Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

NutriQA: A Traceable RAG Question-Answering Benchmark and Empirical Study on Official Chinese Nutrition Data

Yulin Zhao

202334071131@stu.huel.edu.cn

Henan University of Economic and Law https://orcid.org/0009-0005-4623-8465

Research Article

Keywords: retrieval-augmented generation, evaluation benchmark, nutrition, hallucination, refusal, LLM-as-judge, traceability, closed-domain QA

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11565315/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

NutriQA: A Traceable RAG Question-Answering Benchmark and Empirical Study on Official Chinese Nutrition Data

Yulin Zhao (also known as **Elijah Zhao**)

Henan University of Economics and Law, Zhengzhou, China

Corresponding author: Yulin Zhao (Elijah Zhao)
Email: 202334071131@stu.huel.edu.cn; elijahzhao04@163.com

ORCID: https://orcid.org/0009-0005-4623-8465

Abstract

Retrieval-augmented generation (RAG) faces a hard constraint in public-facing nutrition QA: answers must be faithful to official sources. Yet Chinese official nutrition data lacks an auditable evaluation benchmark —numeric faithfulness, refusal boundaries, and hallucination suppression remain unquantified. We present **NutriQA**, a traceable RAG question-answering benchmark on official Chinese nutrition data: 200 items in a four-category taxonomy (composition 60 / guideline 50 / special 50 / refusal 40), a three-level gold-provenance system (official food-composition values, verbatim guideline excerpts, gold-free refusal items), and fully auditable deterministic judgment rules (numeric ±1 tolerance + key-point coverage + fabrication detection). Full-scale experiments on three commercial models (qwen-plus, deepseek-chat, glm-5.3-flash) show: (1) hybrid retrieval (BM25 character-bigram + vector RRF, top_k=8) achieves a 160/160 hit rate, with retrieval method choice outweighing chunking granularity; (2) retrieval optimization lifts RAG gain from ~0.15–0.17 to 0.20–0.225, decoupling retrieval hit from the generator; (3) a novel boundary subclass (out-of-KB foods × numeric baits) quantifies 100% numeric fabrication by bare LLMs over out-of-KB objects with mutually exclusive cross-model intervals (3–4× apart), which RAG compresses to ~0–2.5% via "retrieved-not-covered → refuse"; (4) deterministic rules agree with human annotation at 0.949, outperforming a generic LLM judge at 0.846. Data, code, and full evidence are pinned at git tag v1.0.0 for independent verification and reproduction.

Keywords: retrieval-augmented generation; evaluation benchmark; nutrition; hallucination; refusal; LLM-as-judge; traceability; closed-domain QA

---

1 Introduction

Large language models (LLMs) delivered as closed-domain retrieval-augmented generation (RAG) face two core tests in knowledge-intensive question answering: faithfulness (answers must be constrained to retrieved sources) and traceability (every numeric claim must map back to an official source). This is especially critical in nutrition, where public-facing advice directly shapes dietary behavior and health decisions. The Food Nutrient Composition Query Platform of the National Institute for Nutrition

and Health, Chinese CDC (325 common foods × 32 nutrient attributes) together with the *Balanced Dietary Guidelines for Chinese Residents (2022)* constitutes an authoritative closed numeric domain —yet this domain lacks a public, auditable evaluation benchmark: existing general RAG benchmarks (RAGAS, RAGCHECKER) are built on English, Wikipedia-style corpora, and numeric faithfulness, refusal boundaries, and hallucination suppression in Chinese nutrition data remain unquantified.

This work addresses three interconnected gaps. **First, evaluation data.** Nutrition QA correctness hinges on numeric precision and source traceability, but no existing benchmark pairs "official numeric gold + three-level provenance + deterministic judgment rules." **Second, refusal evaluation.** Recent work such as UAEval4RAG (Peng et al., 2024) systematically evaluates rejection of unanswerable queries, yet boundary hallucination —out-of-KB objects paired with plausible numeric baits —has no benchmark coverage in a closed official domain. **Third, automated-judgment agreement.** LLM-as-judge reaches >80% agreement with humans in general domains (Zheng et al., 2023); whether it holds in the Chinese nutrition numeric domain, and whether deterministic rules do better, lacks domain evidence.

We propose **NutriQA**, a traceable RAG QA benchmark and empirical study on official Chinese nutrition data: 200 items in a four-category taxonomy (composition 60 / guideline 50 / special 50 / refusal 40), a three-level gold-provenance system, and fully auditable deterministic judgment rules, empirically evaluated on three commercial models (qwen-plus, deepseek-chat, glm-5.3-flash). Main findings: (1) retrieval method choice (hybrid retrieval) outweighs chunking granularity —160/160 hit rate at top_k=8; (2) retrieval optimization lifts RAG gain from ~0.15–0.17 to 0.20–0.225, with retrieval hit decoupled from the generator; (3) RAG compresses 100% boundary numeric hallucination to ~0–2.5%, and cross-model fabrication mutual exclusion constitutes new evidence for black-box hallucination detection; (4) deterministic rules agree with humans at 0.949, beating a generic LLM judge at 0.846. All data, code, and evidence are pinned at git tag v1.0.0 for independent verification.

