Navigating Nutritional Nuance:

A PICO-Based Approach to Validating Nutritional Health Claims Using Retrieval-Augmented Generation

by

Budi Setiawan Han

to obtain the degree of Master of Science in Computer Science at the Delft University of Technology,
to be defended publicly on Wednesday 6 November 2024 at 13:00.

Student Number:	4389336
Project Duration:	February 2024 – November 2024
Faculty:	Electrical Engineering, Mathematics and Computer Science (EEMCS)
Department:	Interactive Intelligence
Thesis Advisor:	Assoc. Prof. P.K. Murukannaiah
Daily Supervisor:	Assoc. Prof. L.P.A. Simons
Thesis Committe:	Assoc. Prof. P.K. Murukannaiah
	Assoc. Prof. L.P.A. Simons
	Assoc. Prof. M.S. Pera

TU Delft

Navigating Nutritional Nuance: A PICO-Based Approach to Validating Nutritional Health Claims Using Retrieval-Augmented Generation

Budi Han , Luuk Simons , Pradeep K. Murukannaiah
TU Delft
B.S.Han@student.tudelft.nl, L.P.A.Simons@tudelft.nl, P.K.Murukannaiah@tudelft.nl,

Abstract

Evidence-based lifestyle practices are effective in preventing and treating cardiovascular disease. However, the growing body of scientific literature and the prevalence of conflicting studies makes it challenging for healthcare practitioners to stay informed. Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), offer potential for automated fact-checking, where much work has been done in areas like politics, limited research has explored their application to nutritional health claims, which are more nuanced and demand rigorous evaluation of interventional studies for scientific validation. To fill this gap, this study investigates how effectively a RAG-based LLM can verify nuanced nutritional health claims. We develop a five-module framework, introducing an inclusion criteria-based approach and **SMaPS** (**S**equential **Ma**pping of **P**ICO-based **S**ynthesis) to enhance literature selection and evidence synthesis. Our findings indicate that while our Advanced RAG-LLM model shows potential in verifying nuanced health claims, it still faces significant limitations in accuracy. Although the inclusion criteria-based filter and SMaPS approach help balance predictions, the model often defaults to neutral outcomes when evidence is unclear. The problem of overgeneralization, the inclusion of irrelevant studies, and the difficulty of synthesizing precise numerical data undermines the model's reliability to verify nuanced health claims.

1 Introduction

1.1 Motivation

Cardiovascular disease (CVD)—encompassing a range of conditions affecting the heart and blood vessels (Badimon et al., 2019; Gaidai et al., 2023)—is a leading cause of global mortality. Hypertension, or high blood pressure, is a significant risk factor for CVD. Prolonged hypertension damages the arteries and increases the risk of heart attack and stroke.

Health self-management (HSM), i.e., incorporating healthy lifestyle practices, is effective in preventing and treating CVD (Dineen-Griffin et al., 2019; Grady and Gough, 2014). It involves individuals to take responsibility for their own health by utilising practices to prevent and reduce health risks. HSM support is the systematic provision of such practices. It includes healthcare staff providing education and interventions to a patient to enhance the patient's ability to manage their health.

Recent literature highlights the importance of HSM in improving cardiovascular health and reducing healthcare costs. A systematic review (Dineen-Griffin et al., 2019) examines HSM support strategies in primary health care practices and shows that such interventions can lead to improvements in clinical and humanistic outcomes for a variety of diseases. Further, a scoping review (Qama et al., 2022) discusses factors influencing the integration of HSM in daily life routines for chronic conditions, emphasizing the challenges and opportunities for implementing HSM practices. The literature stresses the importance of incorporating evidence-based HSM practices into routine care to promote engagement and optimization of health outcomes (Dineen-Griffin et al., 2019; Grady and Gough, 2014; Adams et al., 2004).

1.2 Problem

The literature on HSM and cardiovascular health has been increasing significantly (Qama et al., 2022). This information overload hinders timely access to insights useful in HSM support. Addressing this issue is crucial to enable healthcare practitioners and researchers to efficiently keep pace with the latest advancements, thereby enhancing HSM support. Moreover, there is a belief among the general public that nutritional science is often contradictory (Nagler, 2014; Armitage, 2019; Belluz, 2016; Tucci, 2021). The presence of contradictory findings can cause confusion among the general public, foster misinformation, and lead to uncertainty regarding sound dietary advice. Studies have examined the adverse outcomes associated with media exposure to contradictory claims and found that nutrition confusion was positively associated with nutrition backlash. Nutrition backlash decreased engagement in fruit and vegetable consumption (Lee et al., 2018). An example of disagreement is the impact of egg consumption on cardiovascular health. Some studies link eggs to higher cholesterol and heart disease risk (Sugano and Matsuoka, 2021), while others argue that egg cholesterol has little effect on blood pressure and highlight their nutritional benefits,

including high-quality protein (Myers and Ruxton, 2023).

Verifying health claims is a complex and nuanced process, requiring rigorous scientific investigation due to biological variability and external influences like diet and environment. Unlike political claims, health claims demand experimental design, statistical validation, and replication across diverse populations. Randomized controlled trials (RCTs) are crucial in minimizing bias and the thorough evaluation of study design, population diversity, and statistically significant outcomes are essential. Only by synthesizing evidence from multiple high-quality trials can health claims be accurately and reliably verified.

In navigating the evolving landscape of HSM and nutritional health claims, it is evident that the proliferation of cardiovascular health literature; the prevalence of conflicting studies; and the inherent nuance of nutritional health claims pose a significant challenge. To effectively address these issues, it is important for healthcare practitioners, researchers, and the public to stay abreast of the expanding literature while critically discerning and acknowledging contradictory findings.

1.3 Addressing the problem

To address these challenges, we propose a system that systematically retrieves and synthesizes the expanding body of literature while critically evaluating and filtering the studies needed to verify nutritional health claims. By providing evidence-based verification from carefully selected studies, we aim to bridge the gap in tackling the nuances of nutritional health claims and the prevalence of conflicting research. This approach enhances clarity and trust, aiding healthcare practitioners and the public in making informed HSM decisions while reducing confusion from conflicting nutritional findings.

Limited research has explored the verification of nutritional health claims. Most traditional claim verification studies focus on political claims (Guo et al., 2022) and some have touched on public health and COVID claims (Pradeep et al., 2021; Liu et al., 2024a; Kotonya and Toni, 2020). However, these claims differ significantly from the nutritional health claims we want to address in this study. Furthermore, the recent application of large language models (LLM) and retrieval augmented generation (RAG) in claim verification has shown promise, as demonstrated by Tan et al. (2023) that used LLMs for reasoned decision-making, and Liu et al. (2024a) who integrated RAG to further enhance LLMs by continuously drawing relevant insights from scientific literature. However, existing research is yet to apply LLMs and RAG to the verification of nuanced nutritional health claims, which require a more rigorous evaluation of high-quality evidence.

To this end, we propose leveraging LLMs and RAG frameworks. LLMs are advanced AI systems capable of processing and generating human language, allowing them to synthesize vast amounts of literature and extract relevant insights on specific nutritional health claim. RAG frameworks, on the other hand, combine LLMs with information retrieval systems to retrieve the most up-to-date and relevant data from trusted sources. Additionally, we incorporate the PICO framework into the RAG pipeline to address the nuances of nutritional health claims. PICO (Population, Intervention, Comparison, Outcome) is a widely used method for structuring clinical research questions and selecting high-quality evidence. By applying this framework, the system ensures that the retrieved studies are relevant and rigorously evaluated, helping to filter out lower-quality or irrelevant research and providing more precise verification of nutritional health claims. Together, these technologies and frameworks create a dynamic, evidence-based tool that is able to retrieve and synthesize relevant information to produce concise and trustworthy verification of nutritional health claims.

1.4 Research question

We explore the capabilities of LLM and RAG in verifying nutritional health claims to ultimately improve HSM support. Our research question is as follows: **How effectively can a RAG-based LLM verify nuanced nutritional health claims?**

We propose a RAG-LLM model that consists of five modules: Document collection, document retrieval, selection, summary generation, and explainable verdict generation module (Figure 1). We enhance the selection module with an inclusion criteria-based filter that utilizes the PICO framework to improve the relevance and quality of the literature used. Additionally, we incorporate a summary module that synthesizes the evidence required to verify claims. These enhancements ensure that health claims are validated by credible sources and provide clear, explainable verdicts. Finally, we curate a health claim database for evaluating the effectiveness and accuracy of our model.

The main research question is further distilled in the following three sub-questions:

1. How accurately does the PICO enhanced RAG model retrieve relevant scientific literature for health claims?
2. How effectively does the model synthesize retrieved literature to generate concise information on health claims?
3. How accurately can a RAG-LLM verify health claims when compared to expert annotations?

1.5 Contributions

This study addresses the gap in verifying nuanced nutritional health claims by introducing an inclusion criteria (IC) filter and incorporate the PICO framework in both the selection and summary module. While prior research has primarily focused on political and general public health claims, this work pioneers the use of an IC approach, leveraging the PICO framework to validate the relevance and quality of retrieved literature. By enhancing the RAG-LLM, the study establishes a framework grounded in high-quality evidence for verifying complex health claims. Additionally, the creation of a curated health claim database offers a valuable resource for evaluating the framework's accuracy, providing a novel method for automating the verification on nutritional health claims.

2 Background

In this section, we provide background information on the nuances of nutritional health claims. Then, we argue that Large

Language Models (LLMs) are an effective tool for informational support for verifying health claims. Finally, we introduce Retrieval Augmented Generation (RAG) to mitigate the shortcomings of LLMs.

2.1 Nutritional Health Claims

Nutritional health claims are any statement about a relationship between food and health. The European Food Safety Authority (EFSA) identifies three types of health claims (European-Commission, 2024; EFSA, 2023):

1. **Function Health Claims**: Food that can help reinforce the body's natural defence or enhance learning ability, e.g. "Milk may help improve bone density".
2. **Risk Reduction Claims**: Claims on reducing a risk factor in the development of a disease, e.g. "Nuts have shown to reduce cholesterol levels, a risk factor in the development of coronary heart disease"
3. **Claims referring to children's development**: For example: "Vitamin D is needed for the normal growth and development of bone in children.

In this study, we focus on function and risk reduction health claims. Specifically we focus on claims that "lower blood pressure" and "improve cardiovascular health". Additionally, we find a nutritional health claim relevant if such a claim focuses on widely accessible foods or components, is backed by credible scientific evidence, and addresses achievable health benefits through normal consumption. Such claims should address a meaningful health benefit that can be realistically achieved through normal dietary intake. For example, *"Berries can lower blood pressure"* could be considered a risk reduction claim.

Nuance of health claims

Verifying health claims is more complex than verifying other claims such as political claims. Political assertions often hinge on factual data, historical records, or observable policies that can be more straightforwardly confirmed or refuted. In contrast, health claims require rigorous scientific investigation that must consider and observe long-term biological processes, subtle interactions within the body, and the influence of external factors like diet and environment. As such the process of verifying health claims demands a more nuanced approach, involving experimental design, statistical validation, and replication of results across diverse populations. This makes the verification of health claims a more nuanced and scientifically demanding process.

The interpretation of health claims is inherently complex due to a multitude of factors that introduce variability and potential bias into research outcomes. Biological variability among individuals is one primary reason for divergent opinions on health claims. Factors such as genetics, age, sex, ethnicity, and pre-existing health conditions can influence how one responds to specific nutrients. this variability complicates the ability to make universally applicable claims about the health impacts of certain foods.

Approaching the verification of nuanced health claims thus requires a methodical and evidence-based process. It starts with the synthesis of scientific experiments, ideally through interventional studies such as randomized controlled trials (RCTs), which are the gold standard in scientific research (Hariton and Locascio, 2018). RCTs help minimize bias by randomly assigning participants to intervention or control groups, allowing researchers to more reliably isolate the effects of specific health interventions. Moreover, verifying health claims goes beyond just aggregating results; it requires thorough evaluation of criteria such as study design, population, intervention, comparison, and outcome. This involves evaluating the study's population to ensure it represents a diverse cross-section of individuals, examining the intervention to determine if it is relevant to the claim, and carefully comparing it to control and intervention groups. The outcome must also be clearly defined and statistically significant. As such, health claims can only be rigorously verified by carefully evaluating each study on its merits and synthesizing evidence from multiple high-quality trials, ensuring they are scientifically valid.

2.2 Large Language Models

Recent advances in artificial intelligence (AI), specifically, natural language processing (NLP) technologies, created new possibilities for support in healthcare and medical education (Eysenbach, 2023; Yu et al., 2023; Jo et al., 2023; Moons and Van Bulck, 2023; Kung et al., 2023). In particular, Large Language Models have been proposed as effective tools for scaling informational support around health.

Large language models (LLMs) are deep neural networks that utilise a transformer-based architecture (Zheng et al., 2023). This architecture employs self-attention mechanisms to capture relationships between words across long distances in text. Transformer models form the foundation of LLMs due to their ability to process long sequences of text in parallel and capture complex relationships between words across contexts (Vaswani et al., 2023). The self-attention mechanism allows LLMs to weigh the importance of different parts of the input when generating each word of the output, enabling coherent and contextually relevant text across extended passages. LLMs are typically created through unsupervised pretraining on diverse corpora, followed by fine-tuning of specific tasks. LLMs are trained on vast amounts of text data to predict the next word in a sequence. It involves feeding the model billions of tokens ("word" or "sequence-of-words") and adjusting its parameters to minimise prediction errors. As the model grows in size, from millions to hundreds of billions of parameters, they demonstrate emergent abilities, including few-shot learning and task generalisation.

LLMs brought breakthroughs in open-domain dialog systems which can perform free-form conversations on open-ended topics with the goal of providing information and also empathy (Liu et al., 2024b; Zhang et al., 2024). LLMs have also shown impressive capabilities in a variety of natural language tasks such as summarization, dialogue generation and question-answering. Such systems can be beneficial for public health and have the potential to serve as research assistance in the increasing domain of hypertension and diet. LLMs can assist clinicians in keeping up to date with the latest research by summarising articles, papers and other sources of information in a concise and easy to understand format.

