Current Research in Food Science 12 (2026) 101351

Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Large language models in food and nutrition science: Opportunities, challenges, and the case of FoodyLLM

Ana Gjorgjevikj [a], Matej Martinc [b], Gjorgjina Cenikj [a,c], Riste Stojanov [d], Jan Drole [a,c], Gordana Ispirova [e], Giulia Menichetti [e,f,g], Nives Ogrinc [h,k], Dimitar Trajanov [d,i], Sašo Džeroski [b], Barbara Koroušić Seljak [a,j], Tome Eftimov [a,c,*]

[a] *Computer Systems Department, Jožef Stefan Institute, Jamova Cesta 39, Ljubljana, 1000, Slovenia*
[b] *Department of Knowledge Technologies, Jožef Stefan Institute, Jamova Cesta 39, Ljubljana, 1000, Slovenia*
[c] *Jožef Stefan International Postgraduate School, Jamova Cesta 39, Ljubljana, 1000, Slovenia*
[d] *Faculty of Computer Science and Engineering, Ss. Cyril and Methodius - University, Skopje, Rudzer Boshkovikj 16, Skopje, 1000, North Macedonia*
[e] *Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA*
[f] *Network Science Institute and Department of Physics, Northeastern University, Boston, MA, USA*
[g] *Harvard Data Science Initiative, Harvard University, Boston, MA, USA*
[h] *Department of Environmental Sciences, Jožef Stefan Institute, Jamova Cesta 39, Ljubljana, 1000, Slovenia*
[i] *Department of Computer Science, Metropolitan College, Boston University, Boston, MA, 02215, USA*
[j] *Medical Faculty, University of Ljubljana, Vrazov Trg 2, Ljubljana, 1000, Slovenia*
[k] *Research and Innovation Center Pomurje, Jožef Stefan Institute, Lendavska Ulica 5a, Murska Sobota, 9000, Slovenia*

ARTICLE INFO

Handling Editor: Professor Georgios Leontidis

Keywords:
FoodyLLM
FAIR data
Nutrient estimation
Healthy diet
Large language model

ABSTRACT

Background: Reliable nutrient profiling and semantic interoperability are essential for scalable dietary assessment, food labeling (e.g., traffic-light schemes), and FAIR integration of food composition and consumption data. However, general-purpose large language models (LLMs) are not systematically exposed to structured recipe–nutrition mappings and food ontologies, limiting their accuracy and trustworthiness in food and nutrition tasks.

Scope and approach: We review recent LLM advances in life sciences and healthcare and analyze the gap in food and nutrition applications. To address this gap, we introduce FoodyLLM, a domain-specialized LLM fine-tuned on 225k task-aligned QA pairs for (i) recipe nutrient estimation, (ii) traffic-light classification, and (iii) ontology-based entity linking to support FAIR food data interoperability. We benchmark FoodyLLM against strong general-purpose baselines (e.g., Llama 3 8B, Gemini 2.0) under zero-/few-shot prompting across five evaluation folds.

Key findings: Across all tasks, FoodyLLM substantially outperforms general-purpose LLMs for nutrient estimation across all macronutrients (fat, protein, salt, saturates, sugar), accuracy increases from 0.43 to 0.63 to 0.91–0.97; for traffic-light classification across all nutrients and color categories, macro F1 improves from 0.46 to 0.80 to 0.86–0.97; and for ontology-based food entity linking across FoodOn, SNOMED-CT, and Hansard, macro F1 increases from 0.33 to 0.44 (best general-purpose baseline) to 0.93–0.98 on artificial NEL data, and from 0.24 to 0.51 to 0.67–0.84 on real corpora (CafeteriaSA and CafeteriaFCD). Overall, our results demonstrate the practical value of domain-specialized LLMs in food and nutrition research. They enable automated dietary assessment, large-scale nutritional monitoring, and FAIR data integration, while opening new pathways toward sustainable and personalized nutrition.

1. Introduction

Recent advances in Artificial Intelligence (AI), particularly Large Language Models (LLMs) (Shanahan, 2024; Luo et al., 2025a; Luo et al., 2025b), provide powerful tools for tackling complex challenges in food systems. LLMs, trained on diverse datasets, can synthesize scientific literature, model relationships, and generate context-aware insights—capabilities increasingly valuable in food and nutrition

---

* Corresponding author.
E-mail address: tome.eftimov@ijs.si (T. Eftimov).

https://doi.org/10.1016/j.crfs.2026.101351
Received 13 November 2025; Received in revised form 3 February 2026; Accepted 13 February 2026
Available online 16 February 2026
2665-9271/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

sciences.

In the pursuit of the United Nations Sustainable Development Goals (SDGs) (Carlsen and Bruggemann, 2022), especially Zero Hunger (SDG2) (Blesh et al., 2019) and Responsible Consumption and Production (SDG12) (Arora and Mishra, 2023), integrative and adaptive knowledge systems are crucial. LLMs can accelerate innovation across the food value chain, from reducing waste (Schanes et al., 2018) to enabling personalized nutrition (Agostoni et al., 2021) and supporting policy decisions (Guilpart et al., 2022).

However, general-purpose LLMs (e.g., Gemini, ChatGPT, Mistral, Llama3) face limitations in specialized domains, ending up with hallucinosis. This has spurred efforts to fine-tune models or develop domain-specific ones. With the rapid emergence of both general-purpose and fine-tuned LLMs, robust benchmarking has become critical. Platforms like Hugging Face (Jain, 2022) offer leaderboards (e.g., Open LLM Leaderboard[1]), but challenges in fairness and representativeness remain, motivating the creation of domain-specific benchmarks (e.g., in health, finance, food).

In this study, we review LLMs developed for life sciences, healthcare, and food science, identifying key gaps and opportunities. We evaluate the performance of general-purpose LLMs on food-related tasks and introduce a specialized food LLM, benchmarked on nutrient estimation, food traffic-light classification, and data interoperability. Results demonstrate clear performance gains of the domain-specific model over general-purpose ones.