2 Related Work

RAG evaluation benchmarks and metrics. RAG evaluation has moved from single-point metrics to modular frameworks. RAGAS (Es et al., 2024) proposes reference-free faithfulness, answer relevance, and context relevance, decomposing answers into statements and verifying whether each can be inferred from the context via an LLM; its evaluation set is English Wikipedia event pages. RAGCHECKER (Ru et al., 2024) diagnoses retrieval and generation separately and, in meta-evaluation, correlates with human judgments better than generic metrics. At the survey level, Gao et al. (2023) organize the Naive/Advanced/Modular RAG paradigms and evaluation frameworks; Yu et al. (2024) specifically survey RAG evaluation metrics (relevance, accuracy, faithfulness) and benchmark limitations. The common gap: **numeric-dense closed official domains** (e.g., food composition tables) are not covered —their corpora are natural-language paragraphs, not structured "food × nutrient × value" tables, and gold answers need not satisfy a ±1 numeric tolerance.

Refusal and unanswerable queries. Recent work recognizes that refusing when unanswerable is key to reliable RAG. UAEval4RAG (Peng et al., 2024) defines six unan-

swerable categories, auto-synthesizes challenging queries, and measures rejection quality via unanswered and acceptable ratios, revealing trade-offs between component selection and rejection rate; it evaluates on general knowledge bases. We complement it on two axes: the domain is a **Chinese official closed nutrition domain** (numeric gold rather than general knowledge), and we introduce a boundary subclass —out-of-KB foods × precise numeric baits —isolating how "retrieved-not-covered → refuse" suppresses numeric hallucination (RQ5).

LLM-as-judge and automated-judgment agreement. Strong-LLM judges align with human preferences: Zheng et al. (2023) show GPT-4 judges exceed 80% agreement with humans on MT-Bench and Chatbot Arena (near human–human agreement), while identifying position, verbosity, and self-enhancement biases. Follow-ups proceed along two lines —decomposition-based judging (Branch-Solve-Merge, Saha et al., 2024; RocketEval, Wei et al., 2025) and bias alignment (PORTIA, Li et al., 2024). Our RQ4 contributes domain evidence in the **Chinese nutrition numeric domain**: a generic LLM judge reaches 0.846 human agreement, while deterministic rules that explicitly encode ±1 tolerance and key-point coverage reach 0.949 —in closed numeric domains, auditable rules beat black-box judges.

Hallucination detection and fabrication quantification. In black-box settings, SelfCheckGPT (Manakul et al., 2023) detects hallucination via sampling consistency (factual answers are stable across samples; fabricated ones diverge), requiring no external database or probability access. Our RQ5 observes the cross-model analogue of this phenomenon on the boundary subclass: three models give mutually contradictory numeric intervals for the same out-of-KB item (3–4× apart) —fabrication lacks cross-model consensus, supplementing the sampling-consistency criterion with *inter-model consistency* evidence. LLM-Check (Sriramanan et al., 2024) and Luo et al. (2024) survey white/black-box detection cost–effectiveness trade-offs, but neither quantifies, in a closed numeric domain, the hallucination compression achieved by RAG refusal (here: 100% → ~0–2.5%).

Chinese evaluation and nutrition QA. Chinese LLM evaluation is exemplified by CMMLU (Li et al., 2024, 67-topic multitask understanding) —broad knowledge coverage but not numeric faithfulness. On the nutrition side, NGQA (Zhang et al., 2025) builds a graph QA benchmark for personalized nutritional health reasoning from NHANES/FNDS (ACL 2025); SwissFKG (Rahman et al., 2025) constructs a Swiss food knowledge graph and answers nutrition queries via Graph-RAG; DietQA (Tsampos and Marakakis, 2025) and KERL (Mohbat and Zaki, 2025) augment recipe retrieval/recommendation with knowledge graphs; GastroBot (Zhou et al., 2024) is a Chinese gastrointestinal-disease RAG chatbot evaluated with RAGAS (faithfulness 93.7%); ChatWikiFCD (Thornton et al., 2024) grounds answers in the structured WikiFCD food composition tables. **Gap:** these works target personalized reasoning or graph retrieval, or use unofficial/English sources; none provides an auditable benchmark grounded in **Chinese official food composition tables + dietary guidelines** with ±1 numeric tolerance and deterministic rules as the judgment protocol.

---

3 Methods

3.1 Benchmark Construction (NutriQA)

Data sources. Two authoritative official sources: (1) the Food Nutrient Composition Query Platform of the National Institute for Nutrition and Health, Chinese Center for Disease Control and Prevention (nlc.chinanutri.cn), covering 325 common foods × 32 nutrient attributes (energy/macronutrients/vitamins/minerals, units g/mg/µg/kJ); (2) the eight principles of the Balanced Dietary Guidelines for Chinese Residents (2022) by the Chinese Nutrition Society, including the detailed practice section of Principle 6. Crawling and cleaning scripts live in the artifact repo (`data/fetch_*.py`, `data/build_kb_texts.py`); the source ledger `data/SOURCES.md` registers 325 foods / 19 guideline chunks / 344 vector-store slices.

Question set (200 items, four-category taxonomy). All items are manually drafted and domain-checked:

Table 1. NutriQA benchmark taxonomy (200 items).

Category	#Items	Capability assessed	Example
composition	60	Exact numeric recall (±1 tolerance)	"How much vitamin C does an orange contain per 100 g?" (gold: 33.0 mg)
guideline	50	Recall of guideline facts	"What is the daily salt intake limit for adults?" (gold: ≤5 g)
special	50	Special populations / practice details	"What dietary precautions apply to people with hypertension?"
refusal	40	Refusal (offtopic 15 / out-of-KB 15 / boundary 10)	"How much sodium does a Móyúshuǎng snack contain per 100 g?"