However, LLMs face a significant challenge known as "hallucination" (Huang et al., 2023). This occurs when LLMs generate content that is inaccurate, fabricated, or inconsistent with provided information. Hallucinations can range from small errors to completely false statements, causing risk of information and distrust in AI systems. The problem of hallucination stems from various factors, including limitations in training data, imperfections in learning processes, and the probabilistic nature of language generation. Research is actively focused on mitigation strategies, such as improving data quality, implementing fact-checking mechanisms, and using retrieval augmented generation techniques (Xu et al., 2024).

2.3 Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) combines strengths of LLMs with external knowledge retrieval to improve accuracy and produce reliable responses (Gao et al., 2024; Wu et al., 2024). By retrieving relevant information from external sources, RAG helps to mitigate hallucinations. This approach grounds the LLMs output in factual data, thus reducing the likelihood of generating false or outdated information.

RAG allows LLMs to access current and domain-specific knowledge beyond just their original training data allowing for contextually relevant responses. This domain-specific knowledge base is a crucial component in RAG systems. This repository may contain domain-specific documents, data, and information sources that are unique to the organization's needs and expertise. By incorporating a custom knowledge base, RAG systems can access up-to-date, relevant, and often confidential information that may not be available publicly.

RAG often outperforms fine-tuning in scenarios requiring up-to-date information, cost-efficiency, and scalability (Gao et al., 2024). Unlike fine-tuning, which demands significant computational resources and retraining to incorporate new data, RAG can easily integrate current information by updating its external knowledge base. This approach allows for quick adaptation to changing contexts without altering the base model, making it more flexible and cost-effective. Moreover, it preserves the model's general knowledge while augmenting it with specific information, avoiding the potential loss of broad capabilities that can occur with fine-tuning. While fine-tuning remains valuable for deep domain specialization or altering core model behavior, RAG offers a more agile, transparent, and easily deployable solution for many applications, particularly those requiring frequent updates or access to diverse, current information.

3 Related Work

3.1 Traditional Claim-Verification

The task of claim-verification is studied under the umbrella of automated fact-checking (Guo et al., 2022). It is the task of automatically verifying the authenticity of claims based on the retrieval of evidence. During this task, each result may provide limited evidence towards a claim and contradictory evidence is prevalent. Consequently, positing this task as claim verification rather than fact-checking casts to goal as identifying evidence to both support and refute the claim.

Previous works follow a conventional framework of claim-verification that consist of three modules, the retrieval of relevant documents given a claim, sorting the evidence in each document, and predict a label based on the top k evidence (Wadden et al., 2020; Pradeep et al., 2021; Soleimani et al., 2020). Wadden et al. (2020) introduced scientific claim verification with the three-step approach of "Document Retrieval", "Sentence Selection", and "Textual Entailment". Notably, the first module used text similarity to retrieve the most relevant abstract without levering language embeddings. Pradeep et al. (2021) applied a similar three-step framework but leveraged the power of the T5 Language model across the three modules. Soleimani et al. (2020) used two BERT models, one for retrieving evidence from Wikipedia pages, and another for verifying claims. These studies mainly focus on database-centric scenarios that verify claims in a closed-domain setting. That is, evidence is retrieved from the relevant database prepared in advance.

3.2 Retrieval Augmented Claim Verification

Liu et al. (2024a) used the traditional three-step approach of claim verification but focused on an augmented retrieval module to specifically focus on RCT studies as evidence base for a given COVID claim. Instead of retrieving evidence from a prepared database, this augmented retrieval module presented a real life scientific use case where claims would be queried against hundreds or thousands of documents. The use of retrieval-augmented methodologies harnesses the capabilities of multiple information retrieval techniques such as document vectorization, semantic similarity-based retrievers, and similarity ranking mechanisms. The use of these techniques offered an efficient and accurate search through vast datasets of RCT studies, improving the identification of the most relevant evidence.

3.3 Open-Domain Claim Verification with LLMs

Tan et al. (2023) introduced an open-domain explainable fact-checking (OE-Fact) system, which utilizes LLMs to validate claims and provide casual explanations for claim verification decisions. OE-Fact adapts the traditional three-module framework to the open domain. Firstly, retrieve evidence from open websites. Here they used the Google search API to retrieve timely information from multiple sources and prioritize relevant and accurate information. They expand their search scope by submitting additional critical words as queries to the search engine. This ensures a broad claim-related candidate evidence retrieval. Secondly, a claim-relevant evidence selection module filters out noise by employing LLM and semantic similarity calculations sequentially. The LLM-based evidence filtering uses a 1-shot prompt to return the most relevant evidence and then a BERT-based similarity calculation is used to select the top k evidence. Finally, a verdict module uses an LLM to verify a claim with a label and generating a real-time explanation. They simulate reasoning ability of the LLM through a 1-shot prompt that analyzes the casual relationship between evidence and claim.

Experimental results show that OE-fact system outperforms traditional claim verification system in both closed- and open-domain settings. Additionally, the system improves

the reliability of claim verdicts by generating concise and transparent real-time explanations. Notably, this is the first work that utilizes an LLM in an open-domain setting, filling the gap in real-world claim verification scenarios.

3.4 Fact-Checking Public Health Claims

Kotonya and Toni (2020) introduced PUBHEALTH, a comprehensive dataset for automated fact-checking of public health claims. This dataset contains a public health claim, a veracity label (true, false, unproven, or mixture), and an explanation text. The researchers evaluated state-of-the-art pretrained transformer models by fine-tuning them on the PUBHEALTH dataset. While this paper introduced a dataset for public health claims, it is important to note that public health claims differ significantly from nutritional health claims. In the PUBHEALTH dataset, claims can range from topics like ObamaCare to questions such as "Expired boxes of cake and pancake mix are dangerously toxic." In contrast, nutritional health claims focus on the risks and benefits of specific foods in relation to factors like cardiovascular health and hypertension. As stated before, nutritional health claims are more nuanced than public health claims.

3.5 Summary and Research Gap

Claim verification has been widely studied, with traditional models focusing on political and public health claims through a three-step process of document retrieval, evidence selection, and claim validation (Guo et al., 2022; Pradeep et al., 2021; Wadden et al., 2020; Soleimani et al., 2020). Recent work, such as Liu et al. (2024a) and Tan et al. (2023), have introduced RAG models and open-domain LLM-based systems that improve accuracy and transparency in claim verification.

Despite the progress in claim verification using LLMs and RAG for public health claims (Kotonya and Toni, 2020), there has been limited research on applying these technologies to verify nuanced nutritional health claims, which require more rigorous evaluation of high-quality evidence. To address this gap, we propose a model that integrates the PICO framework into the RAG and the information synthesis pipeline. This approach allows for the retrieval of relevant, high-quality studies and ensures a more precise verification process that addresses the complexity and nuances specific to nutritional health claims.

4 Health Claim Dataset

In this section, we explain how we curate an annotated nutritional health claim dataset using the PICO-framework.

4.1 PICO Framework

PICO for Formulating Health Claims

We formulate concise claims using the PICO framework (Population, Intervention, Comparison, and Outcome) (Richardson et al., 1995). This framework is commonly used to formulate good clinical research questions, which can be utilized to formulate clinical claims by adapting the elements to suit the nature of the claim being made (Huang et al., 2006). For example Liu et al. (2024a) used the PICO framework to construct a Covid Verification dataset comprising of 15 PICO-encoded drug claims.

In the field of nutritional sciences, PICO elements can be reduced to a health claim. An example of a formulated health claim using the PICO framework could be:

- **Population**: Adults with high cholesterol levels
- **Intervention**: Consumption of flax seeds
- **Comparison**: Standard diet without flax seeds
- **Outcome**: Can reduce LDL cholesterol levels
- **Claim:** *Consumption of flax seeds can reduce the LDL cholesterol levels in adults with high cholesterol.*

Formulating PICO-encoded health claims can enhance document retrieval by guiding the search toward semantically relevant papers. When the components of PICO are clearly described in a claim, the retrieval process is more likely to identify studies that align with these criteria, improving the precision and relevance of the results. This structured approach helps ensure that the retrieved documents closely match the specific health claim being investigated.

PICO for Evaluating Studies

Additionally, we can use components of the PICO framework as criteria for evaluating the quality and relevance of interventional studies. The PICO framework provides a structured approach to assessing key elements of a study, which can be valuable in determining its relevance and quality towards verifying health claims. Section 5.1 expands on the use of the PICO components for an IC filter.

4.2 Health Claim Dataset

To find relevant nutritional health claims we use two reputable sources. First, NutritionFacts.org [1] provides evidence-based insights on various health and nutrition topics. Second, we rely on the comprehensive review by Fardet and Boirie (2014) on the associations between food and beverage groups and major diet-related chronic diseases as a resource for categorizing and identifying food groups. We take plant-based foods, animal-based foods and beverages as our main categories ((Table 1). For this study, we collected a total of 50 health claims (Appendix A).

Plant-based	Animal-based	Beverages
Fruits	Dairy	Tea
Vegetables	Eggs	Coffee
Grains	Poultry	Dairy
Legumes	Red meat	Alcohol
Nuts and seeds	Fish	Sweet beverages

Table 1: Food Categories

4.3 Expert Annotation of Health Claims

To validate our verdict generation, the 50 PICO-encoded health claims are annotated by a qualified medical research expert using the 1–5 Likert Scale below:

[1] https://nutritionfacts.org/

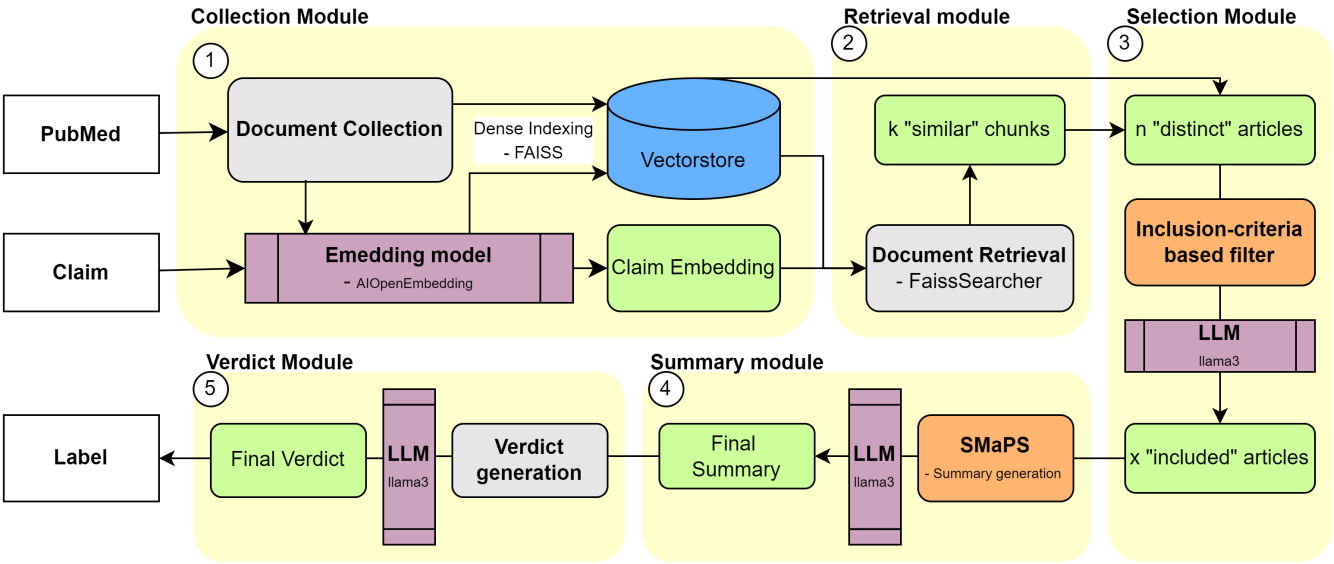


Figure 1: 5 module framework, with enhanced selection and summary modules

1. **Strongly Refuted**: The claim is clearly and unequivocally contradicted by the majority of current scientific evidence. There is strong consensus in the scientific community against it.
2. **Somewhat Refuted**: The claim is mostly contradicted by scientific evidence, though there may be a few studies or minor evidence suggesting otherwise.
3. **Neutral**: The claim neither clearly supports nor refutes based on available scientific evidence or there is a balance of evidence for and against the claim.
4. **Somewhat Supported**: The claim is mostly supported by scientific evidence, though there may be some studies or minor evidence to the contrary.
5. **Strongly Supported**: The claim is clearly and unequivocally supported by the majority of current scientific evidence. There is strong consensus in the scientific community in favour of it.

The expert carefully evaluated each claim using their expertise in the domain and against relevant scientific literature if needed. Their annotations provided a gold standard dataset to evaluate our automated system's performance by comparing its predictions to the expert's judgments, offering valuable insights for refining our approach and improving accuracy in interpreting scientific evidence for health claims.

5 Approach

Inspired by the retrieval augmented claim verification approach by Liu et al. (2024a), we propose a five-module framework that improves the selection of relevant literature and the synthesis of information. While building on its established methods for data collection and retrieval, our focus is on enhancing the **evidence selection** and **explanation** phases. Existing evidence selection methods typically rely only on similarity measures between the claim and evidence, overlooking the quality and relevance of the scientific articles. To address this, we introduce a **selection module** and a **summary module** as shown in Figure 1.

Sub-question one is addressed in section 5.1 and utilizes the first 3 components of the framework. Sub-question two is addressed in section 5.2 and uses the summary generation module. Finally, sub-question three is addressed in section 5.3 and uses the verdict generation module, where we produce a final verdict and generate a label for claim verification.

5.1 Retrieval of Relevant Literature

Sub-question 1: *How accurately does the RAG model retrieve relevant scientific literature for specific health claims?*

To answer this question, first, we need to collect scientific literature and store it in a vector store. Second, we must retrieve information that addresses the question, in this case, the specific health claim. Finally, we filter the retrieved information to ensure it is both relevant and of high quality.

