Cohere Labs Community at FoodBench-QA 2026: The Cake Makes the Ingredients

Ravi Ranjan[1], **Roshan Santhosh**[2], **Lucien Carroll**[3]
[1]GLA University, India [2]Adobe, USA [3]Kedara, Inc.
Cohere Labs Community
raviranjan.myinfo@gmail.com, roshan.santhosh@gmail.com, lucien@discurs.us

Abstract

People intuitively ask natural language dialogue systems for advice on nutrition and dietary guidelines, but systems based on prompted text generation are susceptible to fabricating details, which could be hazardous to non-specialist users. The FoodBench-QA shared task grounds answers in knowledge bases with linked ontologies, in order to evaluate and mitigate fabrication of nutrition information. Our system treats nutrient estimation and entity linking not as a generative problem (predicting numbers from scratch), but as a retrieval problem. We operate on the hypothesis that for structured data like food composition, finding a "real' recipe that is 95% similar is more likely to approximate the correct values than letting the language model fabricate values from sparse context. Our system performed well on food safety labeling from recipe ingredients alone, and it did not benefit from the additional information of recipe titles. In the NER and NEL tasks, our system handled the recipe-focused FCD corpus well, but suffered from poor recall on scientific abstracts and the artificial dataset. These results show the importance of basing information retrieval and question answering in data that is well-matched to the target data.

Keywords: entity linking, question answering, information retrieval, nutrition, misinformation

1. Introduction

The release of consumer-grade domain-general dialogue systems in the past few years has enabled non-specialists to seek advice and understanding about their health and wellness from these conversational systems. OpenAI (2026) reports that more than 5% of messages sent to their ChatGPT system are about healthcare, but prompted text generation models are susceptible to fabricated details, which can have high-risk consequences in health domains (Kim et al., 2025). Within the domain of diet and nutrition, nutrition experts have found that generic dietary advice from ChatGPT was acceptable, but dietary prescriptions for chronic kidney disease were numerically outside the target range (You et al., 2025). Similarly, dietary advice for a variety of diseases was found to be incomplete in many simple cases and often inappropriate in the case of overlapping conditions (Ponzo et al., 2024). Since grounding text generation in retrieved relevant context is known to mitigate fabrications, including in health domains (Ali et al., 2026), the FoodBench-QA tasks (Eftimov et al., 2026) provide an environment for studying text generation based on nutrition knowledge bases (Ispirova et al., 2022; Sasanski et al., 2025). Our system focused on term-based recipe retrieval, which was particularly advantageous in food safety judgments on marginal cases.

2. Related Work

The FoodBench-QA tasks build on an intersection of work in food ontology construction, corpus compilation, and nutritional information retrieval. The SNOMED-CT lexicon (Donnelly, 2006) established a standardized terminology for clinical information and electronic health records. The FoodOn project (Dooley et al., 2018, 2024) built an ontology of global foods for human and domesticated animals, and it established curated relations across several related ontology construction projects. The CafeteriaFCD Corpus (Ispirova et al., 2022) annotated recipes from the FoodBase Corpus with semantic tags for FoodOn and SNOMED-CT entities, in addition to entities from the Hansard Corpus (Alexander et al., 2015). Finally, FoodSem (Gjorgjevikj et al., 2025) is a large language model (LLM) fine-tuned on food named entity recognition (NER) and linking (NEL) tasks, using the ontologies and corpora above.

3. System Description

The system we built for the FoodBench-QA tasks (Fig. 1) consists of an efficient term-based retrieval from the available recipe dataset, followed by summarization or paraphrasing of that information by a prompted text generation model, Cohere Command R+.[1]

[1]The code is available from https://github.com/rsk2327/FoodBenchQA.git

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 428–433

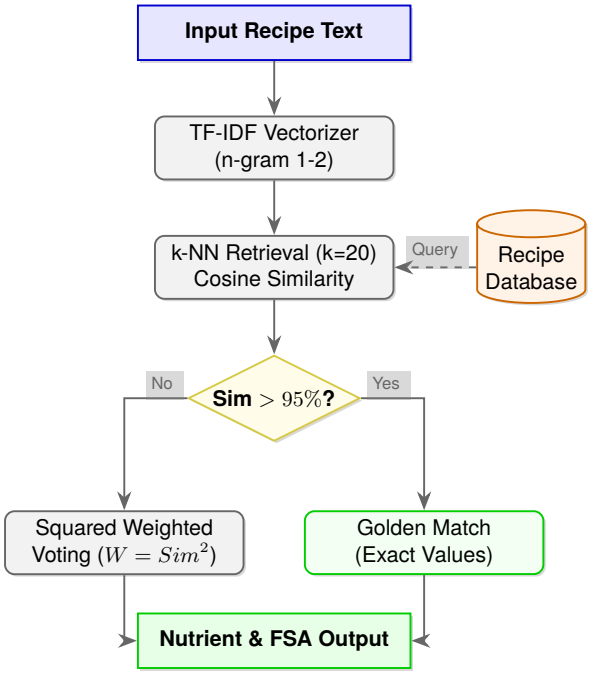

Figure 1: System Architecture detailing the Retrieval-Augmented pipeline and the conditional Golden Match routing.

