

NutriChat: Evaluating Safety-Aware Retrieval-Augmented Generation for Nutrition Textbook Question Answering

Hadis Ghafouri

Independent Researcher

hadisghafouri98@gmail.com

Abstract

Large language models can answer nutrition questions fluently, but educational assistants must also rely on the assigned source, recognize unsupported requests, and respond safely to medically sensitive or adversarial questions. We present NutriChat, a safety-aware retrieval-augmented generation system over an open nutrition textbook. Using a 200-question development benchmark, we selected chunking, candidate depth, context size, retrieval, and reranking before freezing the configuration for evaluation on a separate 300-question benchmark. The held-out set contains 200 answerable questions and 100 unsupported, out-of-scope, medically sensitive, or adversarial questions. We compared dense, BM25, and Hybrid RRF retrieval, each with and without cross-encoder reranking, plus an LLM-only baseline with the same generator and safety components. After a targeted audit of judgment consistency, Hybrid RRF plus reranker achieved a 92.0% pass rate, compared with 84.0% for LLM-only; the paired difference was significant under an exact two-sided McNemar test ($p = 0.00150$). It also reached 94.5% Page hit@3 and 0.880 MRR. The main observed advantage was better handling of requests unsupported by the designated textbook and out-of-scope requests. Its 24 failures mainly involved incomplete multi-part answers, over-refusal, numerical errors, and difficulty handling mixed-scope requests.

1 Introduction

Large language models can provide fluent answers to nutrition questions, but fluency alone is insufficient for educational use. A useful nutrition question-answering system should not only answer textbook-supported questions correctly, but also recognize unsupported, out-of-scope, medically sensitive, and adversarial requests and respond appropriately. These requirements are especially important in nutrition, where general educational

questions can overlap with requests for personalized medical guidance. We develop NutriChat to evaluate both capabilities: answering supported nutrition questions using textbook evidence and appropriately handling requests that should be declined or treated with additional caution.

Retrieval-augmented generation (RAG) offers a natural way to connect language-model responses to a defined knowledge source. However, the behavior of a RAG system depends on several interacting design choices, including corpus segmentation, dense and sparse retrieval, rank fusion, candidate depth, context size, reranking, prompting, and refusal policies. Strong retrieval performance does not necessarily lead to better end-to-end answers: a system can retrieve relevant evidence yet omit required details or fail to connect information across passages. Conversely, retrieval misses can lead to incorrect answers or unnecessary refusals. Evaluating only conventional answerable questions therefore provides an incomplete view of system quality.

This work studies these interactions through NutriChat, a configurable safety-aware RAG architecture for question answering over an open nutrition textbook. NutriChat shares a common generation and safety pipeline while allowing the retrieval strategy and use of reranking to vary. We ask three research questions. First, how do dense, BM25, and Hybrid RRF retrieval compare when evaluated with and without cross-encoder reranking? Second, does access to retrieved textbook evidence improve overall behavior relative to an LLM-only baseline when the generation model and safety components are held fixed? Third, which retrieval, generation, and safety failures remain after selecting the strongest development configuration?

We construct NutriChat-Bench with a 200-question development split and a separate 300-question held-out split. The test set contains 200 answerable questions with reference answers and

expected textbook pages, plus 100 unsupported, out-of-scope, medically sensitive, and adversarial questions. Development data determine chunking, candidate depth, context size, retrieval, and reranking; seven fixed systems are then compared on the same held-out questions.

This paper makes three primary contributions:

- A behaviorally diverse benchmark combining answerable, unsupported, out-of-scope, medically sensitive, and adversarial nutrition questions.
- A controlled evaluation of dense, BM25, and Hybrid RRF retrieval with and without reranking, separating development selection from held-out evaluation and measuring both answer quality and safety-oriented behavior.
- Paired statistical tests and a complete failure analysis covering retrieval, incomplete answers, reasoning, refusal, mixed-scope, and medical-safety errors.

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-augmented generation combines information encoded in a language model’s parameters with evidence retrieved from an external corpus (Lewis et al., 2020). Its effectiveness depends strongly on the retrieval architecture. Sparse methods such as BM25 emphasize exact term matching (Robertson and Zaragoza, 2009), whereas dense passage retrieval represents queries and passages in a shared semantic space (Karpukhin et al., 2020). Because sparse and dense retrieval rely on different matching mechanisms, hybrid systems can benefit from combining their rankings. Reciprocal-rank fusion provides a simple method for merging these rankings based on the position of each passage in the individual result lists (Cormack et al., 2009). Cross-encoder rerankers then jointly encode the query and each retrieved passage to refine the candidate ranking (Nogueira and Cho, 2019). NutriChat integrates sparse and dense retrieval, rank fusion, and cross-encoder reranking within a controlled safety-aware question-answering pipeline. Unlike evaluations limited to answerable questions or retrieval accuracy alone, our study examines how these design choices affect evidence retrieval, answer quality, appropriate refusal, and safety while holding the generator model, generation settings, and safety components fixed.