2. LLMs in life sciences, healthcare, and food science

Life Science. LLMs are reshaping the life sciences by enabling new ways to analyze, model, and interact with complex biological data (Zhang et al., 2025). Recent advances include BioMedGPT, which bridges natural language with molecular and protein data to support biomedical question answering at human-level performance (Luo et al., 2024); InstructCell, a multimodal AI copilot that interprets natural language instructions for tasks such as cell type annotation and drug sensitivity prediction in single-cell RNA sequencing (Fang et al., 2025); and LLMs4Life, which enhances ontology learning through prompt engineering and reuse, producing logically consistent and complete domain ontologies (Fathallah et al., 2024). Beyond LLMs, AI breakthroughs like AlphaFold have revolutionized protein structure prediction, unlocking new possibilities in drug discovery and disease research (Jumper et al., 2021), while GenBio AI's AIDO offers a digital simulation of biological processes—from DNA to cellular functions—paving the way for personalized medicine (Song et al., 2024).

Healthcare. LLMs are also driving major advances in healthcare by supporting medical research, diagnostics, and patient care (Thirunavukarasu et al., 2023). Med-PaLM set a milestone as the first AI to pass the U.S. Medical Licensing Examination, with its successor Med-PaLM 2 achieving even higher accuracy on medical benchmarks (Tu et al., 2024). MEDITRON-70B, an open-source model trained on curated medical literature, outperforms comparable baselines by a notable margin (Chen et al., 2023). Other specialized systems include LLMD, which analyzes longitudinal medical records to structure and interpret complex patient histories (Porter et al., 2024), and Health-LLM, which leverages wearable sensor data for predictive health tasks ranging from activity to mental health, matching or surpassing larger models in performance (Kim et al., 2024). Together, these models demonstrate the growing role of LLMs in making healthcare more precise, predictive, and accessible.

Food and Nutrition Sciences. Recent materials also reveal efforts in the food sector. FoodSky (Zhou et al., 2025) is primarily designed for culinary and gastronomic applications, such as recipe recommendation, cooking assistance, and creative food generation within Chinese cuisine. Its objective is to enhance culinary creativity and provide dietary guidance through recipe-level interactions, and it is trained to perform well in chef and dietetics-style examinations. Beyond culinary systems such as FoodSky, the food domain encompasses a broad range of scientific, regulatory, and nutritional applications that require specialized LLMs tailored to analytical tasks and semantic interoperability rather than recipe generation or gastronomic assistance. These application areas involve fundamentally different objectives and data representations, making general culinary LLMs insufficient and motivating the development of task-specific (i.e., application), domain-specialized food language models. Food and nutrition data present additional domain-specific challenges, including non-standardized portion sizes, vernacular and culturally dependent food names, and mixed dishes with variable ingredient compositions. Food composition (Greenfield and Southgate, 2003) and food consumption data (Leclercq et al., 2019) are inherently heterogeneous, originating from diverse and often isolated studies conducted across different populations, geographies, and contexts. This heterogeneity arises from variations in the data collection methodologies, the representation of food items and nutrients, and the granularity of reported information. Data sources range from dietary surveys, laboratory analyses, and food frequency questionnaires to emerging sources like apps and sensors, each with unique structures and terminologies. These differences make it challenging to combine and standardize data for comprehensive analysis, limiting the scope of insights that can be drawn across studies. Adding to the complexity, many of the nutrient assessment methods currently in use are built on traditional statistical approaches that rely heavily on complete and structured datasets. Missing data (Gjorshoska et al., 2022), inconsistencies, or the lack of standardized formats often lead to biased results or the exclusion of valuable datasets. These traditional methods are not well-equipped to handle the fragmented nature of modern nutrition data, which often contains gaps, textual data, and inconsistencies. In this context, foundation models present a transformative opportunity (Ma et al., 2024). These advanced AI models, particularly LLMs, excel at synthesizing diverse and incomplete data. These capabilities hold immense potential to revolutionize nutrient assessment and food data analysis by making it possible to extract actionable insights from previously incompatible or incomplete datasets. They also open doors to dynamic, real-time assessment methods, enabling a more comprehensive and accurate understanding of nutrition that can inform public health policies, dietary recommendations, and personalized nutrition strategies. In the following section, we introduce FoodyLLM, a domain-specific LLM designed for food and nutrition sciences, to illustrate both the necessity and the advantages of tailoring LLMs to this domain.

3. FoodyLLM - a specialized model for food and nutrition sciences

To compare general-purpose LLMs with a food-specialized model, we introduce FoodyLLM, an LLM designed to i) transform food-related text into interoperable food concepts in support of FAIR data principles by linking food descriptions to established food and health semantic ontologies, and (ii) perform multi-task nutritional inference, including macronutrient estimation and health classification (e.g., traffic-light labeling), based on ingredient lists and quantities (optionally including recipe titles). We construct a dataset of 225k question-answer (QA) pairs covering tasks such as standardizing food terminology, nutrient estimation, and traffic-light classification (see Appendix A, Table 5). Using multi-task fine-tuning of a public LLM, FoodyLLM specializes in nutrient estimation, traffic-light labeling, and ontology linking, with potential for broader applications. Details about its development are available in Appendix A.

We benchmark FoodyLLM against open-source and proprietary models in zero-, one-, and few-shot settings. Across all tasks, it

---

[1] https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/

outperforms state-of-the-art general-purpose models (e.g., Llama 3 8B, Gemini 2.0), even with in-context learning. We selected Gemini 2.0 Flash as representative general-purpose baselines because, at the time of experimentation, Gemini 2.0 Flash achieved performance comparable to or exceeding GPT-4 and Claude on standard benchmarks, while stable API access and the substantially higher computational and usage costs of GPT-4 and Claude made their large-scale evaluation impractical within our resource constraints. These results highlight that domain-specific fine-tuning yields substantial gains in accuracy, robustness, and trustworthiness for food-related tasks. We begin by introducing the food-related tasks for which our specialized food LLM has been fine-tuned and can be effectively applied.

3.1. Assessing recipe nutritional profiles

Nutrient intake assessment evaluates dietary intake to determine whether nutritional needs are met (Neuhouser et al., 2023). It relies on food consumption data (Leclercq et al., 2019), linked to food composition databases (FCDB) (Merchant and Dehghan, 2006), to detect deficiencies or excesses affecting health and well-being. Consumption data informs dietary trends (Ravandi et al., 2025), nutrient intake (Saldanha et al., 2024), and guidelines (Näslund Koch et al., 2024), while also supporting studies on chronic diseases such as obesity (T Kennedy et al., 2001), diabetes (Brega et al., 2025), and cardiovascular conditions (Bermingham et al., 2024). FCDBs, containing nutrient profiles of foods, underpin nutrition labels, fortification programs, and balanced food design. Manual estimation links food entries in FCDBs to chemical analyses, but missing foods and nutrients remain common due to limited coverage. Imputation from studies or ML methods (Ispirova et al., 2020; Cenikj et al., 2023b) can fill gaps, yet complex recipes pose challenges. Predictive models using recipe text and ingredients help estimate nutrient values (Ispirova et al., 2024), though their accuracy depends on training data.

