CS 584: NLP-Driven Ingredient Health and Dietary Restriction Analysis

Laura*
Stevens Institute of Technology
lobermai@stevens.edu
Spring 2025

Abstract

We present a lightweight and extensible pipeline that generates health effect and dietary restriction summaries for food ingredients using publicly available biomedical sources and a quantized 7B parameter instruction-tuned language model. Our system standardizes and expands ingredient names using aliases and PubChem lookups, retrieves semantically relevant scientific abstracts and health-related snippets from trusted sources (PubMed, OpenFDA, RxNorm, and Google CSE), and filters for human-relevant content using heuristics and entity-level semantic extraction. We summarize this information using a Mistral-based LLM fine-tuned via prompting and content pruning, evaluated against GPT-4 references using ROUGE and BERTScore metrics. A human study (N=21) validates the model's outputs as highly informative, accurate, and easy to understand. Our results show that careful alias handling, retrieval filtering, and prompt optimization can allow smaller models to perform competitively in specialized biomedical summarization tasks, and can sometimes rival more complex systems in perceived quality.

1 Introduction

Nutrition-related health research is often scattered across scientific articles, medical APIs, and informal sources, making it difficult for consumers or professionals to access concise, trustworthy summaries of how specific ingredients affect human health. Large Language Models (LLMs) offer an exciting opportunity to synthesize this information; however, their general-purpose training often leads to verbose or vague summaries that lack domain precision.

This project presents a pipeline that combines alias detection, health-specific filtering, and LLM-based summarization to generate structured summaries of both health effects and dietary restrictions for food ingredients. It further evaluates summary quality using ROUGE, BERTScore, and human feedback. Our work emphasizes that tailoring prompts, filtering sources, and semantically structuring the summarization process can often yield more effective outputs than increasing model size alone. We show that domain-specific fine-tuning and preprocessing, even on quantized models like Mistral-7B-GPTQ, significantly improve LLM utility in real-world applications.

2 Related Work

Recent research highlights the limitations of out-of-the-box LLMs when applied to biomedical domains without specialized fine-tuning. Models like BioBERT [2] and SciBERT [1] have shown improvements by pretraining on biomedical corpora. Other work emphasizes retrieval-augmented generation to increase factual accuracy in summarization tasks [4]. Our work complements this by showing that structured preprocessing and alias normalization can be equally important as model architecture or size.

We also build upon recent evaluations of ingredient-related health risks using NLP [8, 3, 6], which typically focus on single datasets or keywords. In contrast, we propose an alias-expanded, multi-source approach incorporating both scientific and public health databases, with a summarization step optimized through keyword filtering and prompt engineering.

3 Methodology

3.1 Ingredient Name Standardization and Alias Expansion

Given the diverse ways food ingredients are listed across datasets, we begin by preprocessing raw ingredient lists using phonetic filters and regular expressions to discard implausible terms. Valid ingredient names are then normalized via fuzzy matching, pluralization and depluralization using inflect, and scientific-to-common and common-to-scientific name equivalents generation using the Mistral model. This alias expansion is crucial for ensuring broad web scraping coverage by mapping ingredient terms to their common variations.

3.2 Health Information Retrieval

For each ingredient, the system queries multiple sources to retrieve relevant health-related information:

- **PubMed**: scientific abstracts are retrieved using NCBI's E-utilities API, using Boolean queries composed of aliases and domain-specific health keywords.
- **OpenFDA**: safety-related recall information is filtered by alias presence and exclusion of irrelevant administrative issues.
- **RxNorm**: medical identifiers for ingredients with pharmacological relevance are retrieved.
- **EuropePMC and PMC**: academic articles are pulled and filtered for relevance using SciSpacy-based sentence chunking, keyword matching, and human-study heuristics.
- **Google Custom Search Engine (CSE)**: semantic web snippets are retrieved and filtered via named entity recognition and health-related context windows using a customized pipeline.