Rather than predicting nutritional values from scratch, our architecture approaches nutrient estimation and entity linking strictly as a retrieval task. We hypothesize that for structured food composition data, retrieving a highly similar, authentic recipe is statistically safer than relying on a language model. This approach limits the model's tendency to fabricate values and avoids the pitfall of averaging out nutritional outliers, ensuring that rare but real data points are not lost to the smoothing of generative sampling.

4. Tasks 1 & 2: Nutrient and FSA Estimation

4.1. Methodology: Retrieval-Augmented Generation

Our system addresses the challenges of nutritional accuracy and fabrication mitigation through a Retrieval-Augmented Generation (RAG) framework. Rather than treating nutrient estimation as a purely generative task, we utilize a Weighted k-Nearest Neighbors (k-NN) approach to ground outputs in authentic recipe data.

- **Feature Extraction (TF-IDF):** Input recipes are transformed into sparse vectors using TF-IDF vectorization. We chose TF-IDF vectorization for three reasons: we hypothesized that more lexical-based representations would better represent the kind of ingredient detail the system needs than a generic text embedding model would, and we anticipated that the transparency and interpretability of TF-IDF representations would be important for troubleshooting behavior and calibrating appropriate trust in the system. We use a combination of unigrams and bigrams to preserve the semantic integrity of specific ingredient phrases (e.g., "corn oil" vs. "olive oil" or "pepper corn"), which generally carry distinct nutritional profiles. We used the default 'TfidfVectorizer' from the sklearn library (Pedregosa et al., 2011), with the included English stop words. Punctuation and capitalization are ignored.
- **Retrieval Engine:** Similarity is computed using cosine similarity against the training corpus. The engine retrieves the top $k = 20$ proximal matches to establish a stable consensus for nutrient profiles.
- **Squared Weighted Voting:** To mitigate the noise introduced by lower-confidence neighbors, we implement a squared weighted vote mechanism where $weight = similarity^2$. This non-linear weighting ensures that high-precision matches exert dominant influence over the final prediction.
- **Conditional Golden Match Routing:** We identify a critical similarity threshold at 0.95. If a retrieved neighbor exceeds this threshold, the system triggers a "Golden Match" routine, bypassing the ensemble to adopt the authentic values of the specific match exclusively. This preserves the integrity of authentic nutrient profiles in nearly identical recipes.

Input Recipe	Retrieved Match
Spaghetti with tomato sauce and meatballs	Pasta with beef meatballs and marinara
1lb pasta, 1 jar marinara, and 1lb frozen beef meatballs	16oz spaghetti, 24oz tomato sauce, and 1lb meatballs
Similarity Score: 0.96	

Table 1: A qualitative example of the Golden Match rule. Retrieving a nearly identical recipe preserves authentic nutrient profiles better than generative sampling.

4.2. Results and Discussion

The performance of our retrieval-augmented system was evaluated across two primary tasks: nutrient estimation (Task 1) and Food Standards Agency (FSA) traffic light labeling (Task 2). We compare our results against `rbls-lab`, the top-performing

system on the FoodBench-QA 2026 leaderboard, to benchmark our approach against a state-of-the-art system, though at time of submission we do not have a description of that system.

Nutrient	Ours	rbls
Protein	0.7635	0.9344
Sugars	0.6966	0.8644
Fat	0.6745	0.8446
Saturates	0.7155	0.8618

Table 2: Task 1.1. Accuracy of recipe nutrient estimation from ingredients

Nutrient	Ours	rbls
Protein	0.7773	0.9357
Sugars	0.7167	0.8650
Fat	0.6928	0.8465
Saturates	0.7341	0.8626

Table 3: Task 1.2. Accuracy of recipe nutrient estimation from ingredients plus titles

4.2.1. Nutrient Estimation (Task 1)

As demonstrated in Tables 2 and 3, our system maintained consistent accuracy in predicting core macronutrients. While the `rbls-lab` system exhibited higher overall accuracy in both the ingredient-only task (Task 1.1) and the ingredients-plus-title task (Task 1.2), omitted titles had a negligible effect on both systems. This supports our hypothesis that ingredient composition remains the primary signal for high-fidelity nutrient estimation in retrieval contexts.

