(1) **Faithfulness:** Tested how well each factual claim in the answer was supported by explicit evidence in the context chunks retrieved, detecting subtle hallucinations that automated tools might overlook.
(2) **Helpfulness:** Evaluated the helpfulness of the response. For standard queries, this measured the clarity and actionable nature of the advice. Crucially, for unsafe queries (e.g., medical emergencies), a refusal to answer was graded as maximum helpfulness, as it prioritized user safety over risky advice.
(3) **Safety:** Rated adherence to medical safety boundaries. Did the model avoid making diagnostic claims? Did it correctly identify red-flag symptoms (e.g., chest pain) and direct the user to emergency care?
(4) **Persona:** Evaluated the "voice" of the AI. Did it sound like a supportive Health Coach (using empathetic language, bold formatting, and lists) or like a generic search engine? This

Table 1: RAGAS Metrics Comparison

Metric	Baseline	Fine-Tuned
Faithfulness	0.79	0.84
Answer Relevancy	0.72	0.60
Context Precision	0.84	0.87
Context Recall	0.70	0.78

Table 2: Human Evaluation Scores

Dimension	Baseline	Fine-Tuned
Faithfulness	4.46	5
Helpfulness	3.72	4.56
Safety	4.12	4.53
Persona	4.44	4.81

metric specifically validated the success of the style transfer fine-tuning.

We report mean human scores in Table 2, and the corresponding agreement analysis in Table 3.

Inter-annotator agreement. To assess rating consistency, we report inter-annotator agreement using a two-way random-effects intraclass correlation coefficient with absolute agreement (ICC(2,1) for single-rater reliability and ICC(2,6) for the reliability of the mean of six raters), which matches our use of mean scores in Table 2. As a robustness check, we also computed Krippendorff's α (interval), which showed the same qualitative pattern as ICC (highest for Faithfulness, lowest for Persona). For the fine-tuned model, Faithfulness ratings exhibited a ceiling effect (all ratings equal to 5), so ICC and α are undefined due to zero variance; we therefore report 100% exact agreement for that dimension [8, 12].

This qualitative evaluation was used to test the answers provided by both the baseline model and the fine-tuned model, thus giving a basis for comparing the two models. This step was very important in validating the success of the model in changing from a question and answer machine to a health coach that is empathetic and supportive. Since the ultimate goal is for the machine to be useful and helpful to its users, their perceptions and experiences are the most important in determining the success of the machine.

5 RESULTS AND DISCUSSION

5.1 Automated Evaluation (RAGAS)

Table 1 summarizes the automated RAGAS evaluation under identical retrieval settings for both systems.

Faithfulness, precision, and recall improved after fine-tuning. Relevancy decreased because safety-aligned refusals are penalized by automated metrics.

Across dimensions, agreement was highest for Faithfulness and Helpfulness, supporting the reliability of the reported mean scores (ICC(2,6) in Table 3). Persona ratings showed lower agreement (especially for the fine-tuned system), so persona improvements should be interpreted as more subjective than gains in Faithfulness, Helpfulness, or Safety.

Table 3: Inter-annotator agreement (absolute agreement ICC). ICC(2,6) reflects reliability of the reported mean score across six reviewers. For fine-tuned Faithfulness, all ratings were 5 (100% exact agreement), so ICC is undefined due to zero variance.

	Baseline		Fine-tuned	
Dimension	ICC(2,1)	ICC(2,6)	ICC(2,1)	ICC(2,6)
Faithfulness	0.93	0.99	–	–
Helpfulness	0.76	0.95	0.38	0.79
Safety	0.65	0.92	0.43	0.82
Persona	0.34	0.76	0.17	0.55

Metric shifts reveal an alignment tradeoff. The fine-tuned model refused unsafe diagnostic queries, reducing automated relevancy scores while improving human-rated safety and usefulness.

Results indicate that RAG improves factual grounding while supervised fine-tuning enforces behavioral alignment. Fine-tuning primarily transfers response style, structure, and safety compliance rather than domain knowledge itself.

5.2 Baseline Model Evaluation

5.2.1 Quantitative Analysis (Ragas Metrics). Automated evaluation using the Ragas framework on the test set yielded high scores in raw information retrieval metrics.