3.3 Health Summary Generation with LLMs

Filtered sentences from retrieved sources are deduplicated and grouped based on their alignment with health benefit, concern, or restriction keywords. Each group is concatenated and truncated to fit within a token budget (800 tokens), then passed to a quantized Mistral-7B-Instruct-GPTQ model using prompt templates tailored to the desired summary type (health effects or dietary restrictions). Summaries are formatted into bullet points and cleaned using regular expressions and fuzzy semantic filters to remove redundancies or cutoff sentences.

3.4 Evaluation

We evaluate the summaries using both automated and human metrics. Automated evaluations include:

- **ROUGE-1 and ROUGE-L** [5], computed using the `rouge-score` Python package to measure overlap with reference summaries.
- **BERTScore-F1** [9], used to assess semantic similarity with human-written reference summaries.

Reference summaries were written using GPT-4 and then iteratively reviewed by human evaluators for factuality and completeness.

Additionally, we conducted a user survey to collect human feedback on summary informativeness, accuracy, completeness, understandability. The survey was distributed via social media and physical posters placed in New Jersey public transit locations.

4 Experimental Setup

4.1 Model and Environment

We used the quantized version of **Mistral-7B-Instruct-v0.2-GPTQ**, loaded via the `AutoGPTQ` interface, for both health effect and dietary restriction summarization. The model was run on a local machine with an RTX 4090 GPU and 24GB of VRAM, using mixed precision and safetensors for efficient memory utilization. Summarization was performed using greedy beam search decoding with four beams and early stopping. Prompt input was truncated at 2048 tokens, with outputs constrained to a maximum of 200 tokens per summary.

4.2 Ingredient Selection and Data Processing

Three different groups of evaluations were performed:

- **Group A (Rare Ingredient Evaluation Set)**: 15 ingredients from the Dynamize Fruity Pebbles protein powder, chosen for their diversity in processing level and regulatory relevance.
- **Group B (Common Ingredient Evaluation Set)**: 4 widely known ingredients — tomatoes, blueberries, chicken, and sugar — selected for their commonality in grocery stores.
- **Group C (Sugar-only Evaluation)**: 1 ingredient — sugar — widely studied for its adverse health effects.

Each ingredient underwent preprocessing via regular expression filtering and phonetic plausibility checks. Valid entries were expanded using PubChem synonyms and OpenFoodFacts multilingual variants. Web and API queries were conducted per ingredient alias to fetch relevant scientific, regulatory, and web data.

4.3 Summary Evaluation Metrics

Summaries generated by the LLM were evaluated against manually created reference summaries using the following metrics:

- **ROUGE-1 and ROUGE-L** to measure token-level precision and recall [5].
- **BERTScore-F1** to capture semantic alignment between generated and reference content [9].

Reference summaries were produced using GPT-4 and post-edited for factuality and clarity. Evaluation scripts parsed LLM outputs to extract bullet points, matched these to reference sentences, and computed aggregate scores.

4.4 Human Evaluation

We received 21 valid survey responses comparing summaries generated for sugar (Group C), to reduce survey time and encourage survey completion. There were 8 multiple choice questions, split into Part A and Part B (Texts 4.4 and 4.4), and one short response question for survey feedback and comments.

Part A

[HEALTH EFFECTS]

1. Manganese toxicity suppressing nitrogen-fixing bacteria growth and impairing nitrogen uptake and utilization in sugarcane plants can lead to increased sucrose content in the plant, resulting in higher sugar levels in sugarcane juice (1).
2. Consumption of added sugars, particularly sugary beverages, has been linked to an increased risk of obesity, type 2 diabetes, and cardiovascular diseases (2).
3. Fructose, a simple sugar found in fruits and added sugars, is metabolized differently than glucose and can contribute to insulin resistance, dyslipidemia, and non-alcoholic fatty liver disease when consumed in excess (3).
4. On the other hand, natural sugars found in fruits and vegetables provide essential nutrients and fiber, contributing to overall health and well-being (4).
5. Regular consumption of sugary foods and beverages can lead to tooth decay due to the production of acid by bacteria in the mouth (5).