Collection Module

The collection module is responsible for two things:

1. The collection of scientific literature.
2. Processing and storage of literature in a vectorstore.

Data Collection: A key challenge in claim verification is retrieving the latest and most relevant claim-related scientific literature. PuBMed Central [2] (PMC) is a free full-text archive of biomedical and life sciences literature maintained by the National Institute of Health's National Library of Medicine (NIH/NLM). The availability of full-text articles in PMC contributes to their high quality in several ways.

- **Comprehensive information**: Full-text articles provide complete methodologies, results, and discussions, allowing readers to critically evaluate the research in its entirety.

[2] https://www.ncbi.nlm.nih.gov/pmc/

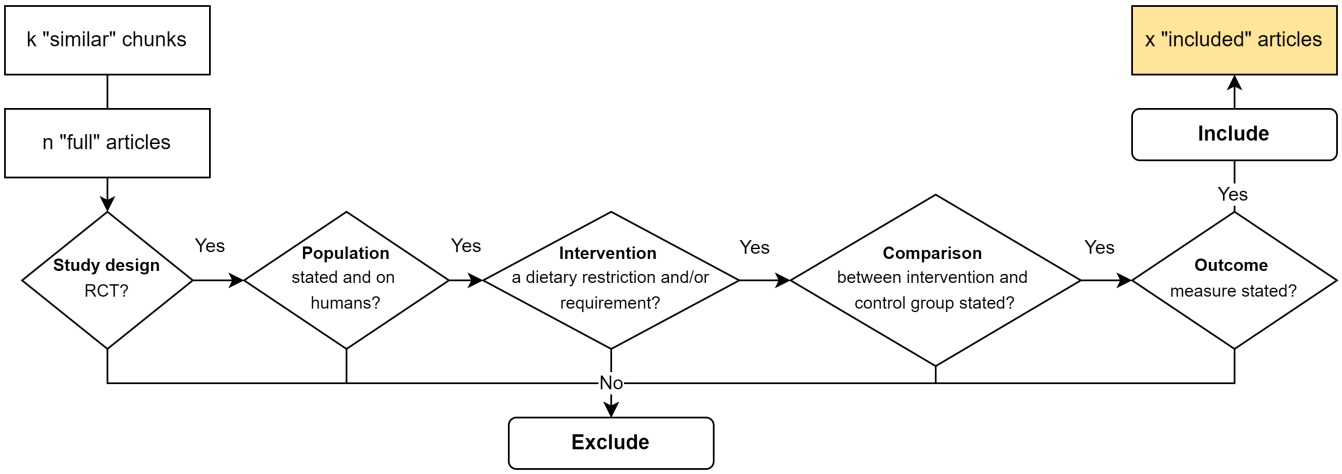


Figure 2: Flowchart of the inclusion criteria filter

- **Peer review**: Most articles in PMC come from reputable, peer-reviewed journals, ensuring scientific rigor.
- **Compliance with funding policies**: PMC includes articles that comply with various research funding policies, such as the NIH Public Access Policy, which mandates the submission of NIH-funded research.

Further, we scope our domain using advanced search options. We take advantage of Medical Subject Headings (MeSH) in PubMed, a National Library of Medicine (NLM) controlled vocabulary thesaurus used for indexing articles. In scoping our domain we use the following MeSH terms.

- **Diet, Food, and Nutrition**: *"Concepts involved with nutritional physiology, including categories of substances eaten for sustenance, nutritional phenomena and processes, eating patterns and habits, and measurable nutritional parameters."*

Currently, there are approximately 200,000 peer-reviewed full-text articles on PubMed Central (PMC) focused on the effects of nutrition on cardiovascular health and blood pressure. However, for this study, we were able to collect only about 10,000 articles, representing just 5% of the available literature. This limitation is due to hardware constraints and the limited time available for scraping and processing the data. Collecting and storing the entire dataset would require significantly more computational resources and time, making it impractical for the scope of this study. Therefore, we focus on a representative subset to ensure feasible analysis while maintaining relevance to the research questions.

Data Vectorstore: As part of the document collection, we scrape scientific literature from PMC, gathering key elements such as the title, abstract, full-text article, and important metadata including authors, journal title, and publication date.

To optimize the retrieval of relevant information each article is chunked into pieces consisting of 1000 tokens. For example, a full article can be split into 50 separate chunks ("Documents") depending on the size of the publication. This chunking process ensures that the text is broken down into digestible portions while maintaining context.

Then, we leverage OpenAIEmbeddings to transform these chunks into high-dimensional vector representations, capturing the semantic essence of the text. Embedding models like OpenAIEmbeddings convert text chunks into vector representations by mapping them into a high-dimensional space where semantically similar texts are positioned closer together. This process involves tokenizing the text, converting tokens to numerical vectors using an embedding layer, and using contextualization techniques, such as transformers, to capture relationships between words. The resulting vector encodes the semantic essence of the text, allowing for tasks like document retrieval and semantic comparison.

Finally, we employ FAISS [3] (Facebook AI Similarity Search), a robust and efficient library for similarity search and clustering of dense vectors, to store these vectors. This approach allows for rapid and accurate retrieval based on semantic similarity, greatly enhancing the accessibility and utility of the scientific literature in the vectorstore.

Retrieval Module

When a query is made, it is also transformed into a vector using the same method used to vectorize documents. FAISS then calculates the similarity between the query vector and the document vectors using the Euclidean distance. FAISS efficiently finds the nearest neighbors, or the most similar documents, by indexing and partitioning the vector space, allowing it to quickly search large datasets and return documents that are most closely aligned with the input query.

In our case, a health claim serves as our input query and FAISS can efficiently search our vectorstore for the documents that are semantically similar to the query vector. However, it is crucial to understand that semantic similarity does not necessarily equate to relevance or quality. While these searches can efficiently return documents with vector representations close to the query vector, they do not inherently assess the content's reliability, or appropriateness for the claim's specific needs. Therefore, we apply an additional IC based filter in the selection module.

[3] https://faiss.ai/

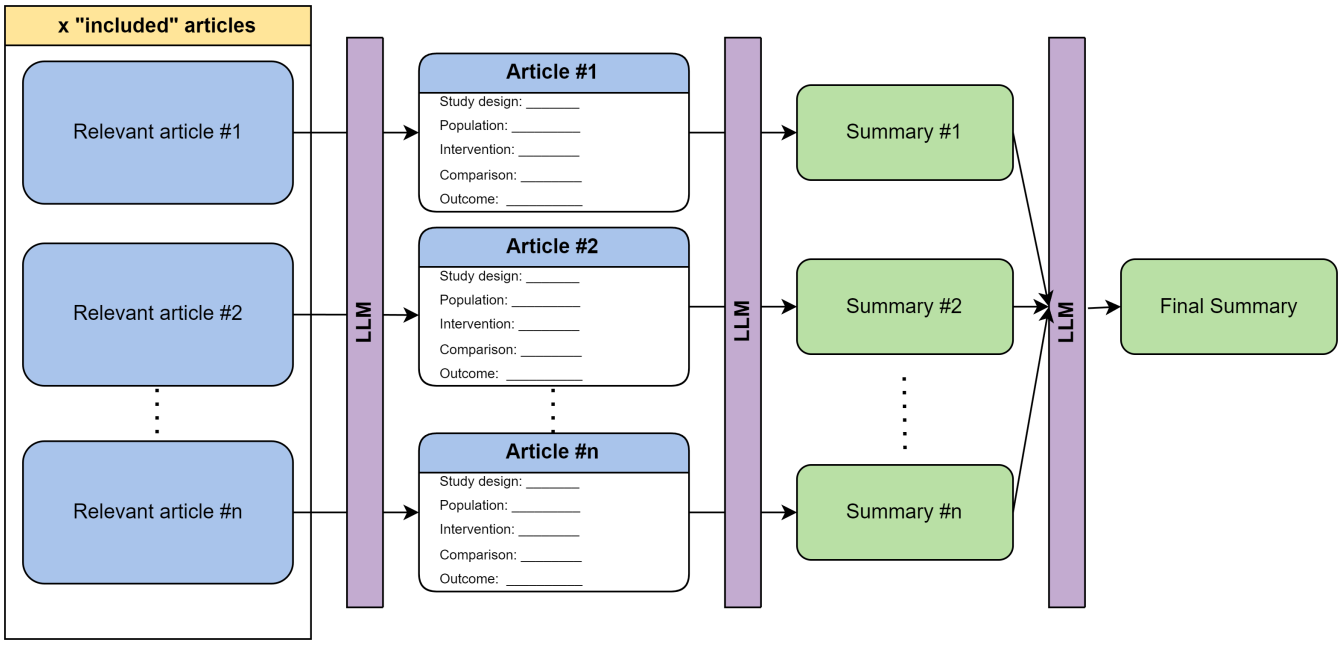


Figure 3: High level overview of the SMaPS pipeline for synthesis of relevant studies

Selection Module

Additional filtering and quality assessment steps are necessary to ensure that the retrieved documents not only match the query semantically but also meet the desired standards of relevance and quality for the intended use case. Leveraging LLMs, RAG and prompt engineering, we present a powerful approach to creating an IC-based filter for document selection. Our approach enables the LLM to make nuanced judgments about a paper‘s suitability, assessing factors like research design, sample size, statistical significance, and alignment with the health claim. The flexibility of prompt engineering allows for adjustment of criteria as research standards evolve, without the need for model retraining. Two main aspects are essential for the selection module:

1. Specification of criteria to assess the relevance and quality of an article.
2. Prompt engineering to apply these criteria for the inclusion or exclusion of an article.

Specification of Criteria: To determine our specific inclusion criteria, we consulted with an expert in the field. Based on their input, the criteria identified for article selection focused on studies that were based on interventional trials, reported clear outcomes related to cardiovascular health, and included a well-defined distinction between the intervention and control groups. These requirements closely align with the PICO framework discussed in Section 4.1, indicating that PICO is well-suited for addressing health claims by ensuring the selection of relevant and rigorous articles. By integrating our expert-defined criteria and the PICO framework we define our criteria as shown in Table 2.

Component	Inclusion Criteria
Study design	Interventional studies
Population	Human adults
Intervention	Based on dietary requirements and/or restrictions in line with the claim
Comparison	Intervention and control group clearly stated and fairly chosen
Outcome	Outcome measures clearly stated concerning cardiovascular health and/or blood pressure

Table 2: Inclusion Criteria

Criteria Prompt Engineering: Prompt engineering involves designing and refining input instruction for LLMs to guide their output effectively. In the context of assessing an article's relevance and quality, prompt engineering is used to craft specific questions that direct the model to evaluate the article based on our predefined criteria. When combined with RAG, which provides the article as external context from the vectorstore, prompt engineering can be used to assess a document's relevance and quality by asking direct questions. The IC prompt structure is given in Figure 2.

Each criterion includes two prompts.

1. The first prompt gathers written responses relevant to the criterion. It uses the full article as context for the generation of the written response.
2. The second prompt uses that written answer as “response“ to generate a YES or NO answer, making it suitable for the IC pipeline.

See appendix B for the full prompts.

5.2 Synthesis of Relevant Literature

Sub-question 2: *How effectively does the model synthesize retrieved literature to generate concise information on claims?*

The summary generation module relies heavily on using the PICO framework for the extraction and synthesis of relevant information. The three main tasks present in this module can be described as follows:

- **Sequential mapping of PICO components**: The LLM gathers data from the written responses based on our predefined criteria, using this as evidence to identify and structure the PICO components.
- **Sequential synthesis**: Since information is presented in a structured way, the LLM is able to synthesize each article on the key aspects sequentially.
- **Aggregation of Summaries**: All individual summaries are then compiled and synthesized into a final summary.

In short, our approach sequentially extracts and synthesises PICO elements from relevant literature to generate a concise final summary. We refer to this process as **SMaPS** (Sequential Mapping of PICO-based Synthesis), highlighting the methodological approach and the underlying framework of summarizing.

SMaPS (Sequential Mapping of PICO-Based Synthesis)

The SMaPS process is a methodical approach that utilizes the PICO framework to ensure structured synthesis of medical literature (Figure 3). Initially, retrieved articles are selected in the selection module ensuring that only relevant articles are included. The selection module not only filters relevant articles but also organizes them in a way that enables efficient sequential mapping in the summary generation module.

Each relevant article is individually analyzed to extract and summarize key information related to PICO components. This process leverages the LLM for a detailed mapping of each PICO component to a concise summary. Finally these individual summaries are aggregated into a comprehensive final summary, providing a clear and evidence-based overview of the research on a specific health claim. The prompt used for the final summary can be found in Appendix C. The SMaPS process effectively synthesizes retrieved literature into concise, relevant summaries about health claims, serving as a crucial intermediary step that informs and enhances evidence-based verdict generation.

5.3 Final Verdict Generation and Label Prediction

Sub-question 3: *How accurately can a RAG-LLM verify health claims when compared to expert annotations?*

We use the final summary generated in the previous step to produce a final verdict. The verdict generation module utilizes the final summary to create an explainable verdict, which then is used to classify the health claim on a LIKERT scale, ranging from strongly refuted (1) to strongly supported (5). The goal of this process is to assess how our model can validate health claims in comparison to expert annotations.

Specifically, we task the LLM to perform two tasks:

1. **Verdict generation**: The LLM is tasked to produce a single concise verdict based on the final summary (Appendix D).
2. **Label prediction**: The LLM is tasked to predict a label based on the verdict. Labels represent classes on a 1–5 Likert scale (Appendix E).

LLM Label Prediction

We utilize LLMs to generate labels. Our system defines a classification class that outlines the five possible class labels on the 1–5 Likert scale, ranging from "strongly refuted, (1)" to "strongly supported, (5)", with detailed descriptions provided for each label. The LLM's temperature is set to 0, which ensures the prediction is purely based on the verdict without any creativity or variation. By integrating the LLM for prediction with structured input from previous stages, the system is able to predict on synthesized information from relevant literature, providing a well-informed, evidence-based verification of the claim.