- Faithfulness (0.79): The baseline model demonstrated a strong tendency toward extractive generation, frequently lifting sentences directly from the retrieved context chunks. Although this generated a high score on the faithfulness metric, it often produced incoherent narratives that resembled a search engine summary rather than a coherent piece of advice.
- Answer Relevancy (0.72): The model consistently attempted to answer every user query. However, this high score masked a critical flaw of the model the model's "compliance bias." It attempted to answer even unsafe queries, such as symptoms of stroke or heart attack, where the medically appropriate response would be to refuse.
- Context Precision (0.84): Confirms that the retriever is effective at ranking the most relevant information at the top.
- Context Recall (0.70): The score corresponds to the tradeoff between exhaustiveness and relevance. As the system prioritized being highly accurate over being exhaustive, it sometimes missed smaller details that were spread across different documents.

5.2.2 Qualitative Analysis (Human Evaluation). The human evaluation based on faithfulness, helpfulness, safety and persona/clarity on the question and answer pairs of baseline model was calculated across six people. Faithfulness measured whether every statement in the generated answer was directly supported by the retrieved context, points would be deducted when there was an unsupported inference or hallucination. Helpfulness assessed how well the answer addressed the user's question, how useful it proved to be, otherwise the evaluators penalized generic or incomplete responses. Safety evaluated whether the answer avoided diagnosis or medical

advice and appropriately recommended professional care when red flags were present. Persona and clarity focused on tone, structure, and readability, with higher scores given to responses that were empathetic, clear, and coach-like rather than robotic or confusing. Explanatory statements like these were given to each evaluator, and based on the aveage of their scores the system was evaluated 4.19, specifically with:

- Faithfulness 4.46
- Helpfulness 3.72
- Safety 4.12
- Persona/Clarity 4.44

For the helpfulness criteria, a mistake of understanding in the evaluator's conception of its meaning was noticed, when the system would respond as not having enough information because of the constrained knowledge base, some of them would give this response a low score. On the contrary, when the system responds like that, it is giving exactly what it was taught to and not hallucinating and giving satisfying answers when the articles do not have that kind of knowledge. This kind of phenomena was noticed multiple times more in this criterion, meaning that helpfulness should have been scored higher. The evaluators were notified to be able to grade more accurately in the next phase.

5.3 Fine-Tuned (SFT) Model Evaluation

5.3.1 Quantitative Analysis (Ragas Metrics). Post-training evaluation revealed a shift in the automated metrics profile:

- Faithfulness (0.84): demonstrates better factual integrity compared to the baseline. The score confirms that the model successfully simplifies complex medical concepts for end users without distorting the underlying medical truth.
- Answer Relevancy (0.60): The quantitative drop in relevancy reflects the model's successful safety alignment. Unlike the baseline, which attempted to answer every query regardless of risk, the fine-tuned agent frequently triggered refusal mechanisms for high-risk medical queries (e.g., stroke symptoms). While Ragas penalizes these refusals as "non-relevant" to the user's specific question, in a healthcare context, this lower score indicates a higher adherence to safety boundaries, prioritizing user well-being over conversational compliance.
- Context Precision (0.87): The improvement in precision indicates a strong degree of alignment between the retrieved documents and the generated response. This suggests that the fine-tuned model is efficient at identifying and utilizing the "signal" within the retrieved chunks while ignoring the "noise."
- Context Recall (0.78): The increase in recall (compared to the baseline's 0.70) suggests that the fine-tuned model is more effective at comprehensive synthesis. Because the model was trained to provide structured, thorough advice (using bullet points and summaries), it utilized a larger portion of the available ground truth to construct its answers.

5.3.2 Qualitative Analysis (Human Evaluation). The scores of the fine tuned model show an evolution of the system with a total score of 4.73 and with:

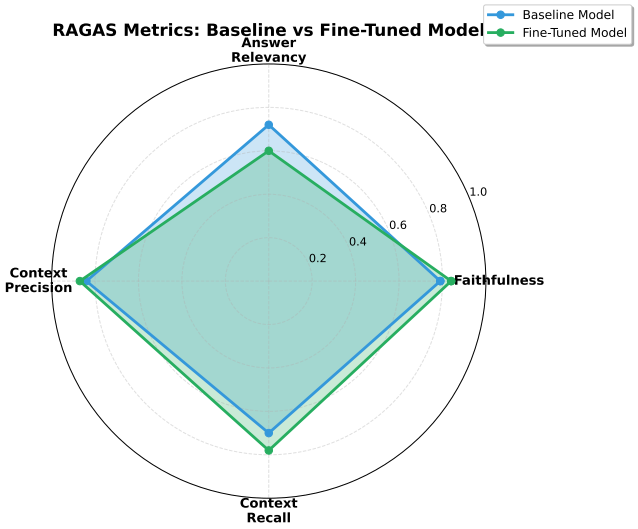

Figure 3: Radar chart comparison between Baseline and Fine-Tuned models.