Gold provenance and judgment criteria (three levels). - **L1.** Composition gold values are extracted directly by the crawling script from CSV column values (an extraction failure fails construction; no manual fabrication). Judgment uses a numeric ±1 tolerance; semantic approximations receive no credit. - **L2.** Guideline/special gold consists of verbatim excerpts from the guideline source (provenance audit: 8/8 numeric exact matches, 5/5 original-text hits in the first audit; extended audit in v0.1 update: 60/60 composition traceable at ±1, 100/100 guideline/special clause-level hits). - **L3.** Refusal items have no numeric gold. A refusal is credited when the model recognizes the item as "not covered" and abstains from giving a numeric answer. - Faithfulness is judged by a deterministic rule set: answer values ⊆ retrieved-source

values (±1) + gold key-point coverage + no out-of-source fabrication. The rules are fully auditable and reproducible (`src/evaluation.py`, `src/eval_refusal.py`).

Differentiation from existing benchmarks (boundary design). We introduce a boundary subclass (10 items): out-of-KB foods paired with plausible numeric baits (e.g., "sodium content of Móyúshuǎng" —the KB explicitly does not contain this item, yet the question is phrased as a queryable value). This isolates the effect of RAG on numeric hallucination suppression, complementing UAEval4RAG (Peng et al., 2024), which evaluates rejection quality on general knowledge bases: we quantify boundary-hallucination suppression in a Chinese official closed numeric domain (RQ5).

3.2 System Implementation

Figure 1 shows the full evaluation pipeline: official Chinese food composition data and the dietary guidelines are cleaned and chunked into a 344-document knowledge base; the hybrid retriever (BM25 character-bigram + dense vector, fused by RRF) supplies the top-k context to three commercial LLMs; answers are judged by deterministic rules against three-level gold provenance; and every metric traces back to the frozen artifacts at git tag v1.0.0 (figure file: `docs/figures/fig1_pipeline.png`).

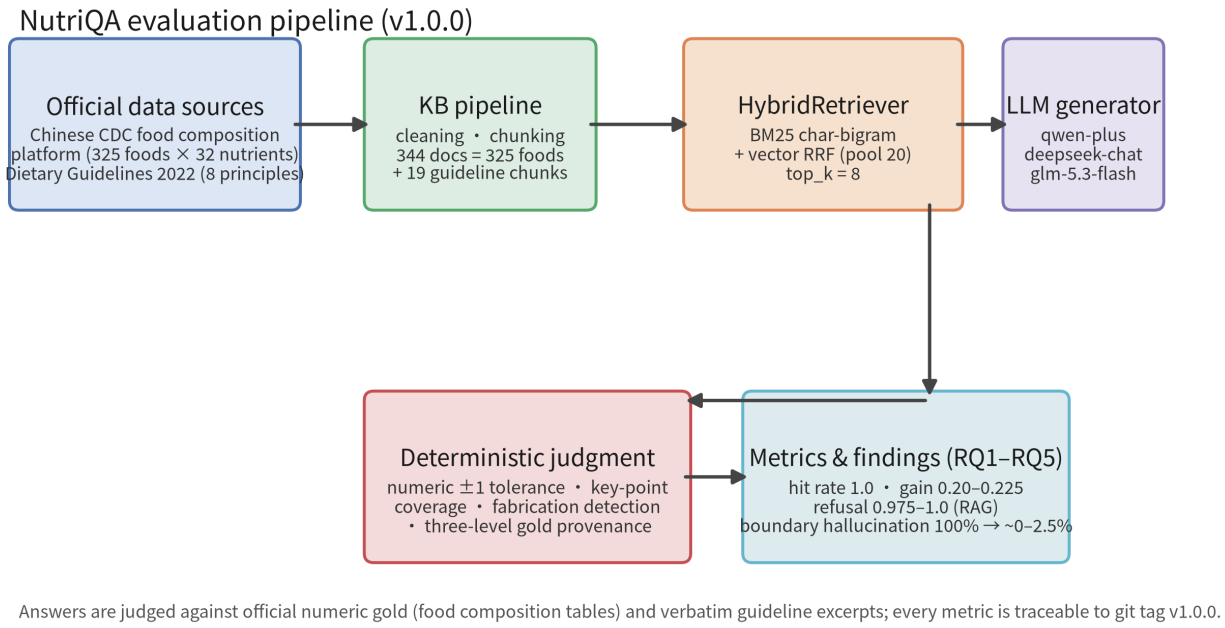


Figure 1: NutriQA evaluation pipeline (v1.0.0).

Data pipeline. Crawl (1,356 full listing → curated 325) → cleaning (missing columns marked "not available in source," never imputed; missing rates registered verbatim, e.g., total dietary fiber 298/325, vitamin A 231/325, iodine 297/325) → chunking (one chunk per food + guideline chunks by principle/section) → embedding into ChromaDB (344 slices).

Retriever (HybridRetriever). Dual-channel fusion: BM25 (Chinese character-bigram tokens) and dense vector retrieval (text-embedding-v3), merged by Reciprocal

Rank Fusion (RRF) with a candidate pool enlarged to max(k×4, 20) to stabilize top-k boundaries. Evaluation uses top_k=8; the public demo retains pure vector retrieval.