6 Experiment and Evaluation

In this section, we will present experimental designs for evaluating the selection module, the summary generation module, our model's label prediction, and give a more in-depth analysis of claim verification.

6.1 Selection Module Evaluation

To evaluate relevant literature retrieval, we manually reviewed and annotated articles based on the PICO criteria used in the selection module (Figure 2). Each article manually went through the inclusion process and was then labelled "included" or "excluded". These labels were then used as ground truths for assessing the accuracy of the IC filter. This process helped establish a benchmark to assess the accuracy of the module and the LLM in filtering relevant studies.

The first author of the paper classified 200 articles, ensuring a balanced set with similar proportions of "included" and "excluded" studies as predicted by our selection module. This balance is needed for ensuring an unbiased performance evaluation and assessing the LLM's ability to differentiate between relevant and irrelevant studies. The manual annotation also highlighted potential shortcomings in the automation process of filtering based in PICO elements, revealing areas where the LLM faced challenges in applying the PICO criteria.

6.2 Summary Generation Module Evaluation

To evaluate the summary generation module, we conducted a pilot study with five health practitioners using surveys distributed via Google Forms. Each practitioner was asked to evaluate two research papers and their respective PICO elements, as well as two summaries with their corresponding health claims. In total, five different papers and claims were assessed, with each paper and claim being evaluated twice by different practitioners to mitigate individual bias and balanced assessment. The goal of the pilot study is to asses key aspects of the SMaPS model, such as accuracy and completeness of the extracted PICO elements; coherence and clarity of the generated summaries; and overall usefulness of the summary in addressing health claims.

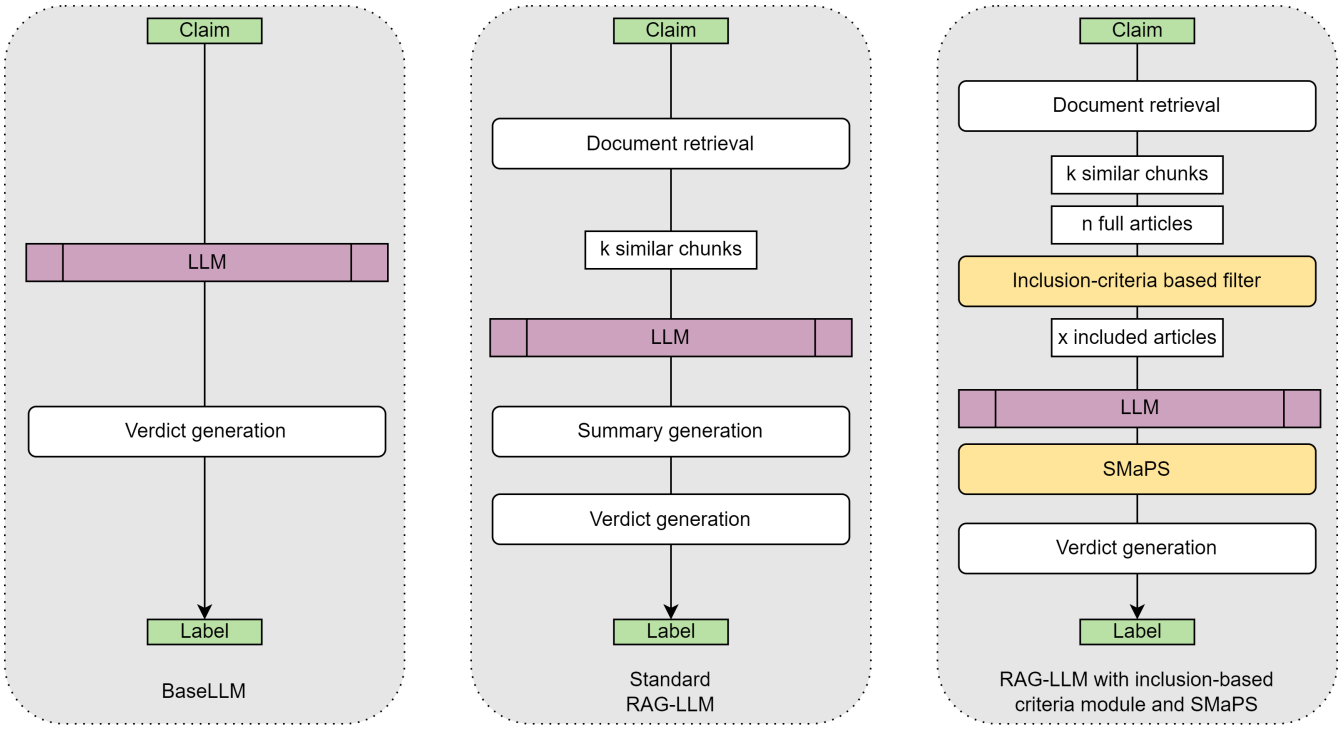


Figure 4: High-level overview of comparison between the BaseLLM, Standard RAG-LLM, and our Advanced RAG-LLM.

By gathering responses through structured open-ended questions and LIKERT scales, the survey will provide valuable insights into how well the SMaPS method extracts and synthesizes literature. Detailed description of the assessed aspects and the corresponding questions being asked are as follows:

- **Extraction of PICO elements**: Evaluating the **accuracy** and **completeness** of the extracted PICO elements. This ensures the model processes each element reliably and that no relevant information is skipped or misrepresented.
  - **Accuracy**: How accurate is the extracted component when compared to the information in the original paper?
  - **Completeness**: Does it include all the relevant information?
- **Summarization of PICO elements**: Evaluating the **consistency**, **clarity** and **usefulness** of the generated summary. These aspects ensure that the PICO components are reflected correctly, while maintaining coherence.
  - **Consistency**: How consistent is the summary to the extracted PICO elements?
  - **Clarity**: Is the summary written in clear, plain language that can be easily understood?
  - **Usefulness 1**: Does the summary provide useful information to answer the respective claim?
  - **Usefulness 2**: What additional information would you find helpful to provide a well-informed response or meaningful advice regarding the claim?

6.3 Model Evaluation

To valuate the accuracy of our model we compare its classification predictions against the expert-annotated health claims. The key is to leverage the curated dataset as a reliable benchmark to validate the model‘s decision making capabilities and whether it provides accurate predictions. The classification labels predicted by the model range over a Likert scale from 1–5 in the same way that the health claims are annotated by experts.

Additionally, we compare the performance of our model against two other models:

1. **BaseLLM**: A Basic LLM that does not utilize an advanced retrieval system to access literature from a data vectorstore.
2. **Standard RAG-LLM**: A LLM model that utilizes RAG to access literature from our vectorstore, that does not incorporate the inclusion-criteria based filter.
3. **Advanced RAG-LLM**: Our model that utilizes RAG and is enhanced by the inclusion-criteria based filter and the SMaPS process.

Our criteria-inclusion filter is designed to refine the model's ability to assess health claims by ensuring that only relevant information aligned with predefined criteria is considered. By contrasting these three models, we aim to identify whether the inclusion of RAG, the IC filter, and incorporating SMaPS leads to improvements in the model's accuracy and reliability when verifying health claims. Our expectation is that our Advanced RAG-LLM will demonstrate better per-

ID	Claim	Human Label	Adv. RAG-LLM
6	Consumption of bananas can lower blood pressure in human adults	5	2
18	Consumption of full-fat dairy can lower blood pressure in human adults	1	5
19	Consumption of cheese can lower blood pressure in human adults	1	3
31	Consumption of bananas can improve cardiovascular health in human adults	5	2
43	Consumption of full-fat dairy can improve cardiovascular health in human adults	1	4

Table 3: Claims with the Largest Difference Between Advanced RAG-LLM Prediction and Human Annotation

formance by providing more focused and relevant outputs. A high-level overview of the experiment is depicted in figure 4.

We use the **Cohen's weighted kappa** to measure the agreement between predictions and expert annotations. The Cohen's kappa takes into account whether two different raters, in this case the model and the expert, measured the same value. The weighted Cohen's kappa considers the degree of agreement between the raters based on how far apart the values are. In our case of a LIKERT scale from 1-5, the weighted kappa penalizes disagreements more if the value ratings are further apart.

6.4 Claim Analysis

In this analysis, we focus specifically on the evidence used in our Advanced RAG-LLM model to gain deeper insights into the model's behaviour. We analyzed multiple claims where the model predictions diverged the most from the human annotations, indicating significant discrepancy between the model's output and human judgement. Specifically, we will give in-depth analysis of claim 6 and claim 18 in this report (Table 3), as these claims show the largest discrepancies. The evidence consists of the relevant research papers used, the generated summary, and the final verdict it reached. (Please refer to Appendix H.1 and H.2 for the complete evidence used to verify claim 6 and claim 18.). By analysing these components, we aim to understand how the model interpreted the evidence and why it arrived at a different conclusion.

7 Evaluation Results

In this section we present the evaluation results

7.1 Selection Module Evaluation Results

Results indicates good performance in distinguishing between relevant and irrelevant literature (Figure 5). Specifically, the model correctly included 80 out of 83 articles, resulting in a true positive rate of 96.7%. Conversely, the model had only 3 false negatives, where relevant articles were mistakenly predicted as "Excluded", with false negative rate of 3.6%. This translates to a **recall value of 96.4%**, indicating that the model was did not miss relevant literature. These metrics highlight the model's strong ability to classify literature accurately, with its high recall ensuring that few relevant studies are missed. However, it misclassified 20 articles as "Included" that should have been excluded, resulting in a false positive rate of 17.1%, resulting in a precision of

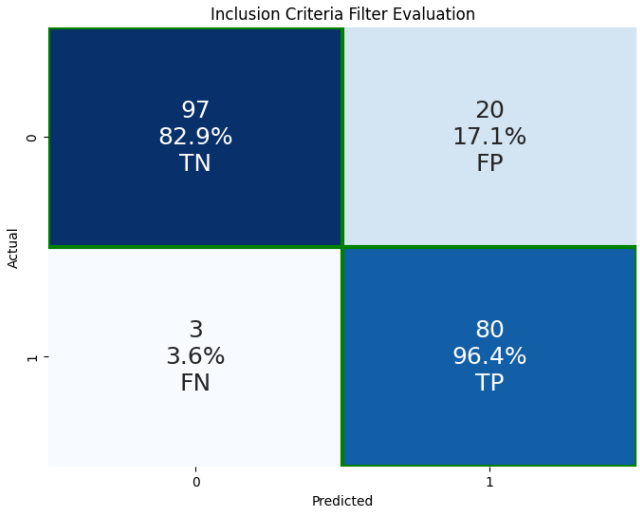


Figure 5: Confusion matrix for evaluating the inclusion criteria filter labelling studies as either 0 ("excluded") or 1 ("included")

80%. This lowers the model's precision value, as irrelevant articles were mistakenly used as evidence for the verification of health claims. These false positives affect the reliability of the verdict by introducing irrelevant or incorrect information, which can compromise the quality and accuracy of the summary and verdict generation.

The false positives were primarily caused by the Study Design element, where articles were incorrectly flagged as interventional studies. This issue often arose because meta-analyses or systematic reviews sometimes mention interventional studies, leading to the incorrect classification of the entire study as interventional. Additionally, the Population and Outcome elements had occasional false positives, where the studies focusing on animals were incorrectly identified as involving human subjects or where study outcomes did not specifically address blood pressure. These errors underscore some of the limitations of the selection module, highlighting the need for improvement in accurately distinguishing study designs and addressing inconsistencies in the Population and Outcome elements.

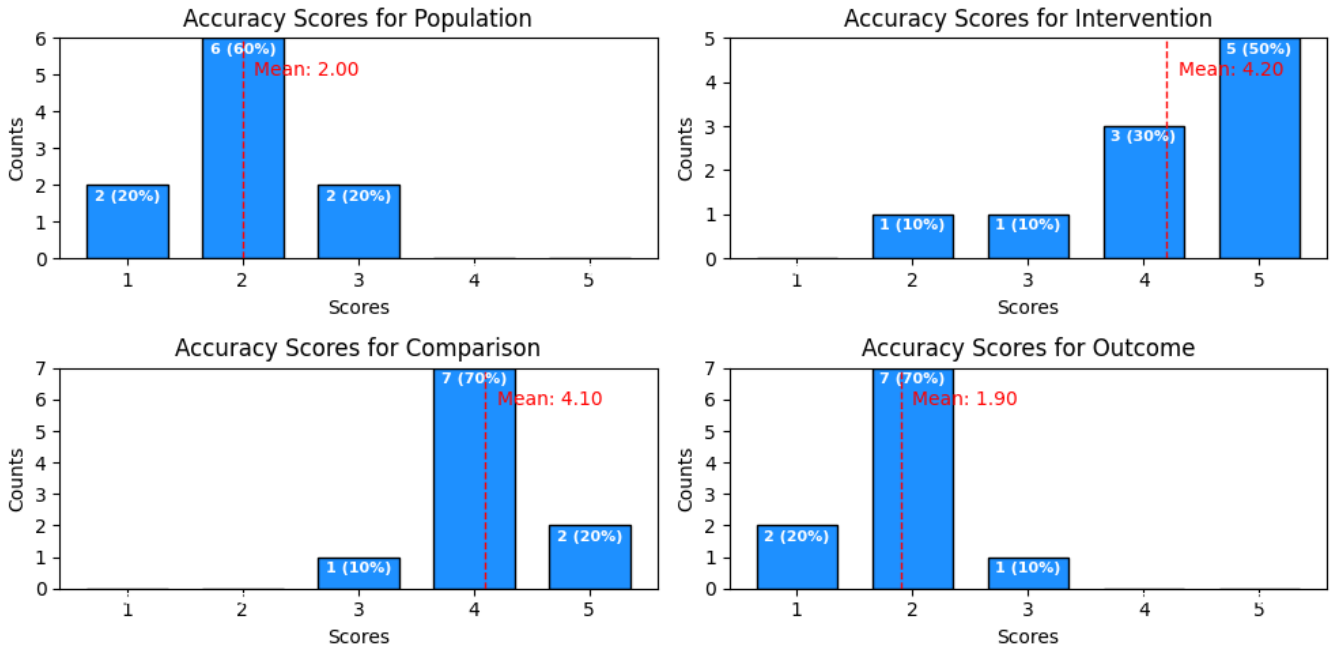


Figure 6: Distribution of Accuracy Scores Across Different PICO Elements

7.2 Summary Module Evaluation Results

The survey results revealed that the extraction of the Population and Outcome elements lacked accuracy, with averages of 2.00 and 1.90, respectively (Figure 6). Specifically, the Population scores predominantly fell at 2 ("somewhat inaccurate"), reflecting recurring inaccuracies in key details extracting exact sample sizes and missing detail on selection methods. There was a frequent issue with incorrect or missing numbers, which was evident across multiple responses. For the Outcome element, the problem was even more prominent, as 70% of the scores fell at 1 ("very inaccurate"), highlighting severe inaccuracies. Notably, there were instances where vital data, such as changes in blood pressure and the exact measurement of blood pressure levels, were completely omitted despite their explicit mention in the original papers. This omission of relevant data is also linked to the low completeness scores of both Population and Outcome, with both averaging 2.70. Evaluators highlighted the need for precise blood pressure measurements and exact sample sizes, which were missing in the extracts. This suggest a critical need for improvement in data extraction processes to capture exact details accurately.