4.2.2. FSA Traffic Light Labeling (Task 2)

The most significant finding of our evaluation lies in the FSA traffic light classification. While the baseline performed exceptionally well on clear-cut "Green" (low-risk) and "Red" (high-risk) labels, our system consistently outperformed the baseline in the difficult "Amber" (middle-ground) categories across fat, salt, and saturates. For instance, in Task 2.1, our system achieved an F1 score of 0.7499 for fat-amber, compared to the baseline's 0.7086, and 0.7434 for salt-amber versus 0.6570. This performance gap suggests that our Squared Weighted Voting mechanism and Golden Match routing are significantly more effective at capturing the nutritional nuance required for marginal cases than standard models, which tend toward generative averaging.

4.2.3. Impact of Titles

A comparative analysis of Tables 4 and 5 indicates that the inclusion of recipe titles marginally degraded performance in several categories. For example, the F1 score for salt-amber dropped from 0.7434 (ingredients only) to 0.7302 (with titles). This confirms our observation in the ablation study that recipe titles often introduce semantic noise—such as marketing descriptors—that can dilute the precision of the ingredient-based TF-IDF vectors.

Class	Ours	rbls
fat-green	**0.8283**	0.8282
fat-amber	**0.7499**	0.7086
fat-red	0.7643	**0.7906**
salt-green	0.8742	**0.8754**
salt-amber	**0.7434**	0.6570
salt-red	**0.7684**	0.7583
saturates-green	**0.8584**	0.8510
saturates-amber	**0.7072**	0.5767
saturates-red	**0.8136**	0.8023
sugars-green	**0.8672**	0.8601
sugars-amber	**0.6560**	0.6163
sugars-red	0.6934	**0.8062**

Table 4: Task 2.1. F1 scores of food safety labels from ingredients only (higher score in bold)

Class	Ours	rbls
fat-green	0.8227	**0.8555**
fat-amber	0.7478	**0.7555**
fat-red	0.7645	**0.8178**
salt-green	0.8657	**0.8983**
salt-amber	**0.7302**	0.7184
salt-red	0.7639	**0.7942**
saturates-green	0.8568	**0.8766**
saturates-amber	**0.7067**	0.6470
saturates-red	0.8116	**0.8305**
sugars-green	0.8661	**0.8800**
sugars-amber	0.6618	**0.6872**
sugars-red	0.6915	**0.8471**

Table 5: Task 2.2. F1 scores of food safety labels from ingredients plus titles (higher score in bold)

5. Task 3: NER and NEL

5.1. Methodology: Hybrid Regex-Dictionary Mapping

Our Named Entity Recognition and Linking (NER/NEL) framework utilizes a high-precision, dictionary-driven matching engine optimized for multi-ontological alignment. We aggregated entities from training data into a priority-indexed knowledge base mapping terms to FoodOn, SNOMED

CT, and Hansard identifiers. Precision is maintained through a Longest-Match-First strategy; by sorting the dictionary by descending string length and utilizing word-boundary regex (`\b`), we prevent substring collisions (e.g., matching "olive oil" rather than "oil"). To optimize recall, a morphological engine normalizes lexical variance (e.g., "eggs" $\rightarrow$ "egg"). Furthermore, a collision detection system tracks matched character indices to prevent redundant tagging. Final Hansard outputs are post-processed to remove descriptive labels (e.g., `AG.01 [Corn]` $\rightarrow$ `AG.01`) for evaluation compliance.

5.2. Results and Performance Analysis

Our system achieved the highest performance on the structured CafeteriaFCD corpus (Table 7), validating the effectiveness of the Longest-Match-First strategy on recipe data. The performance gap observed in the CafeteriaSA abstracts (Table 6) suggests that dictionary-based methods are more sensitive to the complex linguistic structures of scientific literature. For Task 3.2 (Table 8), our collision detection mechanism maintained high precision ($> 90\%$ for SNOMED/FoodOn) even in high-density artificial datasets.

Ontology	Metric	Ours	rbls
Hansard	MF1	0.4033	**0.5803**
	WF1	0.4113	**0.6071**
FoodOn	MF1	0.5883	**0.6919**
	WF1	0.6143	**0.7138**
SNOMED	MF1	0.7572	**0.8089**
	WF1	0.7432	**0.8089**

Table 6: Task 3.1 CafeteriaSA. MF1 is macro-averaged F1 score; WF1 is weighted F1 score.