Here, we frame nutrient assessment as a question answering (QA) task. We fine-tuned an LLM to answer queries about recipe nutrient profiles under two prompt settings: (i) recipe title with ingredients and quantities, and (ii) ingredients with quantities only. Example prompts are provided below.

1. ***Prompt/Instruction:*** Find the nutrient breakdown per 100 g in Fred Harvey Cream Sauce containing the following ingredients: 12 cup soup, swanson chicken broth 99% fat free, 12 cup wheat flour, white, all purpose, unenriched, 3 cup milk, fluid, 1% fat, without added vitamin a and vitamin d, 1 cup soup, chicken broth or bouillon, dry.

Answer/Response: Nutritional values given per 100 g: energy - 121.02, fat - 0.79, protein - 4.16, salt - 1.70, saturates - 0.20, sugars - 1.25. 2. ***Prompt/Instruction:*** Ascertain the nutrient values per 100 g in a recipe that has these ingredients: 12 cup oil, olive, salad or cooking, 34 cup vinegar, cider, 1 1/2 cup sugars, granulated, 3 tablespoon onions, raw, 3 teaspoon mustard, prepared, yellow, 3 teaspoon spices, paprika, 3 teaspoon celery, raw, 3 teaspoon salt, table, 12 cup water, bottled, generic, 1 tablespoon shortening, vegetable, household, composite, 1 cup water, bottled, generic, 2 teaspoon cornstarch, 1 tablespoon corn starch, 1 cup soup, swanson chicken broth 99% fat free, 2 teaspoon cornstarch, 1 tablespoon cornstarch.

Answer/Response: Nutrient values in each 100 g are: energy - 179.57, fat - 18.08, protein - 0.02, salt - 0.15, saturates - 2.51, sugars - 2.34.

3.2. Classifying recipes by traffic light nutrition labels

The food traffic light system is a front-of-pack labeling scheme that uses colors—green (low), orange (medium), and red (high)—to indicate levels of fat, saturated fat, sugar, and salt (Balcombe et al., 2010). This intuitive design helps consumers make healthier choices quickly (Kunz et al., 2020) and has been adopted globally, including in Ecuador (Sarasty et al., 2023) and China (Lin et al., 2024), to address obesity, diabetes, and cardiovascular disease. Evidence shows that color-coded labels reduce cognitive load and increase purchases of healthier products (Sonnenberg et al., 2013), though challenges such as label fatigue and cultural variation remain. In the UK, manufacturers have reformulated products to achieve better ratings (Sacks et al., 2009), but concerns about oversimplification and economic impact persist. Extensions of the system now include eco-traffic lights to guide sustainable choices. Despite high awareness, disparities remain: children, older adults, and marginalized groups may face barriers to using the system effectively (Emrich et al., 2017). Future research explores integration with digital tools, behavioral nudges, and harmonization across countries.

In this study, we frame recipe classification under the traffic light system as a question-answering (QA) task, experimenting with two prompt types: (i) recipe title with ingredients and quantities, and (ii) ingredients with quantities only. Examples are provided below.

1. ***Prompt/Instruction:*** Find the FSA traffic light labels per 100 g in Spicy Peanut Sauce that includes these ingredients: 1 tablespoon shallots, raw, 2 teaspoon spices, garlic powder, 12 cup peanut butter, smooth style, without salt, 3 tablespoon oil, sesame, salad or cooking, 2 tablespoon soy sauce made from soy (tamari), 1 teaspoon spices, ginger, ground, 1 teaspoon roland, seasoned rice wine vinegar, upc: 041, 1/4- 1/2 teaspoon spices, pepper, red or cayenne, 13 cup soup, chicken broth or bouillon, dry.

Answer/Response: The Food Standards Agency traffic lights highlight per 100 g are: fat - red, salt - red, saturates - red, sugars - orange.

2. ***Prompt/Instruction:*** Review the FSA traffic lights per 100 g in a recipe using the following ingredients: 1/2 cup soup, swanson chicken broth 99% fat free, 1 pinch salt, table.

Answer/Response: FSA lights in 100 g: fat - green, salt - orange, saturates - green, sugars - green.

3.3. Making food data interoperable

Integrating food consumption and composition data requires adherence to FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Brinkley et al., 2025; Top et al., 2022), which align with EU policies on sustainable agriculture, health, and data sharing. Interoperability is particularly challenging and relies on ontologies that standardize vocabulary and semantics across datasets. Key food ontologies include FoodOn (Dooley et al., 2018), SNOMED-CT (Donnelly et al., 2006), the Hansard taxonomy (Parliament, 2022), FoodEx2 (EFSA), and FoodOntoMap (Popovski et al., 2019b). These resources enable consistent classification, linking, and analysis of food data across studies.

Natural Language Processing (NLP)-based information extraction supports FAIRification through named-entity recognition (NER) and named entity linking (NEL). Rule-based systems such as drNER (Eftimov et al., 2017) and FoodIE (Popovski et al., 2019a) extract food terms, while corpus-based approaches achieve higher accuracy: BuTTER (Popovski et al., 2019c) (F1 up to 94.6%), FoodNER (Stojanov et al., 2021) (F1 up to 94.3% for NER, 79% for semantic tags), and ensemble models like Food Voting NER (Cenikj et al., 2023a). Domain-specific models such as SciFoodNER (Cenikj et al., 2022a) further link food entities to ontologies. While GPT-based zero-shot NER shows promise (Floridi and Chiriatti, 2020; Achiam et al., 2023), performance drops significantly for ontology linking (Ogrinc et al., 2024).

In this study, we recast NER and NEL as question-answering (QA) tasks. For NER, the model identifies food entities in recipe descriptions or abstracts, then links them to FoodOn, SNOMED-CT, or Hansard. For NEL, the model directly maps provided food entities to ontology concepts. Examples of both prompting strategies are given below.