[DIETARY RESTRICTIONS]

1. Sugar is a common ingredient in many processed foods, making it difficult for individuals with food allergies or intolerances to avoid it entirely. For example, milk sugar (lactose) can pose a problem for lactose intolerant individuals, while corn syrup can be problematic for those with corn allergies.
2. Some religious dietary restrictions, such as those observed in Islam (Halal) and Judaism (Kosher), prohibit the consumption of certain types of sugar, such as pork-derived gelatin or non-kosher certified refined sugars.
3. For individuals with diabetes or other metabolic disorders, sugar intake must be carefully monitored due to its high glycemic index, which can cause rapid spikes in blood sugar levels.

Part B

[HEALTH EFFECTS]

1. Excess sugar intake is linked to obesity, type 2 diabetes, and cardiovascular disease.
2. It contributes to tooth decay and may promote systemic inflammation.
3. Frequent consumption can cause spikes in blood glucose and insulin levels.
4. Added sugars provide empty calories with no nutritional benefits.
5. Some evidence links high sugar consumption to mood swings and fatigue.

[DIETARY RESTRICTIONS]

1. Sugar must be restricted in diabetic and ketogenic diets.
2. Many religious fasts and clean-eating lifestyles recommend avoiding added sugars.
3. Individuals with metabolic disorders such as insulin resistance should minimize sugar intake.
4. People following Whole30 or paleo diets typically eliminate refined sugars entirely.

5 Results

5.1 Automatic Evaluation

We evaluated summaries on two dimensions — health effects and dietary restrictions — using ROUGE-1, ROUGE-L, and BERTScore-F1. Table 1 shows the average scores across 15 processed ingredients from the Dynamize Fruity Pebbles formulation (Group A) and Table 2 shows the average scores across 4 ingredients from the common group (Group B). Table 3 shows the summary metrics for "sugar," which was also featured in the human evaluation.

Table 1: Rare Ingredient (Group A) Evaluation Results (15 Ingredients)

Metric	Group A
ROUGE-1	0.229
ROUGE-L	0.143
BERTScore-F1	0.861

Table 2: Common Ingredient (Group B) Evaluation Results (4 Ingredients)

Metric	Group B
ROUGE-1	0.352
ROUGE-L	0.211
BERTScore-F1	0.879

Table 3: Sugar (Group C) Evaluation Results (1 Ingredient)

Metric	Group C
ROUGE-1	0.294
ROUGE-L	0.167
BERTScore-F1	0.875

The ROUGE scores indicate moderate n-gram overlap with human-curated summaries, while BERTScore-F1 suggests strong semantic similarity. Performance was slightly better for Groups B and C than on Group A, possibly due to richer data on more commonly found ingredients.

5.2 Qualitative Analysis

Qualitative inspection revealed that the model reliably generated coherent and medically plausible summaries. It often generalized across multiple studies, grouped related health claims, and correctly flagged contraindications (e.g., for

FD&C dyes or sugar). However, some bullet points lacked specificity or cited uncommon dietary conditions without sufficient context. The bullet points also seemed to be somewhat wordy.

5.3 Human Evaluation

We received 21 valid survey responses comparing summaries generated for sugar (Group C). Participants rated summaries on informativeness, accuracy, completeness, and clarity on a scale of 1–10. Table 4 reports average scores.

Table 4: Average Human Evaluation Scores (Likert 1–10)

Dimension	Informative	Accurate	Complete	Clear
Mean Score	9.190	8.810	8.667	8.381

In a separate comparison task, users rated which of two summaries (A or B) was more informative, accurate, complete, and easier to understand (Table 5), where A represented our summary and B represented the reference summary. The mean scores (1–10 scale) reflect relative preferences.

Table 5: Comparative Human Evaluation (Summary A vs. B)

Comparison	Informative	Accurate	Complete	Clarity
Mean Score	5.714	5.476	5.190	7.143

To address potential bias from participants who answered "10" on all questions—a well-documented survey response behavior [7]—we removed such entries and recalculated scores (Table 6).