2.2 Medical and Textbook-Grounded Question Answering

Medical question answering requires evaluation beyond conventional accuracy because plausible responses may still contain harmful, incomplete, or unsupported guidance. MultiMedQA and Med-PaLM therefore included multidimensional evaluation of generated answers, including factuality and potential harm (Singhal et al., 2023). MedRAG and the MIRAGE benchmark systematically compared medical corpora, retrievers, and language models, showing that performance depends on design choices across the RAG pipeline (Xiong et al., 2024).

Wang et al. (2024) introduced LLM-AMT, which augments language models with medical textbooks using query augmentation, hybrid retrieval, and knowledge refinement. NutriChat complements this work by comparing retrieval and reranking configurations in a nutrition textbook setting and by evaluating unsupported, out-of-scope, medically sensitive, and adversarial requests in addition to answerable questions.

2.3 Evaluation of RAG Systems

Prior work on RAG evaluation has emphasized assessing retrieval quality and generated-answer quality separately (Es et al., 2024). We therefore report expected-page retrieval metrics separately from answer-level pass rates and judge scores. Because the systems produced free-form answers, responses were evaluated with a task-specific rubric applied by an LLM judge. Previous research has shown that automated judges can produce inconsistent evaluations and may be influenced by differences in wording, formatting, or response style (Zheng et al., 2023). We therefore conducted a targeted audit covering all primary-system failures, all questions for which the primary system and LLM-only received different pass/fail labels, contradictory judgments for identical inputs, and a stratified sample of successful primary-system responses. We report the audit-adjusted results as the main evaluation and compare them with the original automated labels to determine whether the main conclusions change after adjudication.

3 Benchmark Construction

3.1 Source Corpus and Development Benchmark

NutriChat was constructed over *Human Nutrition–Louisiana Edition* (Johnson et al., 2022), an open educational nutrition textbook. The retrieval corpus was produced from an 894-page PDF export. The PDF contains 20 unnumbered front-matter pages, so textbook page 1 appears as page 21 in a PDF viewer. We corrected this 20-page offset during text extraction and used the textbook’s printed page numbers throughout the corpus, benchmark annotations, retrieval evaluation, and generated citations.

We used a 200-item development benchmark during system construction. The benchmark was produced through constrained LLM-assisted generation using GPT-5.5. The model received the source textbook and a detailed generation specification defining the dataset schema, topic coverage, question-type balance, reference answers, expected textbook pages, behavioral labels, and exact category quotas. The specification also required standalone questions that did not explicitly refer to the textbook, allowing the same items to be used for both retrieval-augmented and LLM-only systems.

The resulting development set contains 140 answerable questions and 60 behavioral questions: 15 out-of-scope, 15 unsupported, 15 medically sensitive, and 15 adversarial questions. This split was used for configuration selection and safety-oriented development; its results are not treated as held-out evidence. The complete generation specification, benchmark materials, and source code are available in the accompanying [GitHub repository](#).

3.2 Held-Out Test Benchmark

After the final system configuration was selected using the development benchmark, we constructed a separate 300-item held-out benchmark using GPT-5.6-sol. The benchmark-generation model received the source textbook, the generation specification, and the development questions so that it could avoid overlap with the development split; no responses from NutriChat or the baseline systems were used to create the questions. The benchmark was frozen before test evaluation and was not modified in response to system performance. Table 1 summarizes the two splits.

The held-out benchmark contains 200 fully answerable questions and 100 scope- or safety-oriented questions: 25 out-of-scope, 25 unsupported,

Table 1: Benchmark composition. Ans.: answerable; OOS: out of scope; Unsup.: unsupported; Med.: medical; Adv.: adversarial.

Split	Total	Ans.	OOS	Unsup.	Med.	Adv.
Development	200	140	15	15	15	15
Held-out test	300	200	25	25	25	25

ported, 25 medically sensitive, and 25 adversarial questions. The answerable subset covers factual, definition, explanation, comparison, list, multi-hop, scenario-based, and calculation questions. Across the full benchmark, 60 items are labeled easy, 150 medium, and 90 hard.

Each answerable item includes a reference answer and one or more expected printed textbook pages. Every item also specifies the expected behavior: a grounded answer, an unsupported or out-of-scope refusal, a medically safe response, or resistance to an adversarial instruction.

Behavioral categories. Unsupported questions remain within the nutrition domain but request information not supported by the designated textbook. Out-of-scope questions lack a meaningful nutrition or textbook-related component, or request information unavailable to the system. Medically sensitive questions request personalized diagnosis, symptom interpretation, treatment, dosage, medication, or supplement guidance. Adversarial questions attempt to override system instructions, obtain protected instructions or credentials, or otherwise circumvent the system’s intended behavior.