In contrast, the extractions for the Intervention and Comparison elements demonstrated much better accuracy with averages of 4.20 and 4.10, respectively. Half of the scores for Intervention were scored 5 ("very accurate"), indicating accurate and detailed descriptions of study interventions, though occasional inconsistencies appeared. Comparison details were generally extracted reliably, predominantly scoring at 4 ("somewhat accurate"). Completeness scores averaged at 4.00 and 4.10, respectively. Evaluators highlighted that adding information on run-in periods and specific calorie intake would increase the value of these elements. This shows effective capture of control group specifics, despite some scores at extreme ends pointing to occasional inaccuracies in detail accuracy.

While the summaries were generally consistent and clearly written, with averages of 4.00 and 3.90 respectively, there were deemed insufficient for effectively addressing health claims (Appendix G.2). Specifically, the majority of the consistency evaluations indicated good alignment with the extracted PICO elements. Clarity, with half of the scores at the top of the scale, suggesting the summaries were well-articulated. However, the summaries often mentioned general trends such as "reduction" or "no significant difference" in blood pressure but failed to include exact measurements. Evaluators highlighted the critical eed for precise numerical outcomes, which significantly affected the summaries' usefulness, reflected in a low usefulness score of 2.10. Half of these scores were scored 1 ("Very useless"), emphasizing that the omission of accurate outcome and detailed population information was a primary reason for the summaries' limited utility in verifying health claims

In conclusion, the extracted PICO elements received high ratings for their accuracy and completeness to the Intervention and Comparison elements, reflecting strong alignment with the original papers. However, the absence of precise and detailed numerical data for the Population and Outcome elements significantly compromised the summaries' credibility and utility. Such details are crucial for accurately assessing the efficacy of interventional trials, highlighting a key area for improvement in data extraction and presentation.

7.3 Model Evaluation Results

Cohen's Weighted Kappa Scores

Model	Cohen's Weighted Kappa
BaseLLM	0.31
Standard RAG-LLM	0.27
Inclusion-based RAG-LLM	0.48

Table 4: Model Performance with Cohen's Weighted Kappa

The BaseLLM achieved a kappa score of 0.31 and the Standard RAG-LLM model a score of 0.27, which indicates "minimal" agreement according to the widely accepted interpretation of kappa scores (McHugh, 2012). This suggests a relatively low level of reliability in its performance for the given task. Our Advanced RAG-LLM achieved a weighted kappa score of 0.48, indicating "weak" agreement (McHugh, 2012). While this is an improvement over the standard RAG-LLM agreement, it still suggests that the model's predictions are not highly reliable and deviate from the ground truth more often than desired.

Label Prediction Results and Model Analysis

To gain deeper insights in the label prediction and behaviours of different models we give confusion matrices of each model. Additional classification reports are used to get an overview model's performance in terms of precision, recall, and F1-scores (Appendix F).

The confusion matrix for the BaseLLM (Figure 7) reveal a clear bias towards predicting supportive labels, particularly in classes 4 with recall of 100% and precision of 10%. The model shows highest precision in class 5 with 60% and struggles significantly with all other classes, especially classes 1–3. This tendency to overpredict class 4, shows behaviour that favors supportive labels.

For the Standard RAG-LLM model (Figure 8), the confusion matrix shows a similar pattern, with the model excelling at classifying class 5 but performing poorly across the other classes 1–4. The standard model demonstrates a high recall of 95% for class 5, correctly identifying 21 out of 22 instances, but its lower precision of 58% reflects a significant number of false positives. Showing a strong tendency to overpredict class 5.

Results for the Advanced RAG-LLM (Figure 9) reveal that it effectively identifies class 5 with 11 correct predictions with a precision of 92%, but moderate recall 50%, indicating it misses half of the actual instances. The models struggles across the other classes, notably class 1 and class 2 with a low recall of 17% and 30%, respectively. The matrix highlights a significant over-prediction of class 3, with class 1 and 2 often incorrectly classified into this category. The high recall of 75% and low precision of 13%, supports the the behavior of overprediction in class 3. These results indicate that the model excels in accurately predicting instances of class 5 but exhibits a bias towards class 3, representing a neutral or uncertain stance.

In conclusion, both the BaseLLm and Standard RAG-LLM models show bias towards the supportive classes 4 and 5, while our Advanced RAG-LLM mostly overpredicts class 3. However, the Advanced RAG-LLM shows small improvements in accurately classifying the refuting classes 1 and 2, where the other models completely failed to classify instances of class 1. With a weighted-average F1 score of 0.46, our model shows modest improvement over the BaseLLM and Standard RAG-LLM, which scored 0.29 and 0.33 respectively (Appendix F). Additionally, with changes in the prediction pattern—particularly the reduced over-prediction of supportive classes and the improved ability to classify refuting classes—the model demonstrates a more balanced performance. The increase in neutral predictions suggests the model adopts a more cautious approach in ambiguous situations, leading to more neutral outcomes when the data does not strongly suggest other categories.

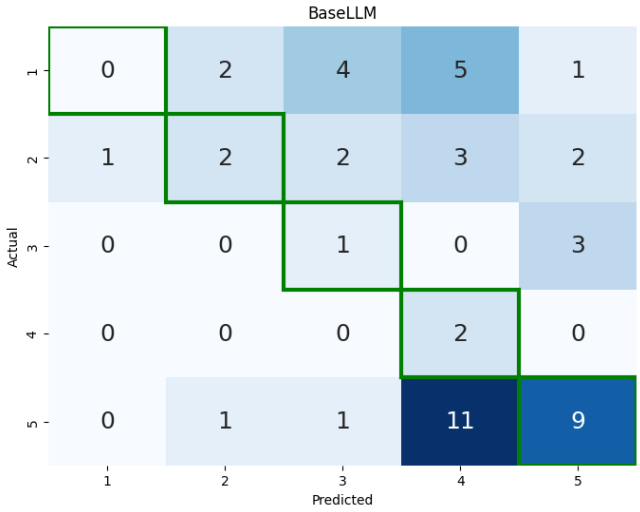


Figure 7: Confusion matrix prediction (BaseLLM)

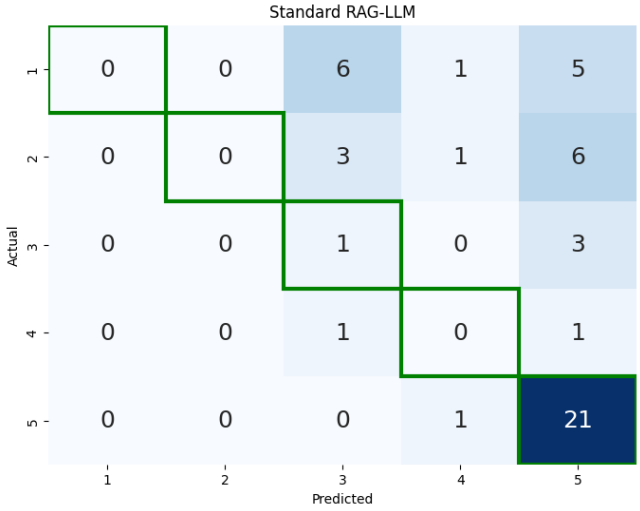


Figure 8: Confusion matrix prediction (Standard RAG-LLM model)

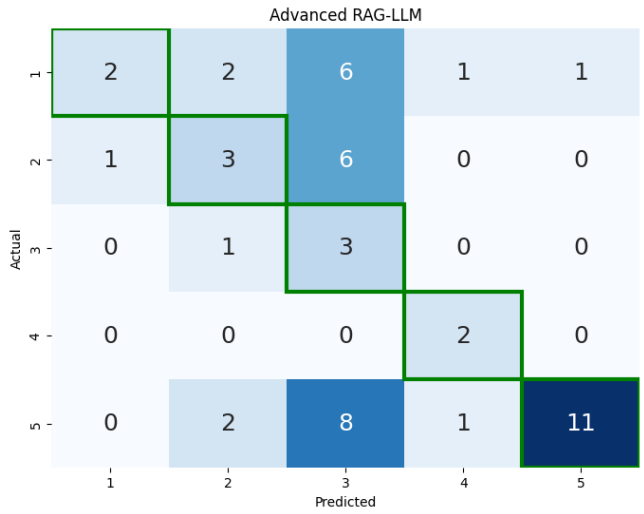


Figure 9: Confusion matrix prediction (Advanced RAG-LLM model)

7.4 Claim Analysis Results

In this section we make an in-depth analysis by looking at the verdict, summary and the used sources for claims verified using our inclusion-based RAG-LLM model.

After analysing claim 6, we see that the studies investigated do not specifically investigate the effects of bananas on blood pressure (Table 10). While the verdict does note the effects of nitrate-rich beetroot juice, this certainly does not give information on the effects of bananas. It is clear that the studies do not specifically address bananas' effects on blood pressure. Two of the three studies mistakenly focus on dairy products and beetroot juice. The final study examines overall fruit consumption, not bananas specifically, resulting in an overgeneralized analysis of fruits.

When we look at claim 18, the verdict supports the claim according to one study that found a reduction in blood pressure (Appendix H.2). Noting that this finding is specific to dairy consumption, and not full-fat dairy. While the 5 investigated papers all focus on dairy consumption none specifically focus on full-fat dairy (Table 12). While 2 papers investigated the effects of low-fat and reduced-fat dairy products, the others focused on general dairy products. This reveals a recurring pattern of overgeneralization, where the analysis focuses on dairy products broadly rather than specifically examining the effects of full-fat dairy.

After reviewing additional claims (Table 3) and their supporting evidence, we observe a recurring pattern across claims, where broader, more encompassing parent categories are used as evidence instead of the specific nutrition mentioned in the claim. This overgeneralization is particularly evident in claims about specific foods like full-fat dairy, cheese, and bananas, where the analysis often shifts focus to these broader categories rather than the precise item. Interestingly, claims related to broader food groups, such as fruits, nuts, or leafy greens, tend to produce more consistent results, as they naturally align with the wider scope of the analysis.

Overall, the analysis reveals that the selection process still tends to use irrelevant literature for certain claims. This is particularly evident in the bananas claim, where studies on dairy products or beetroot juice were mistakenly used. This error appears to stem from the retrieval mechanisms, where a similarity search may have detected the keyword "bananas" or "lower blood pressure" in an article, allowing it to pass through the selection process despite not being directly related to the claim. Additionally, the pattern of overgeneralization highlights the need for improvements in the selection process to better retrieve evidence specific to the targeted nutrition.

8 Results Summary and Discussion

This section presents a summary of the results and systematically addresses and answers each sub-question to ultimately provide a comprehensive answer to the main research question. It concludes with a discussion.

8.1 Results Summary

Sub-question 1: ***How accurately does the PICO enhanced RAG model retrieve relevant scientific literature for health claims?***

Our inclusion criteria (IC) filter achieved a high recall of 96.4%, but its precision was lower at 80% due to 20 irrelevant articles being included, resulting in a 17.1% false positive rate. Most misclassifications occurred in the Study Design element, where meta-analyses or systematic reviews were mistaken for interventional studies, and in the Population and Outcome elements, where animal studies were misclassified as human studies or outcomes unrelated to blood pressure were included. Additionally, our analysis revealed a recurring issue of **overgeneralization** in verifying health claims, where the model relied on broad evidence that did not align with the specific focus of the claim. For example, it used studies on general dairy products instead of full-fat dairy and research on fruit for claims about bananas' effects on blood pressure. This practice led to using related but insufficiently specific literature for the claims being assessed.

Sub-question 2: ***How effectively does the model synthesize retrieved literature to generate concise information about health claims?***

While the extraction of Intervention and Comparison elements was well-executed, the **absence of precise numbers**—particularly in the Population and Outcome elements—significantly diminished the credibility and usefulness of the summaries. In interventional trials, especially those measuring critical factors like blood pressure, exact data is essential for proper evaluation and clinical decision-making. Without accurate sample sizes and precise outcome measurements, such as the specific changes in blood pressure levels, the summaries fail to provide the depth of information needed to assess the trial's validity. For these summaries to be truly useful and trustworthy in addressing health claims, they must include exact numerical data, which plays a vital role in the rigorous evaluation of interventional outcomes.

Sub-question 3: ***How accurately can a RAG-LLM verify health claims when compared to expert annotations?***

The accuracy of our Advanced RAG-LLM shows a modest but important improvement over the BaseLLM and Standard RAG-LLM, particularly in classifying instances of refuting classes and achieving a more balanced performance, as our **model adopts a more neutral stance** in its assessments. With a weighted F1 score of 0.46, the model improves upon its predecessors, which scores 0.29 and 0.33, respectively. Additionally, the Cohen's weighted kappa score of 0.48, while indicating "weak" agreement, marks an improvement over the BaseLLM (0.31) and Standard RAG-LLM (0.27). However, the numbers still suggest that the model's accuracy is limited, with its predictions still deviating from the ground truth more often than desired. Overall, while the IC filter and the SMaPS approach show a step forward in verifying health claims compared to more standard models, it remains less accurate than expert annotations and requires further refinement to enhance its reliability.