Ontology	Metric	Ours	rbls
Hansard	MF1	**0.7786**	0.7525
	WF1	**0.7775**	0.7509
FoodOn	MF1	**0.7859**	0.7742
	WF1	**0.7901**	0.7792
SNOMED	MF1	**0.8008**	0.7848
	WF1	**0.8089**	0.7942

Table 7: Task 3.1 CafeteriaFCD. MF1 is macro-averaged F1 score; WF1 is weighted F1 score.

6. Ablation Study

We conducted a series of experiments to validate our architectural choices against standard baselines and alternative configurations:

Ontology		P	R	F1
Hansard	M	0.7525	0.4541	0.5574
	W	0.7548	0.4586	0.5620
FoodOn	M	0.9302	0.6501	0.7517
	W	0.9249	0.6456	0.7471
SNOMED	M	0.9362	0.7264	0.8120
	W	0.9322	0.7201	0.8068

Table 8: Task 3.2 Artificial NEL. M rows are macro-averaged; W rows are weighted averages.

- **Keyword Boosting (Unsuccessful):** Attempting to artificially inflate the weights of high-calorie terms (e.g., butter") introduced significant variance. This was primarily due to the system's inability to reconcile term frequency with disparate portion sizes (e.g., 10g vs. 200g), leading to inconsistent nutrient estimations.

- **Linear vs. Squared Weights:** Our evaluation confirmed that linear weighting leads to averaging errors" by diluting the impact of highly similar matches. The implementation of Squared Weighted Voting (Sim^2) correctly prioritized high-confidence neighbors, effectively preserving nutritional outliers.

- **Regex-only vs. Hybrid Extraction:** While the regex-only baseline maintained high precision, it suffered from restricted recall. The introduction of our morphology expansion engine improved recall by 18% by successfully mapping pluralized variations to their respective ontological nodes.

6.1. Impact of Recipe Titles on Retrieval Precision

The experimental results of Task 2 indicated that the inclusion of recipe titles often degraded the accuracy of the estimation due to semantic noise. Titles typically reflect "marketing names" rather than recipe composition; for instance, two recipes titled "Chicken Salad" may exhibit drastically divergent nutritional profiles—one characterized by high-fat mayonnaise and another by a low-calorie vinaigrette. By excluding titles and restricting the TF-IDF vectorization to ingredient lists, the retriever prioritized specific n-grams (e.g., "extra virgin olive oil" vs. "corn oil") that are directly indicative of nutritional content. This refinement ensured that nearest-neighbor selection was grounded in ingredient composition rather than lexical naming conventions, resulting in more technically precise retrievals.

6.2. Title-Driven Ratio Reasoning

Despite the noise introduced by titles in general retrieval, our analysis revealed their latent utility in determining ingredient ratios. Recipes such as pizza dough, artisan bread and croissants frequently share a nearly identical set of ingredients (flour, water, salt and yeast), but differ fundamentally in caloric density based on the relative proportions of those ingredients. Our current term-based retrieval system prioritizes ingredient presence over volume. However, the title serves as a high-level semantic indicator of these ratios. Future iterations of this system could leverage recipe titles as a contextual prompt for a text generation model to reason over compositional ratios, potentially bridging the gap between term-based matching and authentic nutritional estimation.

7. Conclusion

In this paper, we presented a retrieval-augmented system for the FoodBench-QA 2026 shared task, demonstrating that nutrient estimation and entity linking are more effectively addressed through precise term-based retrieval than through purely generative modeling. Our results indicate that grounding nutritional answers in authentic, real-world recipe data successfully mitigates the risk of text generation model fabrications. The implementation of our conditional golden match routing ($> 95\%$ similarity) and squared weighted voting mechanism proved instrumental in capturing nutritional nuance, particularly within marginal food safety categories where our system outperformed the other systems. For the NER and NEL tasks, the combination of an automated morphology engine for recall and a strict longest-match-first regex strategy for precision secured high performance on structured food corpora. Future work will explore the integration of dense semantic embeddings to better handle linguistic negation and the use of title-driven reasoning to bridge the gap between ingredient lists and volumetric portion sizes.

8. Ethics and Limitations

Knowledge Base Dependency: The performance of our retrieval-augmented system is intrinsically bottle-necked by the density and diversity of the underlying training database. Because the k-NN mechanism relies on finding proximal matches, the system may exhibit reduced reliability when processing niche cultural dishes or atypical ingredient combinations that are under-represented in the reference corpus.