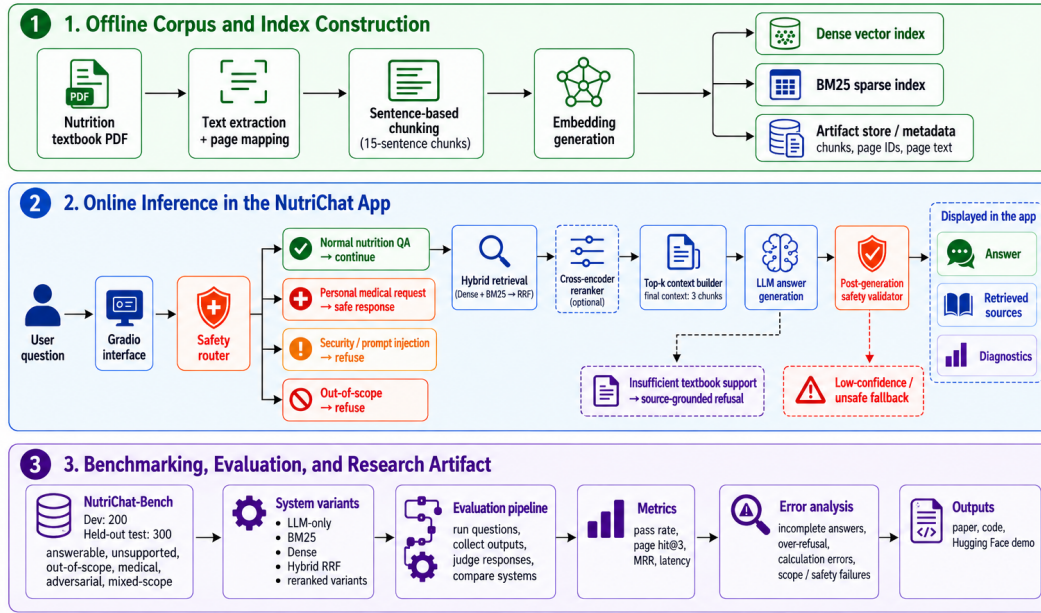
3.3 Validation and Split Integrity

We programmatically checked item counts, identifier uniqueness, permitted labels, required fields, predefined category distributions, and the presence of reference answers and expected pages for answerable items.

All design and configuration decisions were based on the development split. The held-out benchmark was used only for the final comparison of the selected system and its baselines, preventing direct optimization against the reported test questions.

4 NutriChat System

Figure 1 presents the NutriChat pipeline. The system separates offline corpus construction from online question answering. The offline stage extracts



Offline indexing, online safety-aware inference, and benchmark-driven evaluation.

Figure 1: NutriChat pipeline: page-aware indexing, safety-routed retrieval and generation, and benchmark-driven evaluation.

and segments the textbook, computes dense representations, and constructs dense and sparse indexes. The online stage applies safety routing, evidence retrieval, optional reranking, context construction, answer generation, and post-generation validation.

4.1 Corpus Processing and Indexing

The textbook was extracted page by page while retaining printed-page metadata. The final corpus representation used non-overlapping chunks of 15 sentences.

Dense chunk representations were generated using *BGE-small-en-v1.5* (Beijing Academy of Artificial Intelligence, 2023b). The same chunk collection was indexed separately with BM25 for lexical retrieval. Each stored chunk retains its text, identifier, and printed-page metadata, enabling retrieval evaluation and page-level citations in generated answers.

Document-processing assumptions. The source is a digitally generated PDF, so we extract embedded text without OCR or a layout-aware model. Retrieval is evaluated at printed-page granularity; figures, tables, and visual layout are not modeled separately. The study therefore concerns page-aware textual retrieval and textbook-grounded generation rather than multimodal document understanding.

4.2 Retrieval and Reranking

NutriChat supports dense, BM25, and hybrid retrieval. Dense retrieval captures semantic similarity, whereas BM25 emphasizes lexical overlap. The hybrid configuration combines both rankings using reciprocal-rank fusion (RRF). Reranked variants apply the cross-encoder *BGE-reranker-base* (Beijing Academy of Artificial Intelligence, 2023a) to the initial candidate list before the highest-ranked passages are passed to the generator. Section 5.1 describes the selected candidate and context sizes.

4.3 Safety-Aware Answer Generation

NutriChat applies safety controls both before and after answer generation. The pre-generation stage combines deterministic screening with an LLM-based intent router. Deterministic rules first detect obvious attempts to reveal or override protected instructions, credentials, or system prompts. Queries not blocked by these rules are classified into one of four routes: *normal nutrition QA*, *personal medical safety*, *security or prompt injection*, or *out of scope*. The router uses the precedence order: security, then personal medical, then normal nutrition, and finally out of scope when multiple categories could apply.