1. ***Prompt/Instruction:*** Find all food entities present in the text: Mix the cream cheese, beef, olives, onion, and Worcestershire sauce together in a bowl until evenly blended. Keeping the mixture in the bowl, scrape it into a semi-ball shape. Cover and refrigerate until firm, at least 2 h. Place a large sheet of waxed paper on a flat surface. Sprinkle with walnuts. Roll the cheese ball in the walnuts until completely covered. Transfer the cheese ball to a serving plate, or rewrap with waxed paper and refrigerate until needed.

Answer/Response: Certainly, the entities connected with food are outlined as follows: cream cheese, beef, olives, onion, Worcestershire sauce, walnuts, and cheese ball.

Prompt/Instruction: Is it within your capabilities to link the extracted food entities to a Hansard taxonomy?

Answer/Response: Certainly, the entities are associated properly: cream cheese - AG.01.e [Dairy produce]; AG.01.e.02 [Cheese]; AG.01.n [Dishes and prepared food]; AG.01.n.18 [Preserve], beef - AG.01.d.03 [Beef], olives - AG.01.h.01.e [Fruit containing stone], onion - AG.01.h.02.e [Onion/leek/garlic], Worcestershire sauce - AG.01.h [Fruit and vegeta bles]; AG.01.l.04 [Sauce/dressing], walnuts - AG.01.h.01.f [Nut], cheese ball - AG.01.e.02 [Cheese]; AG.01.n.18 [Preserve].

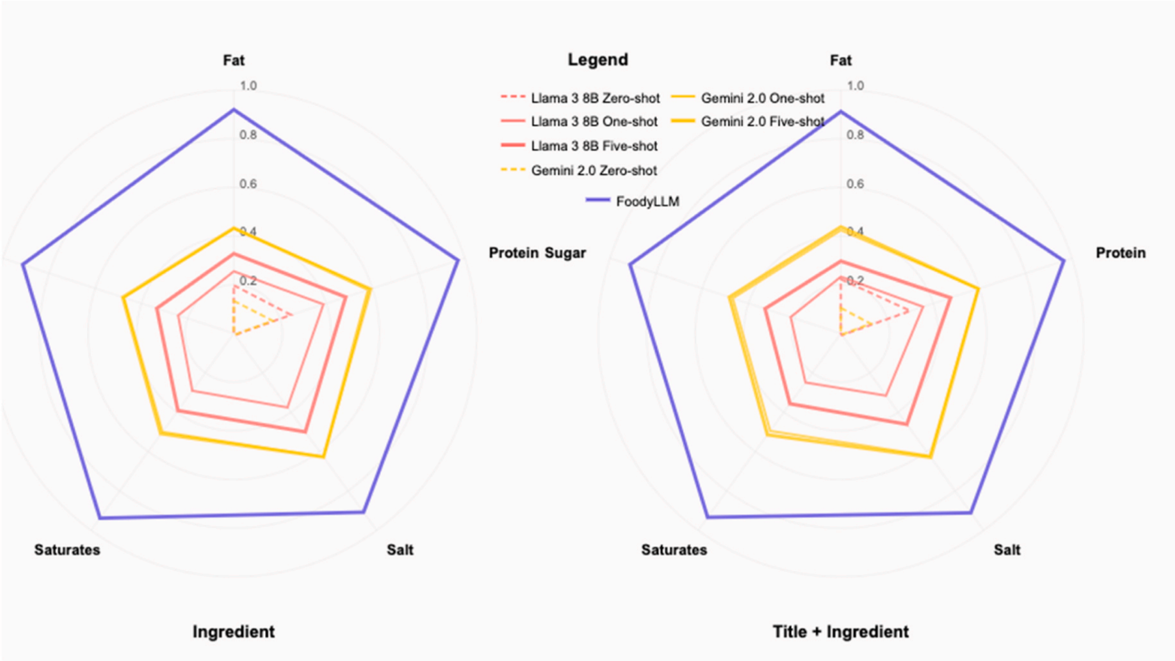
2. **Prompt/Instruction**: Would you mind sharing links to the Hansard taxonomy for these entities given: brandy, capers, leaves, custard, jalapeno pepper, vinegar, Parmesan, flax seeds, bay leaf?

Answer/Response: Definitely, the entities are linked in this manner: brandy - AG.01.y.01.e [Brandy], capers - AE.14.j.04 [Habits and ac tions of horse]; AF.13 [Particular tree/shrub], leaves - AF.10.e [Leaf], custard - AG.01.n.15 [Pastry]; AG.01.n.16 [Egg dishes], jalapeno pep per - AG.01.l.03 [Spice], vinegar - AG.01.ae.01 [Wine-making]; AG.01.l [Additive], Parmesan - AG.01.e.02 [Cheese]; AG.01.ak [Drinking], flax seeds - AF.10.i [Seed]; AF.24 [Yielding fibre/thatching/basket material], bay leaf - AF.10.e [Leaf]; AF.13 [Particular tree/shrub].

It is important to note that, in the case of the first prompt, the text provided for extracting food entities can be either a recipe instruction or a scientific abstract. For the linking part - both in the second part of the first prompt and in the second prompt - the default ontology used is the Hansard taxonomy. However, users have the flexibility to choose either FoodOn or SNOMED-CT for entity linking if preferred.

4. Results and discussion

We evaluate FoodyLLM five times to assess robustness. For recipe nutrient profiling and traffic-light classification, data was split into training and test sets five times with different seeds, using stratified sampling to reduce bias and prevent leakage (see Appendix A). For food data interoperability, the dataset was divided into five chunks with Scikit-learn; in each run, four chunks formed the training set and one the test set. The final training set was built by concatenating four interoperability chunks with one training split from the other tasks, then shuffling. Training data was further enriched with food synonym lists, cooking-unit conversions, nutrient-composition data, and non-food data from Dolly-HHRLHF (MosaicML, 2023) to mitigate catastrophic forgetting (Luo et al., 2023) and overfitting. Using these five enriched sets, we fine-tuned five models, each tested on its corresponding split. Performance was compared to the base (non-fine-tuned) model across all folds, using zero-shot, one-shot, and five-shot prompting. Baselines also included Gemini 2.0 Flash, a proprietary multimodal model with advanced reasoning, context handling, and tool-use capabilities (Zhao et al., 2025). A summary of results is shown in Fig. 1, with task-specific results detailed below and in the appendices.



a)

Fig. 1. a) Nutrient Accuracy Comparison of Zero-, One-, and Five-Shot Baselines versus FoodyLLM on the Ingredient dataset and Title + Ingredient dataset. FSA Unweighted Macro Average F1 Score Comparison of Zero-, One-, and Five-Shot Baselines versus FoodyLLM on b) Ingredient dataset and c) Title + Ingredient dataset. d) Weighted Macro-Averaged F1 Score Comparison of Zero-, One-, and Five-Shot Baselines versus FoodyLLM on NEL evaluation in artificially generated NEL samples, CafeteriaFCD, and CafeteriaSA.