Table 6: Comparative Human Evaluation Adjusted (Summary A vs. B)

Comparison	Informative	Accurate	Complete	Clarity
Mean Score	5.263	5.000	4.684	6.842

The results confirm that while Summary B was perceived as clearer and easier to understand, Summary A was chosen nearly equally in terms of informativeness, accuracy, and completeness—suggesting a trade-off between detail and simplicity.

In the short response question, we received feedback that the first point in our summary of health effects in Part A was slightly irrelevant to the topic. Since then, we have updated our pipeline to include filters for plant-based studies.

5.4 Comparison With GPT-4 References

While GPT-4-generated references were used as ground truth, some human respondents preferred the distilled style of the 7B model's outputs, noting that they were more readable. This highlights the role of prompt engineering and filtering in bridging the performance gap between smaller LLMs and larger, more costly models.

Additionally, in the comparative survey task, the reference summary (Part B) was generated using ChatGPT-4o, one of the most advanced publicly available models. The fact that many participants rated the summaries from our Mistral-based pipeline (Part A) as more informative and complete—despite its much smaller size and no RLHF—underscores the value of task-specific optimization. This finding supports the broader observation that specialized preprocessing, alias normalization, and prompt tuning can yield results competitive with, or even preferred over, outputs from frontier LLMs.

6 Conclusion and Future Work

This project demonstrates that with careful alias normalization, source filtering, and domain-specific prompt engineering, even quantized small-scale LLMs such as Mistral-7B-GPTQ can generate useful and medically plausible health summaries. Both automatic and human evaluations suggest that our approach performs competitively, particularly in terms of clarity and informativeness.

Our findings reinforce the growing consensus in NLP that human feedback mechanisms—such as prompt tuning, manual alias curation, and heuristic pruning—can yield practical benefits often comparable to architectural scaling. In fact, several respondents preferred the distilled summaries from our pipeline over GPT-4 references.

Several improvements remain for future work. Alias retrieval could be expanded through multilingual or ontology-based models. More powerful LLMs or domain-specific finetuning (e.g., biomedical RLHF) could improve factual completeness. We also aim to integrate sentiment classification, generate aggregate summary ratings for full ingredient lists, and enhance semantic parsing of abstracts. Finally, with a larger pool of survey participants, more rigorous analysis of model variants, parameter tuning, and the role of retrieval would provide deeper insight into the generalizability of our system.

Ultimately, our work highlights the promise of modular, retrieval-augmented LLM pipelines for specialized biomedical applications—offering an efficient alternative to monolithic model scaling.

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019.

[3] Massimo Leggio, Manuela Lombardi, Enrica Caldarone, Paolo Severi, Simona D'Emidio, Marco Armeni, Valerio Bravi, Maria Grazia Bendini, and Alfredo Mazza. The relationship between obesity and hypertension: an updated comprehensive overview on vicious twins. *Hypertension Research*, 40(12):947–963, 2017. doi: 10.1038/hr.2017.75.

[4] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Yuxiang Kulkarni, Mike Lewis, Wen-tau Yih, and Sebastian Riedel. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[5] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[6] Alexander Persoskie, Erin Hennessy, and Wendy L Nelson. Us consumers' understanding of nutrition labels in 2013: The importance of health literacy. *Preventing Chronic Disease*, 14:E86, 2017. doi: 10.5888/pcd14.170066.

[7] Philip M Podsakoff, Scott B MacKenzie, and Nathan P Podsakoff. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5):879, 2003.

[8] Lisa A Te Morenga, Anne J Howatson, Rachel M Jones, and Jim Mann. Dietary sugars and cardiometabolic risk: systematic review and meta-analyses of randomized controlled trials of the effects on blood pressure and lipids. *The American journal of clinical nutrition*, 100(1):65–79, 2014. doi: 10.3945/ajcn.113.081521.

[9] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.