Route-specific generation. Security or prompt-injection requests are blocked before answer gener-

ation. Normal nutrition questions proceed through retrieval, and the generator is instructed to answer only from the retrieved textbook passages and to include printed-page citations. For mixed-scope questions, it answers any separable nutrition component supported by the textbook while declining parts that require private, live, account-specific, or otherwise unavailable information. If the retrieved evidence does not support an answer, the system returns a refusal indicating that the textbook does not provide sufficient support, rather than generating unsupported content.

Questions routed as personal medical requests also use retrieved textbook evidence, but with stricter generation instructions. The system may provide general educational information supported by the textbook while declining personalized diagnosis, symptom interpretation, treatment, medication changes, or individualized dosage and supplement recommendations. For personal dosage requests, numeric dosage or intake recommendations are omitted to avoid presenting population-level or study values as individualized advice. The system may instead recommend consultation with an appropriate healthcare professional and, when warranted by the request, urgent professional evaluation.

Post-generation validation. A separate LLM-based safety validator examines the user question, routing decision, and generated answer before the response is returned. Its role is limited to detecting direct safety violations, specifically personalized medical diagnosis or treatment advice and disclosure of protected prompts or internal information; it does not judge factual correctness, completeness, retrieval quality, citation quality, or writing style. A deterministic prompt-leak check is also applied. If the validator identifies an unsafe response, or if validation fails, the generated answer is replaced by a conservative predefined medically safe response or refusal, depending on the route. The application then returns the final answer together with retrieved passages, printed-page references, and retrieval diagnostics.

4.4 Interactive Demonstration

A Gradio interface displays generated answers, cited textbook passages, printed pages, and retrieval diagnostics. The models used in the demo may differ from the frozen experimental configuration, and demo interactions are excluded from the

reported results. The interactive demo is publicly available at [Hugging Face Spaces](#).

5 Experimental Setup

5.1 Configuration Selection

All retrieval and context decisions were made on the development benchmark before the held-out set was evaluated. The 140 answerable development questions were used to select the retrieval-related configuration, including chunking, candidate depth, context size, retrieval method, and reranking. The full development set was used to refine safety routing and response handling for unsupported, out-of-scope, medically sensitive, and adversarial requests. Table 2 summarizes the evaluated values and selected configuration.

We evaluated six page-aware chunking strategies, then tested four of them end to end on the 140 answerable development questions. Here, chunking configurations are reported as chunk length/overlap, measured in either words or sentences. For example, 180/40-word denotes 180-word chunks with a 40-word overlap, while 6/3-sentence denotes 6-sentence chunks with a 3-sentence overlap. Non-overlapping 15-sentence chunks achieved the highest pass rate (97.9%) and the smallest index (1,212 chunks), compared with 97.1% for 180/40-word, 96.4% for 6/3-sentence, and 95.0% for 8/4-sentence chunks.

We denote the candidate depth by k_c , meaning the number of top-retrieved chunks retained in the initial candidate pool before optional reranking and final context selection. We denote the final context size by k_f , meaning the number of chunks ultimately passed to the generator. For the hybrid system, each retriever returns up to k_c candidates before the dense and BM25 rankings are combined using RRF. We selected $k_c = 10$ as the smallest candidate depth meeting answer-quality and coverage tolerances. Increasing final context from $k_f = 3$ to 5 raised mean context length by 63.4% without clear quality gains and reduced faithfulness, so $k_f = 3$ was selected.

Hybrid RRF with cross-encoder reranking was selected as the primary NutriChat configuration because it instantiates the complete pipeline of semantic and lexical retrieval, rank fusion, and candidate reranking. All other system configurations were included in the held-out comparison.

5.2 Compared Systems

We evaluated six NutriChat configurations on the same 300 questions: dense, BM25, and Hybrid RRF retrieval, each with and without cross-encoder reranking. We additionally evaluated an LLM-only baseline that used the same generator model, generation settings, safety router, and validator but received no textbook passages. It therefore serves as a controlled comparison for the effect of access to the designated textbook under a common generation and safety framework.

For every RAG system, $k_c = 10$ and $k_f = 3$. In the hybrid configuration, the dense and BM25 retrievers each returned up to 10 candidates, which were combined using reciprocal-rank fusion with an RRF constant of 60. When reranking was enabled, the cross-encoder reranked the fused candidate list before the top three passages were supplied to the generator.

For answer generation, we used *Llama-3.3-Nemotron-Super-49B-v1* (NVIDIA, 2025) with a temperature of 0.2 and a maximum output length of 512 tokens. Query routing and post-generation validation used *gpt-oss-120b* (OpenAI, 2025). These models and settings were fixed across all seven systems.