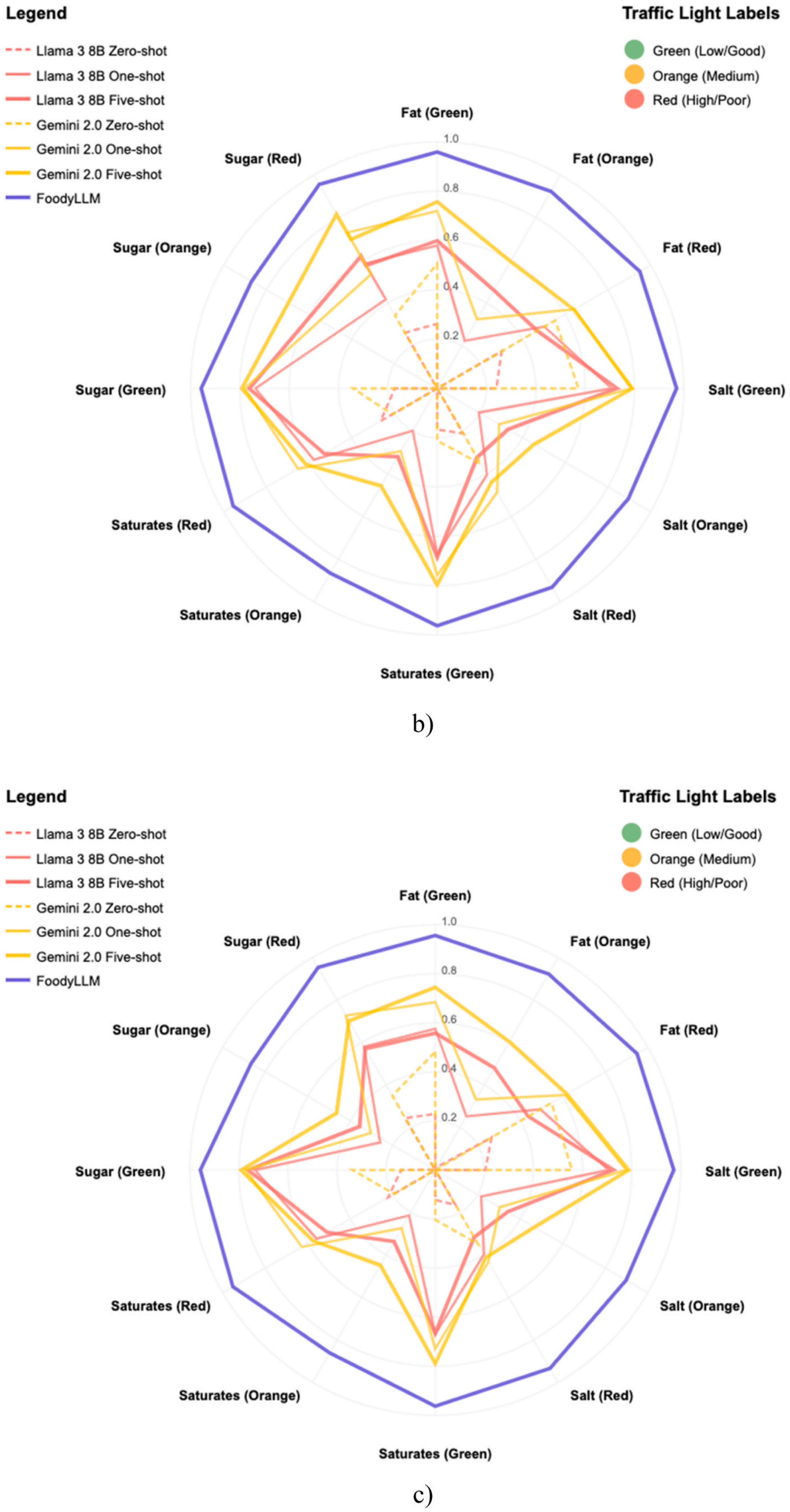
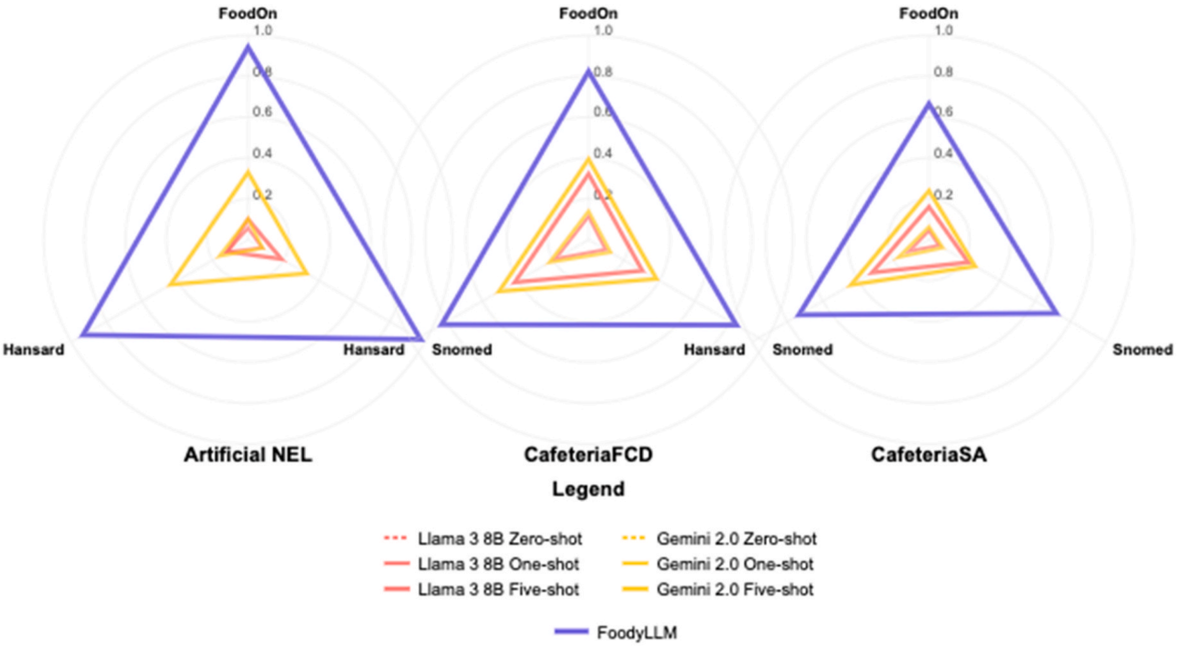


Fig. 1. (*continued*).



d)

Fig. 1. (*continued*).