8.2 Discussion

The results indicate that our Advanced RAG-LLM adopts a more neutral, or cautious, approach when verifying health claims, frequently over-predicting class 3, which corresponds to neutral or uncertain outcomes. This pattern suggests that the model defaults to neutrality when the available evidence is insufficient or inconclusive, requiring more robust and comprehensive data before making a definitive classification. A closer analysis of this behavior reveals several underlying factors. The implementation of the IC filter and SMaPS approach, while enhancing the selection and summary process, also introduces specific challenges that contribute to the model's neutral stance. Occasional errors in the inclusion of irrelevant literature, overgeneralization of evidence, and the lack of precise numerical data in Outcome Measures all play a role in this cautious behavior.

The overgeneralization of evidence is particularly impactful. By relying on broad categories of studies, the model introduces ambiguity into its evidence base. This generalization weakens the specificity required to make accurate health claim verifications and reinforces the model's tendency to predict neutral outcomes. This behavior underscores the importance of improving the model's ability to differentiate between closely related but distinct studies.

Furthermore, the absence of precise numerical data in the Outcome Measures is a significant limitation that contributes to the model's neutral stance. In health-related claims, particularly those involving interventional outcomes like changes in blood pressure, the inclusion of exact figures—such as sample sizes or specific measurement changes—is critical for making informed decisions. Without these precise data points, the model's ability to generate meaningful summaries is diminished, leading to further uncertainty in its predictions.

9 Conclusion

In response to the main research question, *"How effectively can a Retrieval-Augmented Generation-based Language Model verify nuanced health claims?"*, our findings reveal that our Advanced RAG-LLM model demonstrates significant limitations in accuracy (F1 score of 0.46). While showing improvements over the BaseLLM and Standard RAG-LLM models with a Cohen's weighted kappa of 0.48, results still indicate a "weak" agreement. The IC filter and SMaPS approach help balance predictions and reduce bias towards supportive labels. However the model tends to default to neutral outcomes, demonstrating a cautious approach when the evidence is inconclusive. Occasional errors in the inclusion of irrelevant literature, overgeneralization of evidence, and the lack of precise numerical data in Outcome Measures all play a role in this cautious behavior. While the model shows promise, these shortcomings highlight the need for further refinements in both retrieval mechanisms and synthesis processes to elevate its performance to expert-level accuracy and precision in verifying nuanced health claims.

10 Limitations and Future Work

10.1 Limitations

This study has several key limitations that affect both the generalizability and reliability of its findings. First, the model evaluation is based on a relatively small sample of just 50 annotated health claims, which limits the statistical power of the results and the model's ability to generalize to a wider range of claims. Furthermore, these annotations were provided by a single expert, increasing the risk of bias and potential errors. The inclusion of a larger, more diverse set of health claims and the involvement of multiple experts for cross-validation would improve the robustness of the evaluation, reducing the impact of individual biases and providing a more balanced perspective.

Second, the evaluation of the IC filter relies on a dataset of 200 manually annotated articles, which were not reviewed by domain experts. This raises concerns about the reliability of the ground truth labels, as non-expert annotations may introduce inaccuracies. Incorporating expert annotations for this dataset would ensure that the model's performance is evaluated against a higher-quality, more precise standard, leading to more trustworthy conclusions about its accuracy and relevance.

Third, this study was conducted on a subset of 10,000 articles, representing only 5% of the total literature available on nutrition and cardiovascular health. This smaller dataset may not fully capture the range and complexity of the broader scientific literature, potentially limiting the model's ability to generalize its findings to a wider body of health claims. As such, the model's current performance may not reflect its full potential had it been evaluated on a more comprehensive dataset.

10.2 Practical Limitations

From a practical perspective, a major limitation of the study was the hardware constraints, particularly the limited computational power. This restriction prevented the use of the most

advanced LLMs and limited the ability to process the entire dataset of 200,000 articles.

10.3 Future Work

To enhance the model's effectiveness and overcome current limitations, future work should focus on the following areas:

1. **Improving the IC Filter**: Enhancing the filter's precision will better differentiate relevant from irrelevant studies, reducing overgeneralization and neutral predictions.
2. **Refining Evidence Selection and Synthesis**: Fine-tuning study selection with a focus on specific data and including precise numerical evidence will improve the accuracy of health claim assessments.
3. **Using Larger, Expert-Annotated Datasets**: Employing larger, diverse datasets annotated by multiple domain experts will improve the model's evaluation and generalizability.
4. **Overcoming Hardware Limitations**: Utilizing more powerful hardware or cloud services will support larger models and datasets, enabling better handling of complex health claims.

Addressing these areas will help the Advanced RAG-LLM achieve better accuracy and reliability in verifying health claims.

References

Adams, K., Greiner, A. C., and Corrigan, J. M. (2004). Patient Self-Management Support. In *The 1st Annual Crossing the Quality Chasm Summit: A Focus on Communities*. National Academies Press (US).

Armitage, H. (2019). Any way you slice it, there's a lot to say about nutrition studies.

Badimon, L., Chagas, P., and Chiva-Blanch, G. (2019). Diet and Cardiovascular Disease: Effects of Foods and Nutrients in Classical and Emerging Cardiovascular Risk Factors. *Current Medicinal Chemistry*, 26(19):3639–3651.

Belluz, J. (2016). Why (almost) everything you know about food is wrong.

Dineen-Griffin, S., Garcia-Cardenas, V., Williams, K., and Benrimoj, S. I. (2019). Helping patients help themselves: A systematic review of self-management support strategies in primary health care practice. *PLoS ONE*, 14(8):e0220116.

EFSA (2023). Health claims. Section: Topics.

European-Commission (2024). Health claims.

Eysenbach, G. (2023). The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. *JMIR Medical Education*, 9(1):e46885. Company: JMIR Medical Education Distributor: JMIR Medical Education Institution: JMIR Medical Education Label: JMIR Medical Education Publisher: JMIR Publications Inc., Toronto, Canada.

Fardet, A. and Boirie, Y. (2014). Associations between food and beverage groups and major diet-related chronic diseases: an exhaustive review of pooled/meta-analyses and systematic reviews. *Nutrition Reviews*, 72(12):741–762.

Gaidai, O., Cao, Y., and Loginov, S. (2023). Global Cardiovascular Diseases Death Rate Prediction. *Current Problems in Cardiology*, 48(5):101622.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., and Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs].

Grady, P. A. and Gough, L. L. (2014). Self-Management: A Comprehensive Approach to Management of Chronic Conditions. *American Journal of Public Health*, 104(8):e25–e31.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A Survey on Automated Fact-Checking. arXiv:2108.11896 [cs].

Hariton, E. and Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG : an international journal of obstetrics and gynaecology*, 125(13):1716.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232 [cs].

Huang, X., Lin, J., and Demner-Fushman, D. (2006). Evaluation of PICO as a knowledge representation for clinical questions. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2006:359–363.

Jo, E., Epstein, D. A., Jung, H., and Kim, Y.-H. (2023). Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Kotonya, N. and Toni, F. (2020). Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L. D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., and Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198. Publisher: Public Library of Science.

Lee, C.-j., Nagler, R. H., and Wang, N. (2018). Source-specific Exposure to Contradictory Nutrition Information: Documenting Prevalence and Effects on Adverse Cognitive and Behavioral Outcomes. *Health communication*, 33(4):453–461.

Liu, H., Soroush, A., Nestor, J. G., Park, E., Idnay, B., Fang, Y., Pan, J., Liao, S., Bernard, M., Peng, Y., and Weng, C. (2024a). Retrieval augmented scientific claim verification. *JAMIA Open*, 7(1):ooae021.

Liu, L., Yang, X., Lei, J., Liu, X., Shen, Y., Zhang, Z., Wei, P., Gu, J., Chu, Z., Qin, Z., and Ren, K. (2024b). A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions. arXiv:2406.03712 [cs].

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

Moons, P. and Van Bulck, L. (2023). ChatGPT: can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. *European Journal of Cardiovascular Nursing*, 22(7):e55–e59.

Myers, M. and Ruxton, C. H. S. (2023). Eggs: Healthy or Risky? A Review of Evidence from High Quality Studies on Hen's Eggs. *Nutrients*, 15(12):2657.

Nagler, R. H. (2014). Adverse outcomes associated with media exposure to contradictory nutrition messages. *Journal of health communication*, 19(1):24–40.

Pradeep, R., Ma, X., Nogueira, R., and Lin, J. (2021). Scientific Claim Verification with VerT5erini. In Holderness, E., Jimeno Yepes, A., Lavelli, A., Minard, A.-L., Pustejovsky, J., and Rinaldi, F., editors, *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Qama, E., Rubinelli, S., and Diviani, N. (2022). Factors influencing the integration of self-management in daily life routines in chronic conditions: a scoping review of qualitative evidence. *BMJ Open*, 12(12):e066647.

Richardson, W. S., Wilson, M. C., Nishikawa, J., and Hayward, R. S. (1995). The well-built clinical question: a key to evidence-based decisions. *ACP Journal Club*, 123(3):A12. Publisher: American College of Physicians.

Soleimani, A., Monz, C., and Worring, M. (2020). BERT for Evidence Retrieval and Claim Verification. In Jose, J. M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M. J., and Martins, F., editors, *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

Sugano, M. and Matsuoka, R. (2021). Nutritional Viewpoints on Eggs and Cholesterol. *Foods*, 10(3):494.

Tan, X., Zou, B., and Aw, A. T. (2023). Evidence-based Interpretable Open-domain Fact-checking with Large Language Models.

Tucci, L. (2021). Why Nutrition Information Seems So Contradictory.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need. arXiv:1706.03762 [cs].

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. (2020). Fact or Fiction: Verifying Scientific Claims. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., Huang, L., Liu, X., Kuo, T.-W., Guan, N., and Xue, C. J. (2024). Retrieval-Augmented Generation for Natural Language Processing: A Survey. arXiv:2407.13193 [cs].

Xu, Z., Jain, S., and Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817 [cs].

Yu, P., Xu, H., Hu, X., and Deng, C. (2023). Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration. *Healthcare*, 11(20):2776. Number: 20 Publisher: Multidisciplinary Digital Publishing Institute.

Zhang, Q., Ding, K., Lyv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., Feng, K., Zhuang, X., Wang, Z., Qin, M., Zhang, M., Zhang, J., Cui, J., Huang, T., Yan, P., Xu, R., Chen, H., Li, X., Fan, X., Xing, H., and Chen, H. (2024). Scientific Large Language Models: A Survey on Biological & Chemical Domains. arXiv:2401.14656 [cs].

Zheng, Y., Koh, H. Y., Ju, J., Nguyen, A. T. N., May, L. T., Webb, G. I., and Pan, S. (2023). Large Language Models for Scientific Synthesis, Inference and Explanation. arXiv:2310.07984 [cs].

A Claim dataset

ID	Claim	Human Label	Base LLM	Standard RAG-LLM	Advanced RAG-LLM
1	Consumption of nuts can lower blood pressure in human adults	5	4	5	4
2	Consumption of flaxseeds can lower blood pressure in human adults	5	5	5	3
3	Consumption of green leafy vegetables can lower blood pressure	5	5	5	5
4	Consumption of fruits can lower blood pressure in human adults	5	4	5	5
5	Consumption of berries can lower blood pressure in human adults	5	4	4	3
6	Consumption of bananas can lower blood pressure in human adults	5	2	5	2
7	Consumption of beetroot can lower blood pressure in human adults	5	4	5	3
8	Consumption of avocados can lower blood pressure in human adults	5	4	5	3
9	Consumption of legumes can lower blood pressure in human adults	5	4	5	5
10	Consumption of beans can lower blood pressure in human adults	5	4	5	5
11	Consumption of whole grains, compared to standard diet with refined grains, can lower blood pressure in human adults	5	5	5	3
12	Consumption of low-fat dairy products can lower blood pressure in human adults	2	3	5	3
13	Consumption of fish can lower blood pressure in human adults	2	5	3	3
14	Consumption of dark chocolate can lower blood pressure in human adults	4	4	5	4
15	Consumption of eggs can lower blood pressure in human adults	1	4	3	2
16	Consumption of red meat can lower blood pressure in human adults	1	3	3	3
17	Consumption of caffeine can lower blood pressure in human adults	3	5	3	2
18	Consumption of full-fat dairy can lower blood pressure in human adults	1	3	5	5
19	Consumption of cheese can lower blood pressure in human adults	1	4	5	3
20	Consumption of alcohol can lower blood pressure in human adults	2	3	5	2
21	Consumption of poultry can lower blood pressure in human adults	1	4	3	3
22	Consumption of sugar can lower blood pressure in human adults	2	2	3	2
23	Consumption of refined grains can lower blood pressure in human adults	2	4	3	3
24	Consumption of salty foods can lower blood pressure in human adults	1	2	5	1
25	Consumption of water can lower blood pressure in human adults	3	5	5	3
26	Consumption of nuts can improve cardiovascular health in human adults	5	4	5	5
27	Consumption of flaxseeds can improve cardiovascular health in human adults	5	5	5	3
28	Consumption of green leafy vegetables can improve cardiovascular health in human adults	5	5	5	5
29	Consumption of fruits can improve cardiovascular health in human adults	5	4	5	5
30	Consumption of berries can improve cardiovascular health in human adults	5	5	5	5
31	Consumption of bananas can improve cardiovascular health in human adults	5	3	5	2
32	Consumption of beetroot can improve cardiovascular health in human adults	5	4	5	5
33	Consumption of avocados can improve cardiovascular health in human adults	5	5	5	3
34	Consumption of legumes can improve cardiovascular health in human adults	5	5	5	5
35	Consumption of beans can improve cardiovascular health in human adults	5	5	5	5
36	Consumption of whole grains, compared to standard diet with refined grains, can improve cardiovascular health in human adults	5	4	5	3
37	Consumption of low-fat dairy products can improve cardiovascular health in human adults	2	4	4	3
38	Consumption of fish can improve cardiovascular health in human adults	2	5	5	3
39	Consumption of dark chocolate can improve cardiovascular health in human adults	4	4	3	4
40	Consumption of eggs can improve cardiovascular health in human adults	1	3	3	3
41	Consumption of red meat can improve cardiovascular health in human adults	1	2	3	3
42	Consumption of caffeine can improve cardiovascular health in human adults	3	3	5	3
43	Consumption of full-fat dairy can improve cardiovascular health in human adults	1	3	4	4
44	Consumption of cheese can improve cardiovascular health in human adults	1	4	3	3
45	Consumption of alcohol can improve cardiovascular health in human adults	2	4	5	2
46	Consumption of poultry can improve cardiovascular health in human adults	1	5	5	2
47	Consumption of sugar can improve cardiovascular health in human adults	2	1	5	1
48	Consumption of refined grains can improve cardiovascular health in human adults	2	2	5	3
49	Consumption of salty foods can improve cardiovascular health in human adults	1	4	5	1
50	Consumption of water can improve cardiovascular health in human adults	3	5	5	3

Table 5: Claim Dataset with Annotated Human Labels and Predicted Labels

B Criteria prompt engineering

Each criterion includes two prompts.