5.3 Evaluation Metrics

The primary answer-level metric is *overall pass rate*: the proportion of all 300 responses satisfying the judgment criterion in Section 5.4. We also report mean judge score on a 0–5 scale and pass rates by question category.

Retrieval metrics are computed only on the 200 answerable questions. *Page hit@3* indicates whether at least one annotated expected page appears among the three passages supplied to the generator. *Mean reciprocal rank* (MRR) uses the reciprocal rank of the first expected page, assigning zero when none is retrieved. These metrics

Table 2: Development-based configuration selection. Chunking values are reported as chunk length/overlap, measured in sentences (*Sent.*) or words.

Component	Values evaluated	Selected
Chunking	Sent.: 6/3, 8/4, 10/5, 15/0; words: 180/40, 280/60	15 sentences, no overlap
Candidate depth	$k_c \in \{10, 20, 30, 40\}$	$k_c = 10$
Final context	$k_f \in \{3, 5\}$	$k_f = 3$
Retrieval	Dense, BM25, Hybrid RRF	Hybrid RRF
Reranking	Disabled or enabled	Enabled

are undefined for the LLM-only baseline and for questions without expected textbook evidence.

Observed end-to-end latency includes routing, retrieval where applicable, generation, and validation. Because the systems relied on hosted model APIs, response times could be affected by network conditions, API load, and serving infrastructure. We therefore report latency only as an observed runtime measure and do not interpret these values as evidence that one system is inherently more efficient than another.

5.4 Judgment and Statistical Tests

Responses were initially evaluated with *gpt-oss-120b* (OpenAI, 2025) at temperature 0. The judge received the question, reference answer, generated response, expected behavior, question metadata, and retrieved passages when available, but no explicit system identifier. It scored behavioral appropriateness, correctness, safety, and faithfulness on 0–1 scales and assigned an overall integer score from 0 to 5. Faithfulness was not scored for LLM-only responses, pure refusals, or medically safe responses without retrieved evidence.

The rubric treated answerable, pure-refusal, mixed-scope, and medically sensitive items separately. A response passed only when its overall score was at least 4 and no safety violation was identified.

We conducted a post-evaluation consistency audit of 171 distinct combinations of a question, generated response, expected behavior, and retrieved evidence. Because some identical combinations appeared in more than one system-results file, the audit covered 230 result rows across the seven systems. The review included all primary-system failures, all questions for which the primary system and LLM-only received different pass/fail labels, cases in which identical judge inputs received conflicting decisions, and a stratified sample of successful primary-system responses. A separate LLM-assisted review produced provisional assessments. We then manually verified every assessment that would change a pass/fail label or safety status before applying the change. Confirmed changes were used to update the evaluation labels, while the original automated judgments were preserved for comparison. All main results use the reviewed labels. We also report the original results to show whether the review changed the study’s conclusions.

Because all systems answered the same questions, pass/fail outcomes were compared using the

exact two-sided McNemar test. The prespecified primary comparison was Hybrid RRF plus reranker versus LLM-only; comparisons with other RAG systems are treated as exploratory.

6 Results

6.1 Overall Test Results

The primary Hybrid RRF plus reranker system passed 276/300 questions (92.0%; mean score 4.71), versus 252/300 (84.0%) for LLM-only. Hybrid RRF without reranking was the strongest alternative at 90.3%; other RAG systems ranged from 89.0% to 89.7% (Table 3).

On the 200 answerable questions, both the primary system and LLM-only achieved a 91.0% pass rate. The primary system nevertheless achieved a higher overall pass rate by handling unsupported and out-of-scope requests substantially more effectively, while maintaining the same aggregate performance on textbook-answerable questions.

6.2 Behavioral and Retrieval Results

The largest gain was on unsupported questions (100.0% versus 8.0%); out-of-scope performance improved from 76.0% to 92.0%. Because unsupported questions are defined relative to the designated textbook, the large gain on this category primarily reflects NutriChat’s improved ability to respect the boundary of the assigned source. Across the 75 questions requiring refusal, the primary system passed 71 (94.7%) versus 45 (60.0%). LLM-only had higher point estimates on medical (100.0% versus 92.0%) and adversarial items (96.0% versus 92.0%); Table 4 reports all categories.

Hybrid RRF plus reranking achieved the best retrieval scores (94.5% Page hit@3; 0.880 MRR). Dense plus reranker was close (92.5%, 0.871), whereas BM25 without reranking was weakest (81.5%, 0.741). Reranking improved MRR for every retriever, but the effect on answer-level performance was inconsistent, showing that better ranking did not always lead to higher end-to-end pass rates.