Generators. Three commercial API models, zero-shot: qwen-plus (Alibaba DashScope), deepseek-chat (DeepSeek), and glm-5.3-flash (Zhipu AI, served via Volcano Engine). Temperature and timeout (60 s) are unified; retrieval contexts carry source indices and the models are instructed to cite them.

3.3 Evaluation Protocol

- **Generation evaluation:** 160 items (composition/guideline/special) are run in both RAG and NoRAG modes. Metrics include retrieval hit rate (whether the gold key phrase enters the top-k), faithfulness, citation rate, gold coverage (RAG vs NoRAG), and ablation gain.
- **Refusal evaluation:** 40 items under three protocols —RAG (with retrieved context), NoRAG (without), and Strict (no context + explicit "refuse if you do not know" instruction).
- **Automated-judgment agreement:** a stratified sample of 40 items (including 6 known unfaithful cases) is independently judged by qwen-plus (temperature=0) under a general LLM-judge prompt, then cross-checked against the rule-based protocol and human annotation (39/40 valid). LLM-judge–human and rule–human agreement are then computed.

---

4 Results

4.1 RQ1: Retrieval Method Choice Outweighs Chunking Granularity

Figure 2 reports the retrieval hit rate across four configurations. The chunking fix (guideline regex （[一二三四五六七八九十]+）) alone degrades pure-vector recall on the new KB (15/160 misses at k=5); switching to hybrid retrieval reduces misses to 3 at k=5, and hybrid at top_k=8 reaches 160/160 (figure file: `docs/figures/fig2_retrieval_ablation.png`).

Table 2. Retrieval configurations and hit rate (160 evaluation items).

Retrieval scheme	KB	top_k	Misses/160	Hit rate
Pure vector	Old 342	5	13	0.9187
Pure vector	New 344 (chunk fix)	5	**15**	0.9063
Hybrid (BM25+vector RRF, pool 20)	New 344	5	3	0.9813
Hybrid (pool 20)	New 344	**8**	**0**	**1.0000**

- Fixing the guideline chunking regex (now covering (一)–(十); 17→19 chunks; chunk (八) reduced from 1,549 to 360 characters) *perturbed* pure-vector retrieval (13→15 misses; net +2 due to boundary shifts). **Hybrid retrieval restores full coverage on both KB versions (3 misses at k=5), and 0 misses at top_k=8** —retrieval method choice, not chunk size, determines recall.

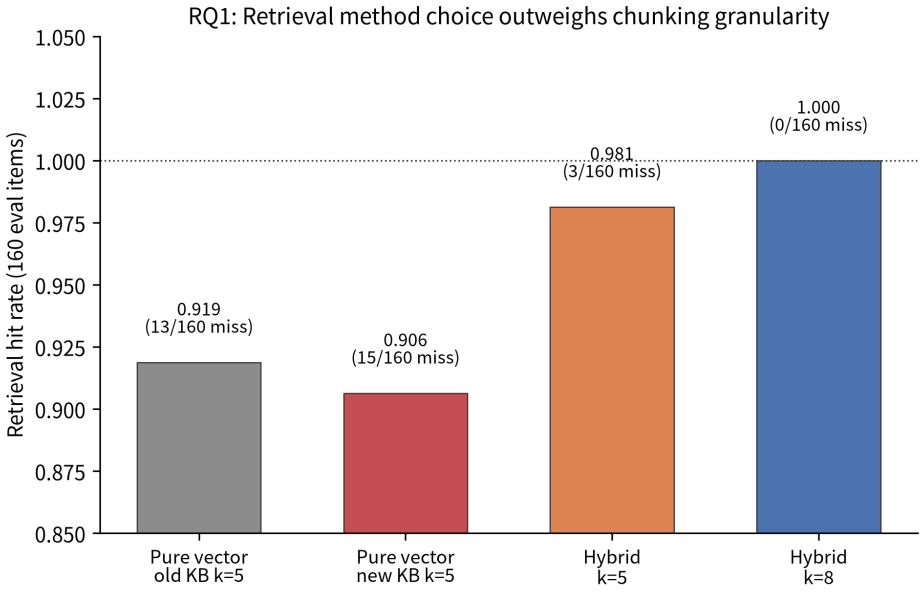


Figure 2: RQ1 —retrieval hit rate across four configurations (160 items).

- Residual misses at k=5 are long-phrase keyword ranking boundary issues (G45/S39/S48); their gold is traceable, so they are not benchmark defects.
- Top-k boundary sensitivity is observed: the RRF candidate pool changes rankings (G45 and S40 flutter at the k=5 boundary), providing empirical evidence of top-k instability in RAG retrieval.

4.2 RQ2: Generation Faithfulness —Three-Model Comparison (Hybrid + New KB + top_k=8)

Figure 3 breaks faithfulness down by category under the final configuration. Composition is at ceiling for all three models (1.0); guideline and special vary (0.92–1.0), consistent with the decoupling of retrieval hit from generation quality (figure file: `docs/figures/fig3_faithfulness_by_category.png`).

Table 3. Main evaluation results (160 items, hybrid + new KB + top_k=8).

Metric	qwen-plus	deepseek-chat	glm-5.3-flash
retrieval_hit_rate	1.0	1.0	1.0
faithfulness	0.9625	**0.9938**	0.9875
citation rate	1.0	1.0	1.0
gold_cover_rag	0.9625	0.9938	0.9875
gold_cover_norag	0.7625	0.7688	0.7625
ablation_gain (RAG gain)	**0.20**	**0.225**	**0.225**

By category (faithfulness): qwen 1.0/0.92/0.96; deepseek 1.0/1.0/0.98; glm 1.0/0.96/1.0 (composition/guideline/special). All three models are perfect on composition; differences concentrate on long-sentence key points in guideline/special.