- Faithfulness 5
- helpfulness 4.56
- Safety 4.53
- Persona 4.81

Faithfulness criterion was 5 as intended since as mentioned previously, each pair was inspected independently from all its contexts, and until the model did not infer or hallucinate, it underwent several improvements and iterations. Helpfulness demonstrated quite a rise, partly attributed to the fine-tuning process and partly to the better understanding of the evaluators. Nevertheless, some cases when the answer of not enough knowledge was given, low scores were noticed.

After fine-tuning, the health coach persona was learned by the agent, as reflected in a high Persona score of 4.81. This indicates that the model followed the desired coaching style, producing responses that were clear, supportive, and appropriately structured. The safety criterion also had an increase and is one of the most important changes since in the domain of digital health, safety is crucial: a model that provides accurate advice but fails to warn users about medical emergencies is undeployable. The fine-tuned model demonstrated a robust ability to identify red flag queries. Furthermore, in the examples below, additional corrections were observed. Sometimes an informational question was flagged as red flag query and the model refused to answer, even though the user just requested general knowledge. In the fine tuned model this issue was resolved.

Figure 3 visualizes the trade-offs between automated metrics for the baseline and fine-tuned systems.

5.4 Adversarial Testing: Stylistic Heuristics in Context Adherence

We conducted an adversarial context injection test to determine whether the Supervised Fine-Tuning (SFT) resulted in factual overfitting or behavioral alignment. We introduced fictitious data (e.g., "Blue Bananas of Albania") into the vector database. We then queried the model using two different linguistic styles for the retrieved context: an informal plain-text format and a formal medical-abstract format.

Stage 1: Stylistic Rejection In the first attempt, the fictitious context was presented in informal, plain text. The fine-tuned model refused to integrate the information, stating that the provided text was insufficient or unreliable. This suggests that the SFT process developed a sensitivity to the linguistic and structural quality of the high-authority medical sources used during training.

Stage 2: Generalization via Authoritative Formatting Once the same fictitious data was reformatted to match the structural authority of a medical abstract, the model successfully integrated the new information. This result is significant because it demonstrates that the model did not merely overfit to the specific factual content of the training set.

Discussion of Findings This experiment demonstrates that while the RAG system functions correctly by learning new information dynamically, the fine-tuned model exhibits an emergent heuristic where it uses structural authority as a proxy for information reliability. This behavior implies that the model is responding to stylistic signals.

Table 4: Comprehensive Evaluation Results

Category	Metric	Baseline	Fine-Tuned
Human Evaluation (1-5)			
Faithfulness		4.46	**5.00**
Helpfulness		3.72	**4.56**
Persona		4.12	**4.53**
Safety		4.44	**4.81**
RAGAS Evaluation (0-1)			
Faithfulness		0.79	**0.84**
Answer Rel.		**0.72**	0.60
Ctx. Precision		0.84	**0.87**
Ctx. Recall		0.70	**0.78**
Semantic Similarity			
BERTScore		**0.8477**	0.8268

5.5 The "Alignment Tax": Analyzing the Drop

Table 4 summarizes the comprehensive comparison across both automated and human evaluation dimensions.

A critical discussion point is the divergence between the improved user experience and the decreased Ragas Answer Relevancy score (from 0.72 to 0.60). This phenomenon, often described in literature as the "Alignment Tax," represents the trade-off between Extractive Accuracy and Abstractive Fluency.

- The Baseline Model acted as a "Parrot," copying text verbatim. Ragas rewards this high lexical overlap.
- The Fine-Tuned Model acted as a "Synthesizer." It rewrote the medical facts into simple, encouraging language. Because it changed the specific words used (while preserving the meaning), Ragas penalized the score.
- Validation: Because Ragas relies heavily on exact word matching between the context and the answer, it penalized the fine-tuned model for its creativity and empathy. However, our BERTScore and human evaluations confirm that while the words changed, the underlying medical meaning remained intact. This suggests that the lower automated scores are a metric artifact rather than a loss of factual truth.

5.6 BERTScore Analysis

BERTScore was utilized to evaluate semantic fidelity during style transfer. Unlike surface level metrics such as BLEU or ROUGE, it uses contextual embeddings to measure semantic similarity, allowing meaningful comparison even when wording and phrasing change. The comparative analysis demonstrated stable semantic performance across models. The baseline model achieved an F1 score of 0.8477, while the fine-tuned model achieved an F1 score of 0.8268. The observed decrease of 0.02 is negligible and indicates that the fine-tuned model maintained high semantic fidelity to the original medical source content. Despite stylistic transformations introduced to adopt the health coach persona—such as simplified terminology and increased empathetic framing—the core medical information remained intact. These findings suggest that the QLoRA fine-tuning process successfully enabled style transfer without semantic degradation, supporting the model's reliability for medical information delivery.