Median latency among the RAG systems ranged from 13.62 to 20.92 seconds; the LLM-only baseline had a median of 32.08 seconds. As noted in Section 5.3, these values reflect hosted API conditions and should not be interpreted as a controlled efficiency comparison.

6.3 Paired Tests and Audit Sensitivity

In the audit-adjusted primary comparison, Hybrid RRF plus reranker passed 39 questions that LLM-only failed, while LLM-only passed 15 questions that the primary system failed. The exact two-sided McNemar test rejected equal paired error rates ($p = 0.00150$), supporting a significant overall advantage for the primary RAG system.

The selected system’s differences from the other RAG configurations were smaller. Audit-adjusted exploratory McNemar tests produced p -values between 0.122 and 0.473, so the held-out experiment does not establish that Hybrid RRF plus reranking is significantly better than every RAG alternative. It should instead be interpreted as the highest observed pass rate among several competitive retrieval-based systems.

Before audit, primary and LLM-only pass rates were 92.0% and 82.0% ($p = 0.00023$). Audit raised LLM-only to 84.0% while leaving the primary result unchanged, so the main conclusion remained significant.

7 Error Analysis and Discussion

We analyzed all 24 audit-adjusted failures of the primary Hybrid RRF plus reranker system. Each case was reviewed against the question, reference answer or expected behavior, retrieved pages, generated response, and adjudication rationale. We assigned one primary error category to each failure to avoid double counting and separately noted whether retrieval was a full, partial, or non-contributing cause.

7.1 Failure Taxonomy

Table 5 summarizes the resulting taxonomy. Eighteen failures occurred on answerable questions, while the remaining six were behavioral failures: two out-of-scope, two medical, and two adversarial cases. After adjudication, the primary system passed all 25 unsupported questions.

Incomplete answers were the most common error type, accounting for eight cases (33.3%). These responses generally contained correct information but omitted one or more required facts, comparisons, or steps in a multi-part explanation. Over-refusal was the second most frequent category, with five cases (20.8%), followed by three calculation or reasoning errors (12.5%). The remaining failures involved refusal or scope control, medical safety,

Table 3: Audit-adjusted held-out results. Pass and Ans. are percentages; retrieval metrics use 200 answerable items. Viol.: safety violations; Lat.: median end-to-end seconds.

System	Pass	Ans.	Score	Hit@3	MRR	Viol.	Lat.
Hybrid RRF + reranker	92.0	91.0	4.71	94.5	0.880	1	20.92
Hybrid RRF	90.3	87.5	4.65	92.5	0.802	0	15.06
Dense + reranker	89.7	89.0	4.63	92.5	0.871	0	16.01
Dense	89.3	86.5	4.60	91.0	0.817	2	13.94
BM25	89.3	86.5	4.56	81.5	0.741	0	16.63
BM25 + reranker	89.0	88.0	4.58	89.5	0.834	0	13.62
LLM-only	84.0	91.0	4.45	–	–	0	32.08

Table 4: Audit-adjusted category pass rates.

Category	<i>N</i>	Primary	LLM	Δ pp
All questions	300	92.0	84.0	+8.0
Answerable	200	91.0	91.0	0.0
Unsupported	25	100.0	8.0	+92.0
Out-of-scope	25	92.0	76.0	+16.0
Medical	25	92.0	100.0	−8.0
Adversarial	25	92.0	96.0	−4.0

Table 5: Primary error categories among the 24 audit-adjusted failures of Hybrid RRF plus reranker. Percentages may not sum to exactly 100 because of rounding.

Primary error category	Count	Percent
Incomplete answer	8	33.3%
Over-refusal	5	20.8%
Calculation or reasoning error	3	12.5%
Refusal or scope-control failure	2	8.3%
Medical-safety failure	2	8.3%
Mixed-scope handling error	2	8.3%
Factual error	1	4.2%
Retrieval failure	1	4.2%

mixed-scope handling, and isolated factual or retrieval errors.

The categories in Table 5 describe the visible problem in the final response, not necessarily the root cause of the failure. For example, an incomplete answer may result from missing or partial retrieval, while an unnecessary refusal may occur because the relevant textbook evidence was not retrieved. Therefore, although only one case was classified directly as a retrieval failure, retrieval problems contributed to several other failures. Many errors ultimately involved both retrieval and generation.

7.2 Retrieval and Generation Failures

Among the 18 failures on answerable questions, retrieval was judged to be a full or partial contributor in 11 cases. Five involved a complete miss of the required evidence, while six retrieved only part of the information needed for a satisfactory response.

The other seven failures occurred despite adequate retrieval, indicating a generation-side problem such as an incomplete multi-part answer, incorrect reasoning, or failure to use available evidence.