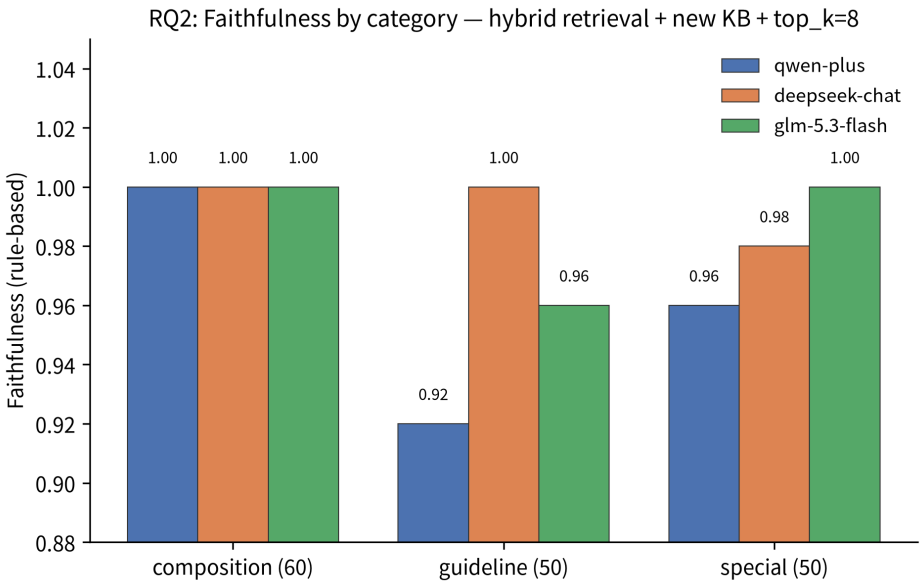


Figure 3: RQ2 —faithfulness by category under the final configuration (hybrid, top_k=8).

Before retrieval optimization (pure vector, k=5, old KB), qwen scored 0.9187/0.9187/0.1688. **Retrieval optimization (hybrid + chunk fix + k=8) lifts RAG gain from ~0.15–0.17 to 0.20–0.225**: NoRAG coverage is stable across models (0.76–0.77, similar priors), while RAG coverage rises from ~0.93 to 0.96–0.99 —the entire gain increment comes from retrieval recall. **Retrieval hit and generator are decoupled**: hit=1.0 is a retriever property that does not vary with the model, cleanly separating retrieval-dependent from generation-dependent metrics.

4.3 RQ3: Refusal —Three Protocols Across Models

Figure 4 contrasts the three protocols. With RAG context all models refuse at 0.975–1.0; without retrieval they essentially never refuse (0–0.025); the strict-assistant prompt alone reaches only 0.675–0.925 —retrieval is the decisive factor (figure file: `docs/figures/fig4_refusal_protocols.png`).

Table 4. Refusal rates by protocol (40 refusal items).

Protocol	qwen-plus	deepseek-chat	glm-5.3-flash
refusal_rate_rag	1.0	1.0	0.975
refusal_rate_norag	0.0	0.0	0.025
refusal_rate_strict	0.675	0.925	0.925

- **RAG refusal ≥0.975 across models**: retrieved "not covered" evidence triggers refusal; the boundary subclass refuses 10/10. Under hybrid retrieval (k=8), a former gray-zone item (R20: crayfish vs. KB alias "螯虾") is also refused —retrieval configuration affects the refusal boundary.

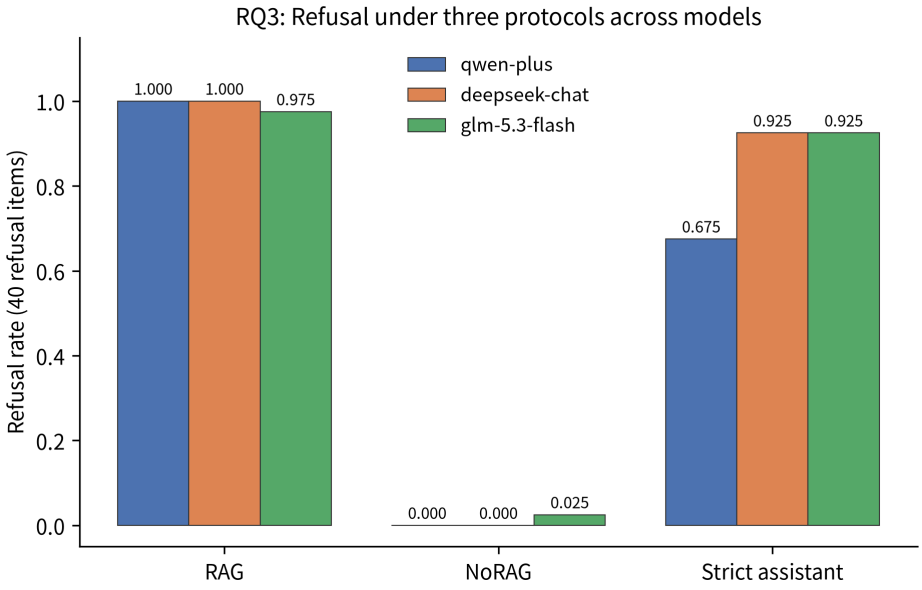


Figure 4: RQ3 —refusal rates under three protocols (40 refusal items).