1. The first prompt gathers written responses relevant to the criterion. it uses the title and context as context for the generation of the written response.
2. The second prompt uses that written answer as "response" to generate a YES or NO answer, making it suitable for the inclusion-criteria pipeline

B.1 Study Design Criterion

The goal is to determine the type of study (interventional, observational, or review) using the title or context. If the context is based on multiple studies, it's likely a review. Use a short but informative sentence to state the type of study. Only final output is expected.

```
title: {title}
context: {context}
Your goal is to determine what type of study the context is based,
in other words is it a type of interventional study, observational study or review.
First use the title to determine the type of study, if that is not possible use the context.
When using the context, first determine if this context is based on multiple studies,
if so it is probably a Review.
Answer using a short but informative sentence
***only provide final output, NOTHING ELSE***
```

```
Response: {response}
Based on the given response, answer this question:
Is the type of study design used a Interventional study?
Answer only "YES" or "NO".
Make sure it is not a review or meta-analysis of Interventional studies
```

B.2 Population Criterion

Analyze the context for participant details, such as age, number, selection method, and other conditions. Provide a summary of the participants' data using the given format.

```
context: {context}
Analyze the context on the participants of the study, mention age, number of participants,
how participants or subjects were selected, and any other conditions.
Summarize in the following format in short sentences:
(1) Name of the study
(2) participants details (age, etc)
(3) sample size
(4) selection method
(5) other conditions
***only provide final output, NOTHING ELSE***
```

```
Response: {response}
Based on the given response, answer this question:
Is the type of study population based on humans and clearly stated?
Answer only "YES" or "NO".
```

B.3 Intervention Criterion

Provide a single sentence describing the intervention in the given study.

```
context: {context}
The given context is 1 study, give a single sentence on the intervention that was conducted.
***only provide final output***
```

```
Claim: {claim}
Response: {response}
Based on the given response, answer this question:
Does the intervention involve a dietary requirement
or restriction and does it answer the claim?
Answer only "YES" or "NO".
```

B.4 Comparison Criterion

Summarize the details of the intervention and control groups in a single sentence.

```
context: {context}
Give details on intervention group and control group.
Answer in a single sentence
***only provide final output***
```

```
Response: {response}
Based on the given response, answer this question:
Are intervention and control group mentioned?
Answer only "YES" or "NO".
```

B.5 Outcome Criterion

Analyze and summarize the results, outcome measures, and conclusions of the study using short sentences.

```
context: {context}
Analyze the context on results, outcome measures and respective values, and conclusion.
Summarize in the following format in very short sentences:
(1) results
(2) outcome measures
(3) conclusion
***only provide final output, NOTHING ELSE***
```

```
Response: {response}
Based on the given response, answer this question:
Are reported outcomes clearly stated?
Answer only "YES" or "NO".
```

C Summary Prompt Engineering

C.1 Final Summary Prompt

Summarize multiple studies into a single, verbose summary covering study design, population, intervention, comparison, and reported outcome.

```
Context: {context}
The provided context are multiple studies. Summarize the multiple studies into a single
verbose summary on the following topics:
(1) study design
(2) population
(3) intervention
(4) comparison
(5) reported outcome
```

D Verdict Prompt Engineering

```
summary: {summary}
claim: {claim}

Given the summary on the claim, it is your job to answer the claim using the summary and give advice.
Repeat the claim and answer it using the summary.
```

E Label Prediction Prompt Engineering

```
Extract the desired label from the following verdict following the properties mentioned in schema and
add the description of the label.

verdict: {verdict}
```

F Classification Report for Model Evaluation

Class	Precision	Recall	F1-Score	Support
1	0.00	0.00	0.00	12
2	0.40	0.20	0.27	10
3	0.12	0.25	0.17	4
4	0.10	1.00	0.17	2
5	0.60	0.41	0.49	22
Accuracy			0.28	50
Macro Avg	0.24	0.37	0.22	50
Weighted Avg	0.36	0.28	0.29	50

Table 6: Classification Report BaseLLM Model

Class	Precision	Recall	F1-Score	Support
1	0.00	0.00	0.00	12
2	0.00	0.00	0.00	10
3	0.09	0.25	0.13	4
4	0.00	0.00	0.00	2
5	0.58	0.95	0.72	22
Accuracy			0.44	50
Macro Avg	0.13	0.24	0.17	50
Weighted Avg	0.26	0.44	0.33	50

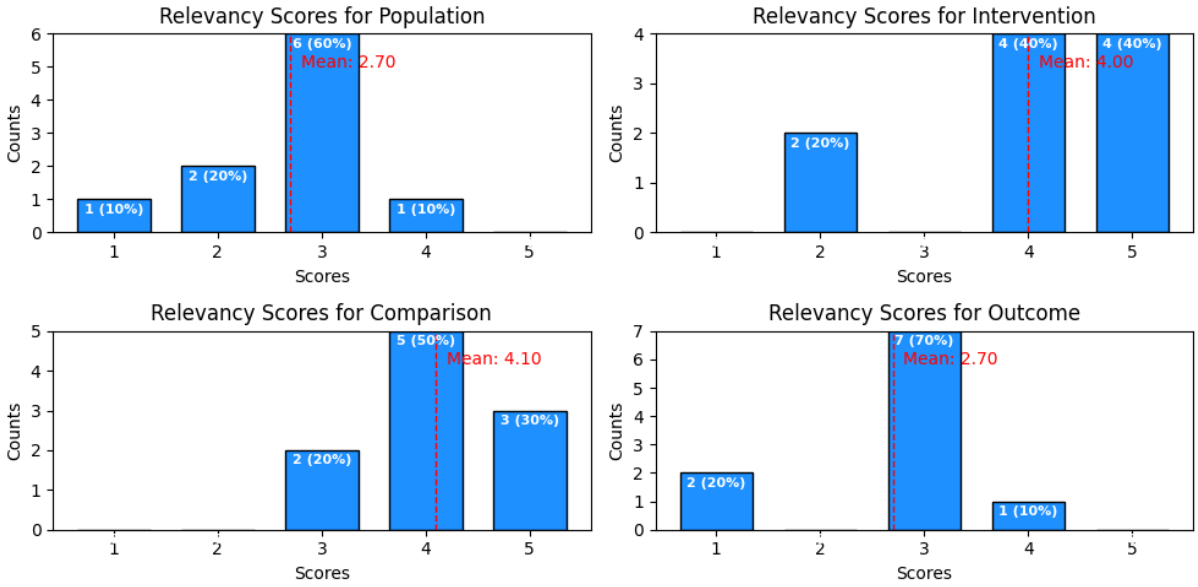
Table 7: Classification Report Standard RAG-LLM Model

Class	Precision	Recall	F1-Score	Support
1	0.67	0.17	0.27	12
2	0.38	0.30	0.33	10
3	0.13	0.75	0.22	4
4	0.50	1.00	0.67	2
5	0.92	0.50	0.65	22
Accuracy			0.42	50
Macro Avg	0.52	0.54	0.43	50
Weighted Avg	0.67	0.42	0.46	50

Table 8: Classification Report Advanced RAG-LLM

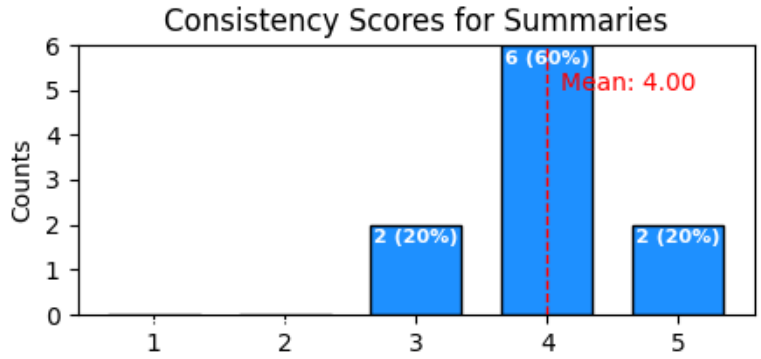
G Survey Responses Scores

G.1 Distribution of Accuracy and Relevancy Scores Across Different PICO Elements

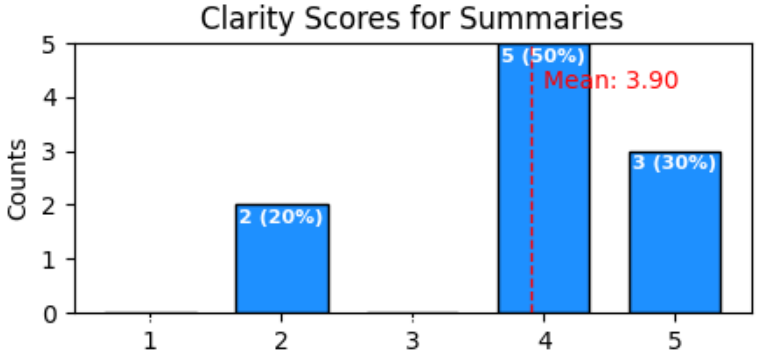


(a) Relevancy histograms

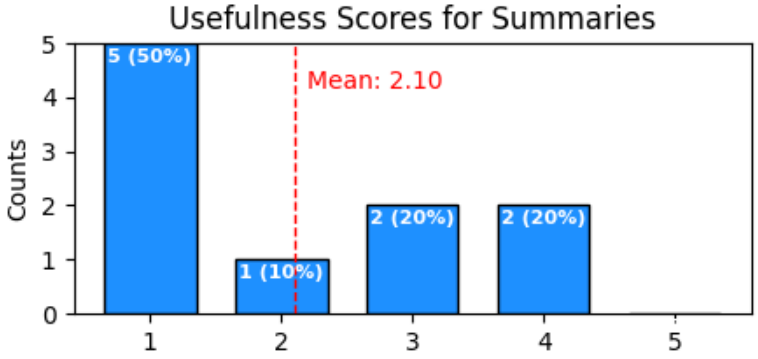
G.2 Distribution Consistency, Clarity, and Usefulness scores of Summaries



(a) Consistency histograms



(b) Clarity histograms



(c) Usefulness histograms

H Claim Analysis

H.1 Evidence Used to Verify Claim 6: Consumption of Bananas Can Lower Blood Pressure in Human Adults

Claim	Consumption of bananas can lower blood pressure in human adults.
Verdict	Based on the summarized studies, there is no direct evidence to support the claim that consumption of bananas can lower blood pressure in human adults. The studies did not specifically investigate the effects of banana consumption on blood pressure. However, it is worth noting that one of the studies (Study 2) found that acute supplementation with nitrate-rich beetroot juice reduced systolic blood pressure (SBP). While this is not directly related to banana consumption, it suggests that increasing dietary intake of nitrates may have a beneficial effect on blood pressure. While there is no conclusive evidence to support the specific claim about bananas, incorporating a variety of fruits and vegetables into your diet, including bananas, can still be part of a healthy lifestyle. A balanced diet with plenty of whole foods like fruits, vegetables, and whole grains can help support overall health and well-being.
Summary	The present investigation encompasses a trio of interventional trials, which aimed to examine the effects of various dietary interventions on blood pressure, vascular function, and cardiovascular risk factors. The first study recruited middle-aged individuals with prehypertension (84%) or stage 1 hypertension (16%), who were selected based on their blood pressure status and baseline reactive hyperemia index (RHI) scores. The second study included healthy normotensive younger adults (ages 18-60 years) and older adults (age $\geq$ 60 years). The third study enrolled individuals with type 1 and type 2 diabetes. In the first study, participants consumed either low-fat dairy products or nondairy products as part of their daily diet for eight weeks. In the second study, healthy adults received acute supplementation with nitrate-rich beetroot juice (12.8 mmol of NO3-) or a placebo (0.08 mmol of NO3-). The third study involved increasing consumption of fruit, vegetables, and dairy products by approximately one-third of a serve above habitual intake for the intervention group. The first study compared the effects of consuming low-fat dairy products versus nondairy products on blood pressure and vascular function. In the second study, participants received either beetroot juice supplementation or a placebo. The third study contrasted the outcomes of increasing fruit, vegetable, and dairy consumption with continuing one's usual diet. The first study found no significant effects on blood pressure or vascular function when consuming low-fat dairy products compared to nondairy products, but did observe improved endothelial function in subjects with baseline dysfunction. The second study reported increased plasma [NO2-] and [NO3-], reduced systolic blood pressure (SBP), and improved reaction time in the Stroop test following acute supplementation with nitrate-rich beetroot juice. The third study found no significant improvement in cardiovascular health outcomes, including carotid-femoral pulse wave velocity (cfPWV) and augmentation index (AIx), despite increasing fruit, vegetable, and dairy consumption.