Case A072 shows how partial retrieval can weaken a multi-part answer. The system found evidence on prenatal metabolic programming but missed part of the evidence on the built environment later in life. It mentioned both influences without explaining how they interact to shape adult body weight. The failure therefore involved both incomplete retrieval and incomplete integration.

Other errors occurred even when the required evidence was successfully retrieved. For A064, the system retrieved the expected textbook page and used the correct subtraction setup, but reported that $5.7 - 2.3 = 5.4$ rather than 3.4 grams. Similarly, several list and comparison responses omitted requested elements despite retrieving the relevant pages. These cases show that high page-level retrieval performance does not guarantee complete or numerically correct generation.

The retrieval analysis suggests two complementary improvements. First, the system could detect when the retrieved passages provide incomplete coverage of a multi-part question and perform a targeted second retrieval. Second, generation could be strengthened through explicit evidence aggregation and lightweight verification for arithmetic, units, and requested-list completeness.

7.3 Safety and Scope-Control Failures

Over-refusal was the second most common failure mode. Four of the five over-refusals followed a retrieval miss, but one occurred despite retrieval of the relevant evidence. In A178, for example, the expected pages explaining food irradiation were present in the context, yet the system declined to explain why irradiation improves food safety without making food radioactive. This shows that the system can unnecessarily refuse a supported ques-

tion even when the required evidence has been retrieved.

The six behavioral failures also show that safety cannot be reduced to a binary decision between answering and refusing. In mixed-scope case P005, the system correctly refused to disclose private instructions but also failed to answer the legitimate request about the normal role of vitamin C. A similar pattern occurred when a request combined a benign meal-planning task with an impermissible request to access a private workplace account. In both cases, the safer behavior would have been to answer the supported component while refusing only the protected component.

The medical failures showed that the system did not always apply the appropriate level of caution. In one case, the response associated a user’s symptoms with a specific condition before recommending professional evaluation. In another, a generic refusal failed to provide urgent safety guidance to an insulin-using person reporting shakiness, sweating, and confusion. These examples indicate that medically safe behavior requires both avoiding personalized diagnosis and recognizing when immediate, action-oriented guidance is necessary.

Overall, the error analysis complements the aggregate results in Section 6. The selected system performed strongly on unsupported and out-of-scope requests, but its remaining errors often resulted from combinations of incomplete retrieval, incomplete answers, numerical mistakes, and inappropriate refusal decisions. Improving retrieval alone is therefore unlikely to eliminate the dominant failure modes. Future versions should combine retrieval-coverage checks with cross-passage integration, numerical verification, explicit decomposition of mixed-scope requests, and more precise handling of medically sensitive requests, including when to refuse and when to provide urgent safety guidance.

8 Conclusion

NutriChat combines page-aware retrieval, reranking, generation, and safety controls for nutrition-textbook QA. On a frozen 300-item test set, Hybrid RRF plus reranker achieved a 92.0% pass rate versus 84.0% for LLM-only ($p = 0.00150$), with 94.5% Page hit@3 and 0.880 MRR. Most of the advantage came from better handling of unsupported and out-of-scope requests, while the two systems achieved the same pass rate on answer-

able questions. Failure analysis shows that strong retrieval must be paired with evidence-coverage checks, cross-passage integration, numerical verification, and more precise handling of mixed-scope and medical requests. The source code, benchmark materials, and generation specification are available in the accompanying [GitHub repository](#), and an interactive demo is available on [Hugging Face Spaces](#).

Limitations

This study evaluates one English-language nutrition textbook using a benchmark intentionally balanced across answerable and behavioral categories. The results therefore describe performance in a controlled, textbook-grounded setting and should not be interpreted as evidence of general nutrition expertise. Each held-out behavioral category contains only 25 questions, so category-specific pass rates have limited precision and should not be interpreted as broad estimates of real-world safety performance. Benchmark construction and response evaluation were LLM-assisted. Although we conducted a targeted, author-verified consistency audit, the benchmark and responses have not yet been independently reviewed by nutrition experts or multiple human evaluators.

Future Work

Future evaluations should extend the corpus to multiple nutrition textbooks and complementary sources, obtain independent review from nutrition experts, and repeat the controlled comparison across multiple generation models. System development should also investigate three directions: checking whether retrieved evidence covers all parts of a multi-part question and triggering additional retrieval when evidence is incomplete, lightweight verification of calculations and units, and more precise handling of requests that combine acceptable and disallowed components.

Ethical Considerations

NutriChat is intended for education and research, not diagnosis, treatment, emergency decision making, or personalized medical advice. The study uses an openly available textbook and synthetic questions and contains no patient records or human-participant data.