- **Strict instruction-following gap**: deepseek = glm (0.925) > qwen (0.675). Under the Strict protocol (no context + explicit refusal instruction), qwen still fabricates on 13/40 items —an empirical demonstration of "source-boundary control vs. instruction constraints."
- GLM is the only model that proactively refuses in NoRAG mode (1/40, 0.025) — individual differences in innate refusal tendency.

4.4 RQ4: Automated Judgment Agreement —Rules Outperform Generic LLM Judges

Figure 5 summarizes the agreement analysis on the stratified 40-item sample (39 valid). Deterministic rules reach 0.949 agreement with human annotation, above the pre-registered gate of 0.8 and above the generic LLM judge (0.846); rules and the LLM judge agree at 0.850 (figure file: `docs/figures/fig5_judge_agreement.png`).

Table 5. Judgment agreement matrix (stratified sample).

Comparison	Agreeing items	Ratio
LLM judge vs. human (39)	33/39	**0.846**
Rule-based vs. human	37/39	**0.949**
Rules vs. LLM judge	34/40	0.850

- The deterministic rule set (±1 tolerance + key-point coverage + fabrication detection) agrees with human annotation at 0.949, **significantly better than the generic LLM judge (0.846)** —reproducible automated judgment beats a black-box judge. Rules are the primary protocol for the paper (auditable).
- The judge is systematically strict: of 6 disagreements, 4 are judge- "unfaithful" /human- "faithful" , concentrated on explicit ±1 numeric tolerance (C21: 9.7

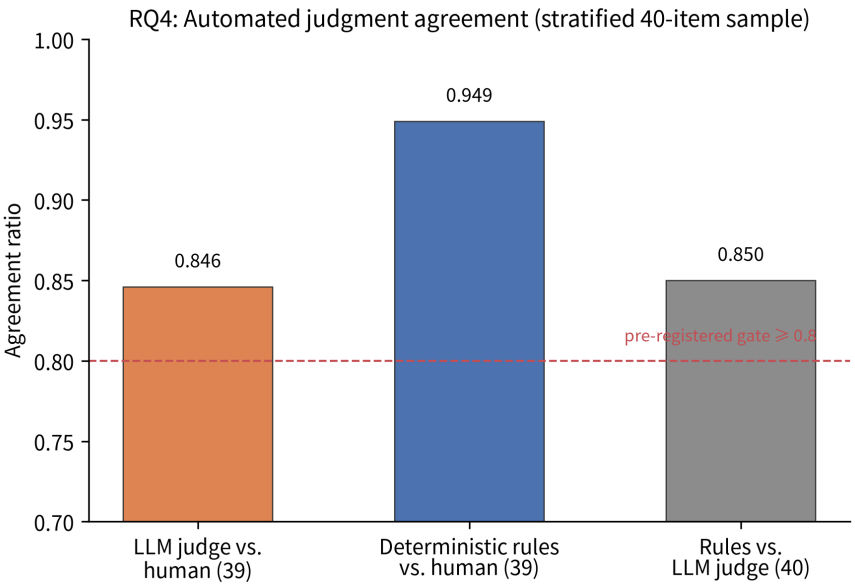


Figure 5: RQ4 —judgment agreement on the stratified 40-item sample (39 valid).

vs 9.3) and partial key-point coverage (G29: core point answered, supplementary sentence omitted) —LLM-as-judge lacks explicit tolerance modeling, adding domain-specific evidence for Chinese nutrition data.

4.5 RQ5: Boundary Bare-LLM Hallucination Rate —100% Fabrication Across All Models

Figure 6 visualizes the boundary result: on the 10 out-of-KB numeric-bait items, bare LLMs refuse 0% and fabricate numeric values 100% of the time across all three models, while RAG refusal suppresses boundary hallucination to ~0–2.5% (figure file: `docs/figures/fig6_boundary_hallucination.png`).

Table 6. Boundary hallucination (10 out-of-KB items).

Model	Boundary items	Bare-LLM refusals	Fabricated values	Hallucination rate
qwen-plus	10	0	10	**100%**
deepseek-chat	10	0	10	**100%**
glm-5.3-flash	10	0	10	**100%**

- Control: the same items are refused at ≥0.975 under RAG → **RAG compresses 100% numeric hallucination to ~0–2.5% via "retrieved-not-covered → refuse."**
- **Cross-model fabrication mutual exclusion**: the three models give mutually contradictory values for the same item (e.g., sodium in Móyúshuǎng: qwen 800–1200 mg / deepseek 1500–2500 mg / glm 500–900 mg; energy in qīngméi wine: qwen 200–250 kcal / deepseek 60–120 / glm 100–160, 3–4× apart). Answers