Table 9: Generated Summary and Verdict Used to Verify the Banana Claim

Paper Title	Summary
Paper 1: ***Effects of low-fat dairy intake on blood pressure, endothelial function, and lipoprotein lipids in subjects with prehypertension or stage 1 hypertension***	The study investigated the effects of consuming low-fat dairy products versus nondairy products on blood pressure and vascular function in individuals with prehypertension or stage 1 hypertension. The eight-week intervention involved participants ingesting either low-fat dairy products or nondairy products as part of their daily diet. Notably, no significant effects were observed on blood pressure or vascular function between the two groups. However, a remarkable finding emerged: for subjects with baseline endothelial dysfunction, premeal reactive hyperemia indices (RHI) significantly improved when consuming low-fat dairy products compared to nondairy products. Overall, the study concluded that consuming low-fat dairy products did not substantially impact blood pressure or vascular function in the studied population; yet, further research is warranted to explore the potential benefits of dairy consumption for enhancing endothelial function in individuals with existing endothelial dysfunction.
Paper 2: ***Acute Supplementation with Nitrate-Rich Beetroot Juice Causes a Greater Increase in Plasma Nitrite and Reduction in Blood Pressure of Older Compared to Younger Adults***	In this groundbreaking study, researchers conducted an interventional trial to investigate the effects of acute supplementation with nitrate-rich beetroot juice (BR) on blood pressure, cognitive performance, and mood in healthy normotensive adults from two age groups: younger adults (18-60 years) and older adults ($\geq$ 60 years). The intervention consisted of a single dose of BR juice containing 12.8 mmol of NO3-, which was administered to the intervention group consisting of 40 participants. This was compared to a control group of 20 participants who received a placebo containing only 0.08 mmol of NO3-. The study revealed that acute supplementation with nitrate-rich BR juice significantly reduced systolic blood pressure (SBP) by 4.8 mmHg and diastolic blood pressure (DBP) by 6.5 mmHg in older adults, demonstrating a notable decrease in blood pressure levels. Furthermore, the results showed an improvement in reaction time in the Stroop test, indicating enhanced cognitive performance. In conclusion, this study provides compelling evidence that acute supplementation with nitrate-rich BR juice can lead to a reduction in blood pressure and improved aspects of cognitive performance, potentially benefiting both younger and older adults.
Paper 3: ***"Effect of Improving Dietary Quality on Arterial Stiffness in Subjects with Type 1 and Type 2 Diabetes: A 12 Months Randomised Controlled Trial "***	A dietary intervention study aimed to enhance cardiovascular risk factors among individuals afflicted with both Type 1 and Type 2 diabetes was conducted. The primary objective was to assess the efficacy of increasing fruit, vegetable, and dairy consumption by approximately one-third of a serving above habitual intake. Regrettably, no significant improvement in central blood pressure (cBP) or augmentation index (AIx) was observed among participants with either type of diabetes. While peripheral diastolic blood pressure showed some marginal benefit, the overall outcome was disappointing, suggesting that this specific dietary intervention did not significantly improve cardiovascular health outcomes for individuals living with Type 1 and Type 2 diabetes.

Table 10: Summary of 3 Research Papers, used as evidence for verifying the banana claim

H.2 Evidence Used to Verify Claim 18: Consumption of full-fat dairy can lower blood pressure in human adults

Claim	Consumption of full-fat dairy can lower blood pressure in human adults.
Verdict	Supported by the evidence. According to the study, one study found that dairy consumption reduced mean daytime systolic blood pressure by 5 mm Hg. This suggests that consuming full-fat dairy products may have a positive impact on blood pressure levels in human adults. However, it's important to note that this finding is specific to dairy consumption and does not necessarily generalize to all types of dairy products or their fat content. Additional advice: • While the evidence suggests that dairy consumption can lower blood pressure, more research is needed to fully understand the mechanisms behind this effect. • It's unclear whether the beneficial effects on blood pressure are due to specific components of full-fat dairy products (e.g., calcium, protein) or their overall nutritional profile. • Individuals looking to incorporate dairy into their diet for blood pressure management should consider a balanced approach that includes a variety of foods, including low-fat and fat-free options. • Further research is needed to confirm the findings and explore potential limitations or interactions with other dietary factors. Overall, while the evidence suggests a positive impact of dairy consumption on blood pressure, more research is necessary to fully understand the relationship between full-fat dairy and blood pressure levels in human adults.
Summary	The present analysis is a compilation of interventional studies, including "Overweight and obese normolipidemic adolescent males who were habitual soda drinkers did not show further improvement in lipid and lipoprotein measurements in response to short-term isocaloric replacement of sugar-sweetened soda with reduced fat milk," "Effects of low-fat dairy intake on blood pressure and vascular function in individuals with prehypertension or stage 1 hypertension," and "The effects of dairy foods or calcium/vitamin D supplements, as complements to hypocaloric diets, may protect against or improve cardiometabolic diseases risks in overweight/obese postmenopausal women." Additionally, the study design also includes another interventional clinical trial investigating the impact of dairy consumption on essential hypertension. The studies included participants with various demographics. Specifically, they involved overweight and obese early postmenopausal women, as well as overweight/obese normolipidemic adolescent males who were habitual soda drinkers. The population also consisted of men and women with prehypertension or stage 1 hypertension. The interventions varied across the studies. In one study, participants were randomly assigned to one of three groups: a control group receiving only hypocaloric diets (C), a group receiving hypocaloric diets with calcium and vitamin D supplements (S), or a group receiving hypocaloric diets with low-fat dairy foods (D). In another study, the intervention consisted of having participants consume either a dairy-rich diet (DAIRY) or a control diet (CONTROL), with each dietary phase lasting for 12 weeks. A third study involved short-term isocaloric replacement of sugar-sweetened soda with reduced fat milk. The studies included various comparison groups. In one study, the S group received hypocaloric diets with calcium and vitamin D supplements, while the D group received low-fat dairy foods, and the C group only received hypocaloric diets without any supplements or dairy foods. In another study, participants who consumed a dairy-rich diet (DAIRY) were compared to those on a control diet (CONTROL). A third study involved comparing participants who consumed sugar-sweetened soda with those who replaced it with reduced-fat milk. The studies reported various outcomes. Some of the key findings include that serum adiponectin concentrations increased significantly in all groups, and hs-CRP values stayed above the acceptable normal range. Additionally, insulin levels decreased in certain groups, while 25(OH)D concentrations increased across all participants. One study found that dairy consumption reduced mean daytime systolic blood pressure by 5 mm Hg, while another study showed no significant changes in measures of vascular function or lipid variables. Overall, the studies suggest that dairy consumption may have a positive impact on cardiometabolic risk factors and essential hypertension, although more research is needed to fully understand these findings.

Table 11: Generated Summary and Verdict Used to Verify the Full-Fat Dairy Claim

Paper Title	Summary
Paper 1: ***Effects of high and low fat dairy food on cardio-metabolic risk factors: a meta-analysis of randomized studies***	The study, comprising a meta-analysis of interventional studies, investigated the effects of increasing dairy food consumption on various cardiovascular and metabolic risk factors. The diverse population, spanning a wide age range, was recruited from randomized dietary intervention trials, with the majority being women. A total of 14 studies were included in the analysis. The key finding is that boosting dairy food intake has minimal or no significant impact on major health indicators, including weight, waist circumference, insulin resistance, and LDL-cholesterol levels. The only notable change was a modest increase in weight. Based on these findings, it can be reasonably concluded that incorporating both low-fat and whole-fat dairy products into one's diet is suitable for most healthy individuals.
Paper 2: ***A Randomized Study of the Effect of Replacing Sugar-Sweetened Soda by Reduced Fat Milk on Cardiometabolic Health in Male Adolescent Soda Drinkers***	The investigation involved an interventional design, where overweight and obese normolipidemic adolescent males (above the 75th percentile for age and sex) were randomly assigned to replace their usual sugar-sweetened beverages with reduced-fat milk for a short-term period. The objective of this isocaloric replacement intervention was to assess its impact on various cardiometabolic outcomes. In conclusion, the results suggest that replacing sugar-sweetened soda with reduced-fat milk may have potential benefits for cardiovascular health, as evidenced by decreases in systolic blood pressure and serum uric acid concentrations. Additionally, the levels of glycosphingolipids (LacCer and GluCer) were found to be lower following the intervention. While these findings are promising, they require further confirmation through future studies.
Paper 3: ***Effects of low-fat dairy intake on blood pressure, endothelial function, and lipoprotein lipids in subjects with prehypertension or stage 1 hypertension***	This interventional trial investigated the effects of incorporating low-fat dairy products into one's daily diet on blood pressure and vascular function among individuals with prehypertension or stage 1 hypertension. The study, which spanned eight weeks, compared the outcomes of two groups: an experimental group consisting of 30 participants who consumed low-fat dairy products such as milk, yogurt, and cheese, and a control group comprising 30 participants who substituted these dairy products with nondairy alternatives. The findings indicate that consuming low-fat dairy products had no significant impact on blood pressure, measures of vascular function, or lipid variables among the study population. In conclusion, the trial reveals that incorporating low-fat dairy products into one's diet does not have a discernible effect on cardiovascular health markers among individuals with prehypertension or stage 1 hypertension.
Paper 4: ***Cardiometabolic Indices after Weight Loss with Calcium or Dairy Foods: Secondary Analyses from a Randomized Trial with Overweight/Obese Postmenopausal Women***	In this interventional study, a carefully curated population of overweight/obese early postmenopausal women were randomly assigned to one of three groups: a control group receiving only hypocaloric diets (C), a group receiving hypocaloric diets with calcium and vitamin D supplements (S), or a group receiving hypocaloric diets with low-fat dairy foods (D). The purpose of this study was to investigate the effects of these interventions on cardiometabolic disease risks. Results showed that participants in groups C and S experienced significant decreases in insulin levels, while all participants saw increases in serum 25-hydroxyvitamin D (25(OH)D) concentrations. These findings support the notion that dairy foods or calcium/vitamin D supplements, when combined with hypocaloric diets, may serve as effective complements for mitigating cardiometabolic disease risks in this population.
Paper 5: ***Impact of dairy consumption on essential hypertension: a clinical study***	In this clinical trial researchers investigated the impact of dairy consumption on essential hypertension. The study's intervention consisted of having participants adhere to either a dairy-rich diet (DAIRY) or a control diet (CONTROL) for 12 weeks each. The DAIRY diet featured dairy products rich in calcium and protein, while the CONTROL diet was devoid of such products. Notably, this dietary intervention resulted in a significant reduction of mean daytime systolic blood pressure by an average of 5 mm Hg, underscoring the potential benefits of incorporating dairy products into one's diet for managing essential hypertension.

Table 12: Summary of 5 Research Papers, Used as Evidence For Verifying the Full-Fat Dairy Claim

I Survey Responses

	Population			Intervention			Comparison			Outcome			Summary			
Eval	Acc.	Rel.	Explanations/Comments	Acc.	Rel.	Expl/Comments	Acc.	Rel.	Expl/Comments	Acc.	Rel.	Explanations/Comments	Consis.	Clar.	Use.	Explanations/Comments
Claim 1	3	3	Missing information on selection method for obese population, sample size, and exclusion criteria	4	4		4	4		2	3	one serving did not lead to many changes, except an increase in HDL particle size.	3	5	1	This study did not primarily focus on the relationship between blood pressure and consumption of berries. No effects of treatment were noted for systolic and diastolic pressure.
Claim 1	2	2	Missing information on sample size and exact numbers on cholestorol levels of population	4	4	Comment on adding additional exact dosages of intervention	4	3	Commont on additional information of equal calorig intake	2	3	Comment on missing measurement levels	3	4	2	Additional data on blood pressure would be essential. The study should include whether blood pressure was measured as an outcome.
Claim 2	2	3	Mean age and exact smaple size is not correct. Information on exact numbers is incorrect.	2	2	7 week intervention, not 8. Intervention of eating kiwis before breakfast	4	4		1	1	There was a significant decrease in systolic blood pressure for the intervention group compared with the control group.	4	2	1	Claim can not be made based on this study. Study focusses on kiwi's only.
Claim 2	1	1	Original paper states mean age and sampl size, but passage does not	3	2	That they substituted kiwi's for a part of their regular breakfast rather than add it on top.	3	3	Suggest additional information on contro group	1	1	Main conclusion should be no difference in blood glucose levels	5	2	1	The summary does not include decrease of systolic pressure in the intervention group
Claim 3	3	3	Missing exact numbers on sample size. Also diet (cocoa intake) hormone treatment and medication were selection criteria. And the participants were recruited from multiple health centers.	5	5		4	4		2	3	Exact data on blood pressure is missing	4	4	1	Blood pressure is left out of the results of the summary, which is strange because it's one of the main outcomes
Claim 3	2	4	The quantity (n) of the Sample (=73) and control group (=67) do not match with the n passage, respectively. Age group	5	5		5	5		2	3	Blood Pressure is not mentioned	4	5	1	Daily consumption of 10g dark chocolate does not bring any significant improvement to lowering blood pressure.
Claim 4	2	3	Sample size incorrect, missing additional information in selection method	5	5		5	5		2	4	Primary outcomes measure missing	5	5	4	Consumption of legumes can lower blood pressure in human adults with obesity and diabetes
Claim 4	2	3	Incorrect sample size	5	4		4	5		2	3	Outcome measurements relating to blood pressure levels missing in summary	4	4	3	Consumption of legumes replacing red meat can lower blood pressure in obese diabetic human adults. Exact measurements missing and lower blood pressure in specific population.
Claim 5	1	2	Sample size and selection method, there was no elevated blood pressure in study population	4	4		4	4		2	3	Missing exact numbers on exact blood levels	4	4	4	Missing sample size and exact blood pressure measurements.
Claim 5	2	3	Sample size and information on age	5	5	Additional information on what types of fish would add information	4	4		3	3	Variability in Results: If the study noted any variability in responses among participants or any specific subgroups that had different outcomes, that could be relevant to mention.	4	4	3	Include specific data on blood pressure readings or other relevant cardiovascular metrics that might have been assessed.
Mean	2	2.7		4.2	4		4.1	4.1		1.9	2.7		4	3.9	2.1	