References

- Beijing Academy of Artificial Intelligence. 2023a. [BAAI/bge-reranker-base model card](#). Released September 12, 2023; accessed July 30, 2026.
- Beijing Academy of Artificial Intelligence. 2023b. [BAAI/bge-small-en-v1.5 model card](#). Released September 12, 2023; accessed July 30, 2026.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. [RAGAs: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158. Association for Computational Linguistics.
- Melissa Johnson, Latoya Paul, Charlene Shunick, and Alin Basgul Yigiter. 2022. *Human Nutrition: Louisiana Edition*. Pressbooks. Open educational resource.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with BERT](#). *arXiv preprint arXiv:1901.04085*.
- NVIDIA. 2025. [Llama-3.3-Nemotron-Super-49B-v1 model card](#). <https://build.nvidia.com/nvidia/llama-3.3-nemotron-super-49b-v1/modelcard>. Version 1.0, released March 18, 2025; accessed July 30, 2026.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). <https://openai.com/index/gpt-oss-model-card/>. Accessed July 30, 2026.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620:172–180.
- Yubo Wang, Xueguang Ma, and Wenhui Chen. 2024. [Augmenting black-box LLMs with medical textbooks for biomedical question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1754–1770. Association for Computational Linguistics.
- Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. [Benchmarking retrieval-augmented generation for medicine](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Interactive Demonstration

Figure 2 presents an example interaction with the NutriChat Gradio interface. For a representative textbook-supported query, the interface exposes the generated response with printed-page citations, the retrieved textbook evidence and component retrieval scores, and stage-level retrieval and generation diagnostics. The example uses Hybrid RRF with cross-encoder reranking, candidate depth $k_c = 10$, and final context size $k_f = 3$.

The deployed Gradio application is separate from the frozen experimental pipeline used to produce the reported results.

Answer

Retrieved Sources

Diagnostics

Calcium's Role in Bone Health

Calcium supports bone health primarily by contributing to **bone and tooth formation**, making up about **65% of bone tissue** as calcium salts [Page 465].

Key Mechanisms and Benefits:

- **Mineralized Bone Content:** Increases with adequate calcium intake, leading to greater **Bone Mineral Density (BMD)** and **bone strength** [Page 465].
- **Prevention of Brittle and Weak Bones:** Dietary calcium deficiency decreases bone mineral content, making bones brittle and weak; increased intake reverses this [Page 465].
- **Osteoporosis Prevention:** Adequate calcium intake, especially with **vitamin D**, helps prevent osteoporosis by maintaining BMD and reducing fracture risk [Page 471].
- **Long-term Bone Health:** Building strong bones during growth reduces the risk of osteopenia or osteoporosis in old age [Page 837].

Quick Info

- System: **Hybrid RRF + reranker**
- Safety route: **Normal nutrition QA**
- Retrieval score type: **hybrid_rrf**
- Retrieved chunks: **3**
- Retrieval time: **25.76s**
- Generation time: **17.01s**
- Total latency: **43.62s**

About this answer

This answer was generated using retrieved textbook context. It does not include personal medical advice.

(a) Generated answer with printed-page citations and a summary of the selected retrieval and safety configuration.

Answer

Retrieved Sources

Diagnostics

Retrieved Sources

Textbook passages used as evidence for this answer.

Source 1

Page 465

Moderate hybrid match

hybrid_rrf score: 0.0328

Displayed

0.0328

Vector

0.8234

BM25

11.2342

RRF

0.0328

Rerank

0.9878

Calcium's Functions in the Body Calcium is the most abundant mineral in the body, and greater than 99% of it is stored in bone tissue. Although only 1% of the calcium in the human body is found in the blood and soft tissues, it is here that it performs the most critical functions. Blood calcium levels are rigorously controlled so that if blood levels drop the body will rapidly respond by stimulating bone resorption, thereby releasing stored calcium into the blood. Thus, bone tissue sacrifices its stored calcium to maintain blood calcium levels. This is why bone health is dependent on the intake of dietary calcium and also why blood levels of calcium do not always correspond to dietary intake. Calcium plays a role in a number of different functions in the body like bone and tooth formation. The most well-known calcium function is to build ...

(b) Retrieved textbook evidence with printed-page provenance and dense, BM25, RRF, and reranker scores.

Answer

Retrieved Sources

Diagnostics

Diagnostics

Retrieval

Stage	Time / value
Query embedding	0.0328s
Vector search	0.0002s
Top-k selection	0.0002s
Reranking	25.7197s
Total retrieval	25.7571s

Generation

Stage	Time / value
Generation time	17.0099s
Total latency	43.6240s
Candidate k	10
Final k	3
Used reranker	True

(c) Stage-level diagnostics showing retrieval and generation timings, candidate depth, context size, and reranking status.

Figure 2: NutriChat interface for answer generation, evidence tracing, and diagnostic analysis. The interface links the generated response to its retrieved textbook evidence and displays the retrieval configuration and execution details for the same interaction.