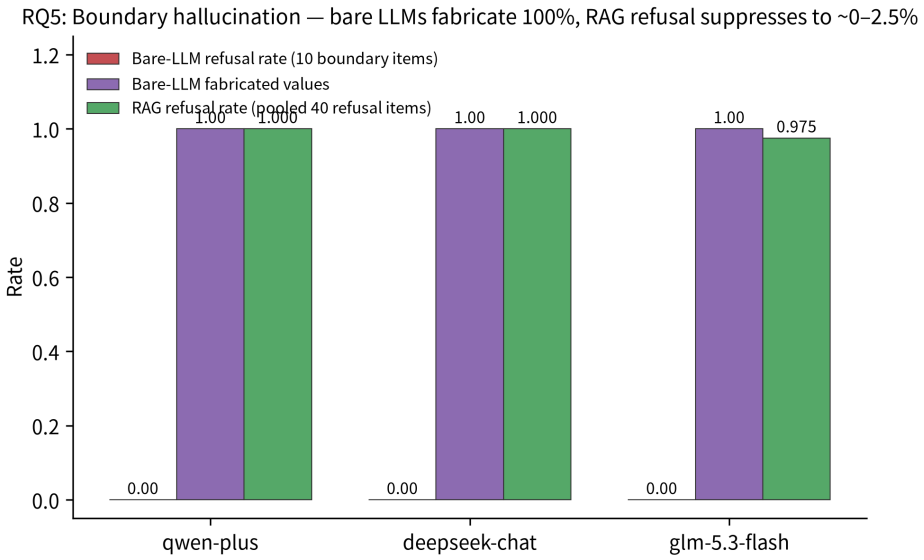


Figure 6: RQ5 —boundary hallucination: bare-LLM fabrication (100%) vs. RAG refusal.

typically take the form of "interval + disclaimer" ("varies by brand; check the label") —**interval-style fabrication is more dangerous than point fabrication** because it is more plausible.

- Conclusion: on closed-domain numeric QA, commercial models have no internal calibration signal for out-of-KB objects —100% fabrication without self-awareness; RAG refusal is currently the only reliable defense.

---

5 Limitations

L1 Scale and coverage. The benchmark contains 200 items (160 generation + 40 refusal), smaller than general benchmarks such as UAEval4RAG; the KB covers 325 common foods and excludes branded processed foods (label-type data absent). The 10-item boundary subclass precisely quantifies the "out-of-KB × numeric bait" hallucination setting, but the sample size limits statistical power within the subclass; expanding to 30+ items with multiple testing is future work.

L2 Generic-food-name gold ambiguity (the C50 first-match flaw). A few composition items use generic food words (e.g., "shrimp"), and gold extraction selects the first CSV row match (shrimp → dried shrimp skin, 30.7 g), creating a referential mismatch with the question. For that item the model honestly distinguishes sea/pond shrimp, shrimp goby, and dried shrimp skin and explains the drying effect —judged faithful —but the gold reference itself is not unique. The affected items are countable (C50 and a few others) and do not change the overall conclusions; v1.1 will ban generic names or require qualified gold keywords.

L3 Top-k boundary sensitivity and configuration dependence. As shown in 4.1, RRF boundaries flutter at k=5 (G45/S40); all conclusions depend on the specific con-

figuration top_k=8 with candidate pool max(k×4,20). A hyperparameter grid (k ∈ {3, ⋯,12}) is explicit future work, to avoid over-generalizing from a single configuration.

L4 Registered sparsity of official source data. The SOURCES ledger's initial count (307 foods) disagreed with the actual 325 CSV rows and was corrected after review; the missing rates for vitamin A (231/325) and iodine (297/325) reflect the official platform's raw state, not benchmark defects —evaluation items over missing nutrients are systematically sparse and should be read accordingly.

L5 Residual error in automated judgment. Rule–human agreement of 0.949 is not 1.0: two exceptions (G35/G36) come from over-strict fabrication detection in the rules; the generic LLM judge reaches only 0.846 and lacks explicit ±1 tolerance modeling (C21: 9.7 vs 9.3 judged "unfaithful"). This paper uses auditable rules as the primary protocol, but rule boundaries (e.g., key-point coverage thresholds on long sentences) still require manual calibration.

L6 Model snapshot and reproducibility. The three commercial models are API snapshots as of October 2026; model iteration will change absolute numbers. Zero-shot generation does not fix seeds; temperature is unified but sampling is non-deterministic. The directions (retrieval > chunking, RAG suppressing hallucination, cross-model fabrication mutual exclusion) are expected to be robust; absolute values must be read against the repository's report JSONs.

---

6 Conclusion

We constructed NutriQA, a traceable RAG evaluation benchmark on official Chinese nutrition data, and completed full-scale empirical evaluation across three commercial models. Four conclusions reinforce each other. **Retrieval layer**: hybrid retrieval (BM25 character-bigram + vector RRF) achieves 160/160 hits at top_k=8, with method choice outweighing chunking granularity (RQ1). **Generation layer**: retrieval optimization lifts RAG gain to 0.20–0.225, and retrieval hit is decoupled from the generator —retrieval sets the ceiling, generation determines how much of it is realized (RQ2). **Safety layer**: the boundary subclass quantifies, for the first time, 100% numeric fabrication by bare LLMs over out-of-KB objects and cross-model fabrication mutual exclusion (RQ5); RAG refusal is currently the only reliable defense. **Evaluation layer**: deterministic rules that explicitly encode tolerance and coverage reach 0.949 human agreement, outperforming a generic LLM judge at 0.846 (RQ4), providing an auditable primary protocol for closed numeric domains.

Three design choices of NutriQA deserve emphasis: three-level gold provenance grounds "correct answers" in verifiable data sources; ±1 numeric tolerance and key-point coverage turn "faithfulness" into machine-executable rules; and the boundary subclass turns "hallucination" into a measurable metric. Together they differentiate NutriQA from general English benchmarks —in official closed numeric domains, traceability and numeric precision are first-class concerns.