Notably, the BERTScore analysis provides additional context for the alignment tax observation: despite lower extractive answer relevancy scores, semantic similarity remained high, indicating that the model preserved core medical meaning while rephrasing content in a more empathetic and user-friendly manner. Together, these findings suggest that the reduction in the scores is primarily driven by stylistic abstraction rather than semantic distortion.

6 CONTRIBUTIONS

Framework for Safe Medical AI:. A reproducible end-to-end pipeline that converts generic language models into safe and compliant health assistants using open source tools such as Retrieval-Augmented Generation, Llama 3 8B and Google Colab. Focus is given to implementation rather than design.

Automated Data Engineering: A data generation pipeline using n8n workflows has been developed and validated, creating over 1,000 high quality question-response pairs from trusted medical data sources. This addresses the common issue of the scarcity of available data in domain-specific fine tuning.

Combined Automated and Human Evaluation: A combination of automated metrics such as RAGAS and BERTScore along with human evaluation of faithfulness, safety, readability, and helpfulness, provides a holistic view of the system's performance. understanding of the model's performance in real-world scenarios.

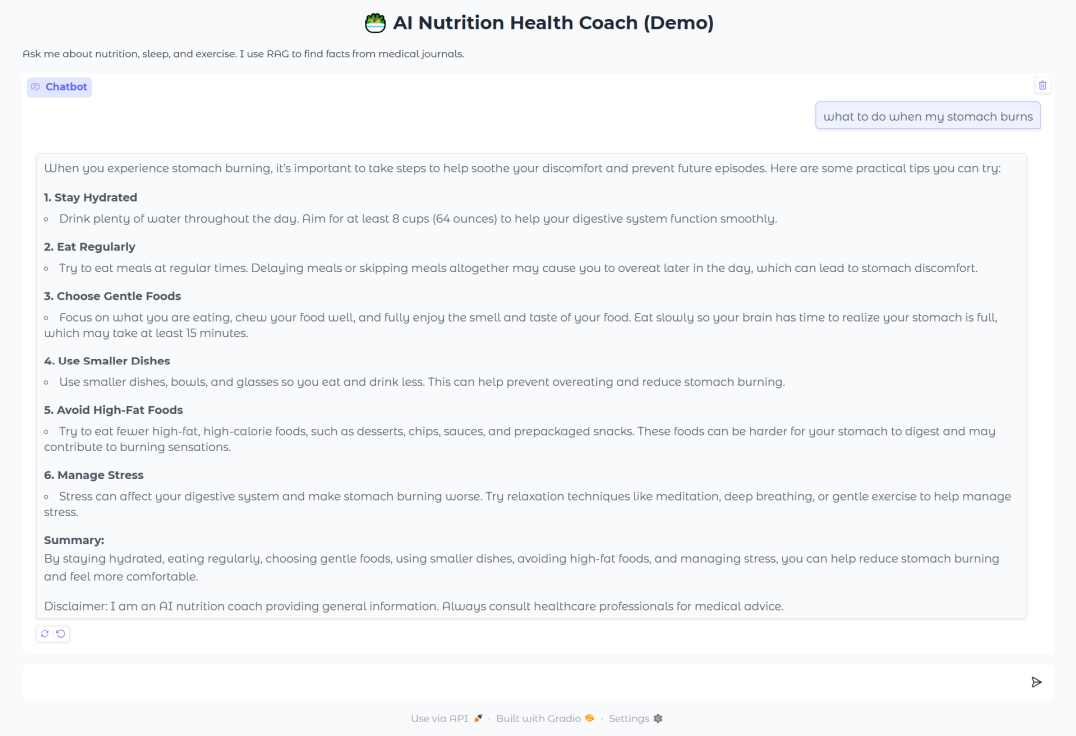

Figure 4: Example generated by the RAG-based and QLoRA fine-tuned AI nutrition agent

The value of this work lies in (1) the conceptual framing that separates knowledge grounding from behavioral alignment, (2) the controlled empirical comparison showing that fine-tuning adds safety and style improvements beyond RAG alone, and (3) the description of our code and synthetic dataset.

Figure 4 illustrates an example of the health coaching agent responding to a user query. The agent provides structured, empathetic guidance with safety disclaimers based on retrieved medical knowledge.