4.1. *Assessing recipe nutritional profiles*

We evaluated FoodyLLM for nutrient value estimation using two prompt variants: (i) recipe title with ingredients and quantities (Title + Ingredient dataset) and (ii) ingredients with quantities only (Ingredient dataset). Predictions for fat, protein, salt, saturates, and sugar were compared to ground truth values.

Error tolerances followed EU Regulation 1169/2011 (Bairati, 2017), which specifies acceptable deviations for standard food products (see Appendix Tables 6 and 7). For each instance i, with true value t_i and model prediction r_i, we compute the absolute difference:

$$d_i = |t_i - r_i|. \tag{1}$$

A prediction is marked allowed if it satisfies tolerance thresholds (protein, sugar, fat in Eq. (2); saturates in Eq. (3)):

$$\begin{aligned} &allowed = 1 \ if: \\ &t_i \leq 10 \wedge \{ \ d_i \leq 2, for\ protein\ and\ sugar; \quad d_i \leq 1.5, for\ fat \\ &10 \leq t_i \leq 40 \wedge d_i \leq 0.2 \times t_i \\ &t_i > 40 \wedge d_i \leq 8 \end{aligned} \tag{2}$$

$$\begin{aligned} &allowed = 1 \ if: \\ &t_i < 4 \wedge d_i \leq 0.8 \quad for\ saturate \\ &t_i \geq 4 \wedge d_i \leq 0.2 \times t_i \end{aligned} \tag{3}$$

Accuracy is the proportion of predictions within tolerance:

$$Accuracy = \frac{\sum_{i=1}^{n} allowed}{n} \tag{4}$$

where n is the number of test instances.

Before the results are presented, we need to point out that the nutrient estimates produced by FoodyLLM should not be interpreted as direct recall of nutritional values from the training set, nor as explicit deterministic calculations based on ingredient proportions. At inference time, the model does not access a nutritional database or retrieve nutrient tables. Instead, as a transformer-based conditional language model fine-tuned on paired ingredient lists and nutritional profiles, it learns a latent mapping that approximates the conditional distribution of nutrient values given ingredient compositions and quantities. During training, ingredient and quantity tokens are embedded into a shared representation space in which correlations between ingredient patterns and target nutrient distributions are encoded through gradient-based optimization. Consequently, the model's predictions can be viewed as high-dimensional interpolation within a learned manifold of ingredient–nutrition relationships: for familiar ingredient combinations, outputs are guided by proximity to similar examples in the training data, while for novel combinations the model extrapolates from partially overlapping patterns. This mechanism differs fundamentally from rule-based nutritional computation, as no explicit nutrient-per-gram tables or symbolic summation procedures are encoded; instead, numerical outputs emerge from statistically learned associations captured in the model parameters. Accordingly, the predictions should be interpreted as learned estimates rather than analytically computed values, and reduced accuracy may be expected for rare ingredients or out-of-distribution recipes. Because FoodyLLM is formulated as a question-answering model rather than a conventional predictive pipeline, its behavior is expressed through conditioned input–output mappings on recipe descriptions, and model-level feature importance or parameter-based explainability is therefore not directly applicable in this setting.

Table 1 reports FoodyLLM's average accuracy in estimating fat, protein, salt, saturates, and sugar across five folds for the Ingredient and Title + Ingredient datasets. We compare it against a non-fine-tuned LLM and Gemini 2.0 using zero-, one-, and five-shot prompts, each including the test instruction plus zero, one, or five randomly selected training examples. Example prompts from the Ingredient dataset are shown below.

4.1.1. *Zero-shot prompt*

[INST]Review the nutrient values per 100 g in a recipe using these ingredients: 180 g wheat flour, bread, unenriched, 180 g wheat flour, white, cake, enriched, 200 ml sugars, granulated, 200 g butter, without salt, 200 ml water, bottled, generic, 1 pinch salt, table.[/INST]

Nutritional values in each 100 g: energy - 364.03, fat - 17.96, protein - 4.09, salt - 0.05, saturates - 10.95, sugars - 18.33.

Table 1
Tolerance-Based Average Accuracy by Class for the Ingredient and Title + Ingredient datasets on Nutrient Estimation.