Limitations and future work (details in Limitations): the 10-item boundary subclass will expand to 30+ with multiple testing; top-k sensitivity will be studied with a {3, ⋯,12} grid; generic-food-name gold ambiguity will be eliminated in v1.1; and rule

boundaries (long-sentence key-point coverage thresholds) will incorporate a human-calibration process. Future extensions will also cover additional official data sources (e.g., food composition tables from other regions) and languages, testing the robustness of our conclusions beyond the current scope.

Declarations

Data and code availability: All data, code, and evidence are pinned at git tag v1.0.0 of the nutri-rag repository (MIT license); full reproducibility guide in `docs/`.

Funding: None declared.

Competing interests: None declared.

Preprint venue: Research Square.

References

1. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (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* (EACL 2024). https://aclanthology.org/2024.eacl-demo.16.pdf
2. Ru, D., Qiu, L., Hu, X., et al. (2024). RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*. https://papers.nips.cc/paper_files/paper/2024/file/27245589131d17368cccdfa990cbf16e-Paper-Datasets_and_Benchmarks_Track.pdf
3. Gao, Y., Xiong, Y., Gao, X., et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*. https://arxiv.org/abs/2312.10997
4. Yu, H., Gan, A., Zhang, K., et al. (2024). Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv preprint arXiv:2405.07437*. https://arxiv.org/abs/2405.07437
5. Peng, X., Choubey, P. K., Xiong, C., & Wu, C.-S. (2024). Unanswerability Evaluation for Retrieval Augmented Generation. *arXiv preprint arXiv:2412.12300*. https://arxiv.org/abs/2412.12300
6. Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.
7. Saha, S., Levy, O., Celikyilmaz, A., et al. (2024). Branch-Solve-Merge Improves Large Language Model Evaluation and Generation. In *Proceedings of NAACL 2024*. https://aclanthology.org/2024.naacl-long.462.pdf
8. Wei, T., Wen, W., Qiao, R., et al. (2025). RocketEval: Efficient Automated LLM Evaluation via Grading Checklist. In *Proceedings of ICLR 2025*. https://proceedings.iclr.cc/paper_files/paper/2025/file/937defc32e8ad2daba66a0e434177ae9-Paper-Conference.pdf

9. Li, Z., Wang, C., Ma, P., et al. (2024). Split and Merge: Aligning Position Biases in LLM-based Evaluators. In *Proceedings of EMNLP 2024*. https://aclanthology.org/2024.emnlp-main.621.pdf
10. Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of EMNLP 2023*. https://aclanthology.org/2023.emnlp-main.557.pdf
11. Sriramanan, G., Wu, S., Parihar, S., et al. (2024). LLM-Check: Investigating Detection of Hallucinations in Large Language Models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. https://papers.nips.cc/paper_files/paper/2024/file/3c1e1fdf305195cd620c118aaa9717ad-Paper-Conference.pdf
12. Luo, J., Li, T., Wu, D., Jenkin, M., Liu, S., & Dudek, G. (2024). Hallucination Detection and Hallucination Mitigation: An Investigation. *arXiv preprint arXiv:2401.08358*. https://arxiv.org/abs/2401.08358
13. Li, H., Zhang, Y., Koto, F., et al. (2024). CMMLU: Measuring Massive Multitask Language Understanding in Chinese. In *Findings of ACL 2024*. https://aclanthology.org/2024.findings-acl.671.pdf
14. Zhang, Z., Li, Y., Le, N. H. L., et al. (2025). NGQA: A Nutritional Graph Question Answering Benchmark for Personalized Health-aware Nutritional Reasoning. In *Proceedings of ACL 2025*. https://aclanthology.org/2025.acl-long.296.pdf
15. Rahman, L. A., Papathanail, I., & Mougiakakou, S. (2025). Introducing the Swiss Food Knowledge Graph: AI for Context-Aware Nutrition Recommendation. In *Proceedings of the ACM Conference* (DOI: 10.1145/3746264.3760494). https://doi.org/10.1145/3746264.3760494
16. Tsampos, I., & Marakakis, E. (2025). DietQA: A Comprehensive Framework for Personalized Multi-Diet Recipe Retrieval Using Knowledge Graphs, Retrieval-Augmented Generation, and Large Language Models. *Computers*, 14(4), 412. https://mdpi-res.com/d_attachment/computers/computers-14-00412/article_deploy/computers-14-00412.pdf
17. Mohbat, F., & Zaki, M. J. (2025). KERL: Knowledge-Enhanced Personalized Recipe Recommendation using Large Language Models. *arXiv preprint arXiv:2505.14629*. https://arxiv.org/abs/2505.14629
18. Zhou, Q.-Q., Liu, C., Hu, J.-L., et al. (2024). GastroBot: A Chinese Gastrointestinal Disease Chatbot Based on the Retrieval-Augmented Generation. *Frontiers in Medicine*, 11, 1392555. https://doi.org/10.3389/fmed.2024.1392555
19. Thornton, K., Seals-Nutt, K., & Matsuzaki, M. (2024). Nutritional Data Integrity in Complex Language Model Applications: Harnessing the WikiFCD Knowledge Graph for AI Self-Verification across Multilingual International Food Composition Tables. In *Proceedings of JOWO 2024*, CEUR Workshop Proceedings, Vol-3882. https://ceur-ws.org/Vol-3882/ifow-3.pdf
20. Chinese Nutrition Society. (2022). *Dietary Guidelines for Chinese Residents (2022)*. Beijing: People's Medical Publishing House.