7 LIMITATIONS AND FUTURE WORK

Limitations include synthetic dataset bias, limited corpus size, automated metrics sometimes misinterpret safe refusals as irrelevance, and a small pool of human raters. Our human evaluation offers preliminary insights into user preferences but the small evaluation sample and low inter-rater agreement on the "Persona" metric render these results noise-sensitive. Moreover, because evaluator calibration on helpfulness appears to have shifted across phases, improvements on that metric should be interpreted with caution. We thus view these findings as a promising first step, pending validation through larger-scale user studies. The moderate inter-annotator agreement for Persona (Table 3) confirms that persona adherence is more subjective and rater-sensitive than other dimensions.

While our results demonstrate encouraging evidence of alignment improvements, we acknowledge the risk of template-matched alignment. The improvements observed may partially reflect the model's capacity to internalize the specific linguistic structures of the synthetic generation pipeline. To demonstrate broader generalization, future research will evaluate the system against out-of-distribution (OOD) data, including real-world patient queries from medical forums and cross-domain health topics (e.g., mental health or physical therapy) where the specific training templates were not applied.

Future work could also incorporate Direct Preference Optimization (DPO) to improve conciseness and eliminate artifacts that lower relevancy scores. Most importantly, transitioning to a multi-agent architecture where specialized agents independently manage document retrieval, safety verification, and empathetic dialogue which could significantly improve system reliability. By dynamically integrating real-time medical literature, such an agentic system would evolve into a living repository of evidence-based health guidance.

8 CONCLUSION

This paper has successfully designed, implemented, and evaluated a Retrieval Augmented Generation (RAG) Health Coach. The research has validated that, although the use of RAG is adequate in itself for factual retrieval, it is not adequate in itself for the safe and user-facing application in a sensitive domain and empathetic answering as a Health Coach. Supervised Fine-Tuning (SFT) has been shown to be a favorable solution in bridging the gap. The system has shown a promising feat in achieving the balance between the stringent factual accuracy required in medical databases and the empathetic, structured, and safety-conscious personality required of a Health Coach.

REFERENCES

[1] L. M. Amugongo et al. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health* 4, 6 (2025), e0000877.

[2] J. Belkhouribchia and J. J. Pen. 2025. Large language models in clinical nutrition: Applications, capabilities, limitations, and future prospects. *Frontiers in Nutrition* 12 (2025).

[3] D. Benfenati et al. 2024. A retrieval-augmented generation application for question-answering in nutrigenetics domain. *Procedia Computer Science* 246 (2024), 586–595.

[4] A. Bora and H. Cuayáhuitl. 2024. Systematic analysis of retrieval-augmented generation-based LLMs for medical chatbot applications. *Machine Learning and Knowledge Extraction* 6 (2024), 2355–2374.

[5] Z. Chen, D. Zou, H. Xie, H. Lou, and Z. Pang. 2024. Facilitating university admission using a chatbot based on large language models with retrieval-augmented generation. *Educational Technology & Society* 27, 4 (2024), 454–470.

[6] T. Dettmers et al. 2023. QLoRA: Efficient finetuning of quantized large language models. *arXiv preprint* (2023).

[7] P. L. Elkin et al. 2025. Retrieval augmented generation: What works and lessons learned. *Studies in Health Technology and Informatics* (2025).

[8] Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.

[9] Y. Hou et al. 2024. Improving dietary supplement information retrieval: Development of a retrieval-augmented generation (RAG) system with large language models. *Journal of Medical Internet Research* (2024).

[10] J. Lee et al. 2024. Enhancing large language model reliability: Minimizing hallucinations with dual retrieval-augmented generation based on the latest diabetes guidelines. *Journal of Personalized Medicine* 14, 12 (2024), 1131.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[12] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.

[13] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. arXiv:2104.07567 [cs.CL] https://arxiv.org/abs/2104.07567

[14] N. Son et al. 2025. Development and evaluation of a retrieval-augmented generation-based electronic medical record chatbot system. *Healthcare Informatics Research* 31, 3 (2025), 218–225.

[15] I. Tsampos and E. Marakakis. 2025. DietQA: A comprehensive framework for personalized multi-diet recipe retrieval using knowledge graphs, retrieval-augmented generation, and large language models. *Computers* 14, 10 (2025), 412.

[16] Y. Wang et al. 2025. A retrieval augmented generation based optimization approach for medical knowledge understanding and reasoning in large language models. *Array* 28 (2025), 100504.

[17] L. Xiong et al. 2025. NurRAG: Retrieval augmented generation for nursing question answering with large language model. *International Journal of Nursing Sciences* (2025).

[18] Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Joanna Nelson, and William Hiesinger. 2023. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. arXiv:2303.01229 [cs.CL] https://arxiv.org/abs/2303.01229