Class	LLM	Experiment	Ingredient	Title + Ingredient
Fat	Llama 3 8B	Zero-shot prompting	0.198 ± 0.004	0.223 ± 0.002
	Llama 3 8B	One-shot prompting	0.255 ± 0.003	0.233 ± 0.003
	Llama 3 8B	Five-shot prompting	0.328 ± 0.009	0.297 ± 0.006
	Gemini 2.0	Zero-shot prompting	0.134 ± 0.003	0.105 ± 0.004
	Gemini 2.0	One-shot prompting	0.431 ± 0.008	0.424 ± 0.005
	Gemini 2.0	Five-shot prompting	0.433 ± 0.006	0.438 ± 0.006
	FoodyLLM		0.920 ± 0.004	0.912 ± 0.003
Protein	Llama 3 8B	Zero-shot prompting	0.247 ± 0.005	0.301 ± 0.004
	Llama 3 8B	One-shot prompting	0.388 ± 0.008	0.356 ± 0.006
	Llama 3 8B	Five-shot prompting	0.485 ± 0.006	0.475 ± 0.007
	Gemini 2.0	Zero-shot prompting	0.169 ± 0.002	0.129 ± 0.005
	Gemini 2.0	One-shot prompting	0.596 ± 0.005	0.596 ± 0.006
	Gemini 2.0	Five-shot prompting	0.585 ± 0.006	0.596 ± 0.005
	FoodyLLM		0.972 ± 0.002	0.967 ± 0.002
Salt	Llama 3 8B	Zero-shot prompting	0.004 ± 0.001	0.004 ± 0.001
	Llama 3 8B	One-shot prompting	0.376 ± 0.009	0.316 ± 0.003
	Llama 3 8B	Five-shot prompting	0.500 ± 0.005	0.462 ± 0.006
	Gemini 2.0	Zero-shot prompting	0.003 ± 0.001	0.003 ± 0.001
	Gemini 2.0	One-shot prompting	0.626 ± 0.005	0.621 ± 0.005
	Gemini 2.0	Five-shot prompting	0.628 ± 0.004	0.629 ± 0.007
	FoodyLLM		0.909 ± 0.005	0.910 ± 0.007
Saturates	Llama 3 8B	Zero-shot prompting	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	One-shot prompting	0.291 ± 0.004	0.249 ± 0.005
	Llama 3 8B	Five-shot prompting	0.393 ± 0.009	0.358 ± 0.003
	Gemini 2.0	Zero-shot prompting	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	One-shot prompting	0.501 ± 0.006	0.495 ± 0.001
	Gemini 2.0	Five-shot prompting	0.513 ± 0.004	0.515 ± 0.005
	FoodyLLM		0.938 ± 0.004	0.934 ± 0.004
Sugar	Llama 3 8B	Zero-shot prompting	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	One-shot prompting	0.242 ± 0.005	0.218 ± 0.008
	Llama 3 8B	Five-shot prompting	0.335 ± 0.007	0.329 ± 0.005
	Gemini 2.0	Zero-shot prompting	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	One-shot prompting	0.477 ± 0.006	0.469 ± 0.006
	Gemini 2.0	Five-shot prompting	0.482 ± 0.007	0.484 ± 0.009
	FoodyLLM		0.917 ± 0.005	0.915 ± 0.008

4.1.2. *One-shot prompt*

[INST]*The following are examples of questions (with answers) about nutrition.* Question: Determine the nutritional profile per 100 g in a recipe that uses these ingredients: 1 cup cheese, gouda, 4 tablespoon butter, without salt, 3/4 cup wheat flour, white, all-purpose, unenriched, 1/2 teaspoon salt, table, 1/2 teaspoon spices, pepper, red or cayenne, 1 tablespoon cream, fluid, heavy whipping Answer: Nutrient values highlighted for 100 g: energy - 426.12, fat - 29.33, protein - 14.52, salt - 1.92, saturates - 18.33, sugars - 1.18 *Respond to the following question in the same manner as seen in the examples above.* Question: Review the nutrient values per 100 g in a recipe using these ingredients: 180 g wheat flour, bread, unenriched, 180 g wheat flour, white, cake, enriched, 200 ml sugars, granulated, 200 g butter, without salt, 200 ml water, bottled, generic, 1 pinch salt, table [/INST]

Nutritional values in each 100 g: energy - 364.03, fat - 17.96, protein - 4.09, salt - 0.05, saturates - 10.95, sugars - 18.33.

4.1.3. *Five-shot prompt*

[INST]*The following are examples of questions (with answers) about nutrition.* Question: Gauge the nutri ent values per 100 g in a recipe prepared with the following ingredients: 2 cup cream, fluid, heavy whipping, 1 tablespoon spices, cardamom Answer: Per 100 g, the nutrient values are: energy - 339.02, fat - 35.36, protein - 3.04, salt - 0.07, saturates - 22.49, sugars - 2.85 Question: Establish the nutrient profile per 100 g in a recipe containing these ingredients: 1 tablespoon shallots, raw, 2 teaspoon spices, garlic powder, 12 cup peanut butter, smooth style, without salt, 3 tablespoon oil, sesame, salad or cooking, 2 tablespoon soy sauce made from soy (tamari), 1 teaspoon spices, ginger, ground, 1 teaspoon roland, seasoned rice wine vinegar, upc: 041224705142, 1/4-1/2 teaspoon spices, pepper, red or cayenne, 13 cup soup, chicken broth or bouillon, dry Answer: Nutrient profile for every 100 g: energy - 494.83, fat - 40.58, protein - 20.22, salt - 17.01, saturates - 8.28, sugars - 12.29 Question: Verify the nutrient values per 100 g in a recipe prepared with these ingredients: 16 ounce milk, fluid, 1% fat, without added vitamin a and vitamin d, 8 ounce beverages, almond milk, unsweetened, shelf stable, 13 cup sugars, granulated, 14 cup cornstarch, 12 teaspoon vanilla extract, 14 teaspoon shortening confectionery, coconut (hydrogenated) and or palm kernel (hydrogenated) Answer: Nutrient facts per 100 g: energy - 340.40, fat - 1.30, protein - 0.41, salt - 0.03, saturates - 1.11, sugars - 50.82 Question: Identify the nutritional composition per 100 g in a recipe with these ingredients: 500 g ground turkey, raw, 1 cup onions, raw, 12 cup bread crumbs, dry, grated, plain, 12 cup carrots, raw, 12 cup sauce, barbecue, 2 teaspoon sauce, worcestershire, 1 teaspoon spices, garlic powder, 34 teaspoon spices, pepper, black Answer: Nutrient profile for each 100 g: energy - 180.22, fat - 1.96, protein - 4.72, salt - 1.63, saturates - 0.42, sugars - 18.20 Question: Find the nutri tional breakdown per 100 g in a recipe that uses the following ingredients: 12 pound pretzels, soft, unsalted, 21 ounce corn, sweet, white, raw, 12 ounce cookies, graham crackers, plain or honey, lowfat, 32 ounce nuts, walnuts, english, 7 ounce cookies, graham crackers, plain or honey, lowfat, 12 ounce cookies, graham crackers, plain or honey, lowfat, 1 pound butter, without salt, 12 drop sauce, ready-to-serve, pepper, tabasco, 8 tablespoon sugars, brown, 2 teaspoon spices, chili powder, 2 teaspoon sauce, worcestershire, 2-3 tablespoon spices, garlic powder Answer: The nutrient breakdown per 100 g is: energy - 383.98, fat - 14.10, protein - 7.92, salt - 0.59, saturates - 3.85, sugars - 3.92 *Respond to the following question in the same manner as seen in the examples above.* Question: Review the nutrient values per 100 g in a recipe using these ingredients: 180 g wheat flour, bread, unenriched, 180 g wheat flour, white, cake, enriched, 200 ml sugars, granulated, 200 g butter, without salt, 200 ml water, bottled, generic, 1 pinch salt, table [/INST]

Nutritional values in each 100 g: energy - 364.03, fat - 17.96, protein - 4.09, salt - 0.05, saturates - 10.95, sugars - 18.33.