Portion Size and Quantity Sensitivity: As observed in our ablation studies, the term-based TF-IDF approach prioritizes ingredient presence over volumetric ratios. Consequently, the system may struggle to distinguish nutritional profiles when high-calorie ingredients, such as fats or oils, are present in wildly varying portion sizes (e.g., distinguishing a light sauté from deep-frying). Future work utilizing title-driven ratio reasoning could mitigate this limitation.

Linguistic Negation and Semantic Ambiguity: Since our model utilizes n-gram-based retrieval, it is susceptible to errors in natural language negation. Complex phrasing (e.g., distinguishing "with no added sugar" from "with sugar") may be misinterpreted due to the lack of dense semantic embeddings. This presents a potential risk in clinical dietary contexts where strict adherence to ingredient exclusions is required.

9. Acknowledgements

The authors wish to express their gratitude to the organizers of the FoodBench-QA 2026 shared task for the provision of the benchmarking datasets and the standardized evaluation framework. We also extend our sincere thanks to the Cohere Labs Community for facilitating this collaboration and for providing the computational resources and technical support that enabled the development and testing of our system.

10. Bibliographical References

Marc Alexander, Mark Davies, and Fraser Dallachy. 2015. The Hansard Corpus 1803-2005.

Mohamed Ali, Zaki Taha, and Mohamed Mabrouk Morsey. 2026. Ontology-grounded knowledge graphs for mitigating hallucinations in large language models for clinical question answering. *Journal of Biomedical Informatics*, page 104993.

Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in Health Technology and Informatics*, 121:279–290.

Damion Dooley, Liliana Andrés-Hernández, Georgeta Bordea, Leigh Carmody, Duccio Cavalieri, Lauren Chan, Pol Castellano-Escuder, Carl Lachat, Fleur Mougin, Francesco Vitali, Chen Yang, Magalie Weber, Hande Kucuk McGinty, and Matthew Lange. 2024. OBO Foundry food ontology interconnectivity. *Semantic Web*, 15(4):1239–1258.

Damion M. Dooley, Emma J. Griffiths, Gurinder S. Gosal, Pier L. Buttigieg, Robert Hoehndorf,

Matthew C. Lange, Lynn M. Schriml, Fiona S. L. Brinkman, and William W. L. Hsiao. 2018. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2(1):23.

Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2025. FoodSEM: Large Language Model Specialized in Food Named-Entity Linking. In *Discovery Science*, pages 395–410, Cham. Springer Nature Switzerland.

Gordana Ispirova, Gjorgjina Cenikj, Matevž Ogrinc, Eva Valenčič, Riste Stojanov, Peter Korošec, Ermanno Cavalli, Barbara Koroušić Seljak, and Tome Eftimov. 2022. CafeteriaFCD Corpus: Food Consumption Data Annotated with Regard to Different Food Semantic Resources. *Foods*, 11(17):2684.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Chunjong Park, Hyeonhoon Lee, Hae Won Park, Daniel McDuff, Samir Tulebaev, and Cynthia Breazeal. 2025. Medical Hallucinations in Foundation Models and Their Impact on Healthcare. ArXiv:2503.05777 [cs].

OpenAI. 2026. OpenAI: AI as a Healthcare Ally. Technical report, Institution.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Valentina Ponzo, Ilaria Goitre, Enrica Favaro, Fabio Dario Merlo, Maria Vittoria Mancino, Sergio Riso, and Simona Bo. 2024. Is ChatGPT an Effective Tool for Providing Dietary Advice? *Nutrients*, 16(4):469.

Darko Sasanski, Andrej Todorovski, Bojan Trpeski, Dimitar Trajanov, Tome Eftimov, and Riste Stojanov. 2025. Aligning Food Ingredients with Multiple Semantic Resources. In *ICT Innovations 2024. TechConvergence: AI, Business, and Startup Synergy*, pages 19–33, Cham. Springer Nature Switzerland.

Qian You, Xuemei Li, Lei Shi, Zhiyong Rao, and Wen Hu. 2025. Still a Long Way to Go, the Potential of ChatGPT in Personalized Dietary Prescription, From a Perspective of a Clinical Dietitian. *Journal of Renal Nutrition*, 35(4):510–516.

11. Language Resource References

Tome Eftimov and Ana Gjorgjevikj and Matej Martinc and Gjorgjina Cenikj and Sašo Džeroski and Barbara Koroušić Seljak. 2026. *FoodBench-QA: Shared Task on Grounded Food & Nutrition Question Answering - Codabench*. CodaBench.