Table 1 shows that FoodyLLM consistently outperforms the non-fine-tuned LLM (zero-, one-, five-shot) on the Ingredient dataset, achieving the highest accuracy for protein (97.2%), followed by saturates (93.8%), fat (92.0%), sugar (91.7%), and salt (90.9%). Similar results are seen on the Title + Ingredient dataset (protein 96.7%, saturates 93.4%, sugar 91.5%, fat 91.2%, salt 91.0%), though adding titles slightly reduces accuracy (1%). These results highlight the limitations of general-purpose LLMs and the need for domain-specific fine-tuning in food science. FoodyLLM also surpasses Gemini 2.0 across all nutrients, with tolerance-based accuracies from 90.9% to 97.2% (vs. Gemini's 43.3–62.8% even with five-shot prompts). Improvements are substantial: fat (92.0% vs. 43.3%), protein (97.2% vs. 59.6%), salt (90.9% vs. 62.8%), saturates (93.8% vs. 51.3%), sugar (91.7% vs. 48.2%). Comparable trends are observed for the Title + Ingredient dataset. These findings, visualized in Fig. 1a and detailed in Fig. 2 (Appendix C), confirm the effectiveness of domain-specific fine-tuning for nutrient estimation.

Generalization of FoodyLLM to branded foods: We further evaluated FoodyLLM on branded food products, which typically list ingredients without quantities, unlike our recipe-based training data. This tests the model's ability to generalize beyond its training context. We used 2070 branded products from the Open Food Facts API,[2] each including product name, ingredients, nutrient values, and food group. FoodyLLM was prompted with product name and ingredients only, under zero-, one-, and five-shot settings, using tolerance-based evaluation. Zero-shot accuracies were 29.3% (fat), 46.1% (protein), 39.6% (salt), 41.4% (saturates), and 34.1% (sugar). One shot slightly improved results to 30.1%, 44.6%, 39.3%, 42.0%, and 37.7%, while five-shot produced 29.1%, 45.4%, 40.0%, 41.2%, and 35.6%, showing marginal differences. Overall, one-shot offered slight gains, but additional examples did not substantially improve performance. The lower performance on branded food products mainly reflects differences in how these foods are described compared to recipes. FoodyLLM was trained primarily on recipe-style data, where ingredients are listed together with their quantities, which provides sufficient information to estimate nutrients and health labels. In contrast, branded food products usually report only ingredient lists without quantities, resulting in much less informative inputs. When only partial information is available, accurate nutritional inference becomes inherently more difficult, which explains the observed drop in performance. This indicates that FoodyLLM is currently better suited to recipe-style inputs, while improved performance on packaged foods will require either estimating ingredient quantities or further fine-tuning the model on branded food data.

4.2. Classifying recipes by traffic light nutrition labels

Tables 2 and 3 report FoodyLLM's performance on the Ingredient dataset (see Tables 8 and 9 from Appendix D for Title + Ingredient). We frame this as a multi-class classification task, computing macro precision, recall, and F1 across fat, salt, saturates, and sugar with traffic-light labels (red, orange, green). Macro scores average per-class values, ensuring fair evaluation under class imbalance. FoodyLLM consistently outperforms all non-fine-tuned baselines (zero-, one-, five-shot) across nutrients and labels. For the Ingredient dataset, macro F1 scores reached 94.9% (fat-red), 93.2% (salt-red), 95.5% (saturates-red), 95.6% (sugar-red); 92.3% (fat-orange), 89.5% (salt orange), 86.5% (saturates-orange), 86.8% (sugar-orange); and 95.9% (fat green), 97.1% (salt-green), 96.3% (saturates-green), 95.7% (sugar-green). Adding recipe titles yields similar trends with slightly lower scores. Compared to Gemini 2.0, FoodyLLM achieves markedly higher macro F1 (86.8–97.1% vs. 45.3–79.7%), even though Gemini performs best with five-shot prompts. Results are visualized in Fig. 1b,c, and Figs. 3 and 4 from Appendix D. FoodyLLM's main weakness lies in the orange category, where synonym use (e.g., “amber", “yellow") lowers apparent accuracy. To ensure fairness, we normalized synonyms only for baselines (e.g., “low"→green, “high”→red, “am ber/yellow/medium/moderate"→orange), yielding slight improvements (see Tables 13–16 from Appendix I), but FoodyLLM still maintains a clear performance lead. Fig. 8 (Appendix I) shows the distribution of synonyms used by the Llama3 8B baseline model to represent the "orange" traffic light label across all four macronutrient classes (fat, salt, saturates, sugars) aggregated over all 5 cross-validation folds. Since in practical use the model is intended to support expert judgment. Consequently, this limitation does not necessarily require additional fine-tuning of the model, as it can be effectively addressed through post-hoc normalization and expert interpretation.

4.3. Making food data interoperable

We evaluated food interoperability on two tasks using three ontologies – SNOMED-CT, FoodOn, and the Hansard taxonomy. In the NEL-only task, the model linked given food entities to ontology concepts. In the NER + NEL task, it first extracted food mentions from text data and then mapped them. This dual setup assessed both concept-mapping accuracy and entity recognition, highlighting strengths and limitations in automated food knowledge integration.

4.3.1. Named entity linking

The five-fold cross-validation NEL results are shown in Table 10 (Appendix E). In this task, the model links food entity mentions to ontology concepts from FoodOn, SNOMED CT, and Hansard. Baselines included a non-fine-tuned Llama 3 with zero-, one-, and five-shot prompts sampled from training splits. Llama 3 failed to link entities in zero-shot, showed minor gains with one-shot, and achieved better precision with five-shot, but low recall, mapping only a small portion of entities. FoodyLLM clearly outperforms both Gemini 2.0 and non-fine-tuned Llama 3, reaching macro F1 scores of 0.942 (FoodOn), 0.975 (SNOMED), and 0.932 (Hansard), versus Gemini's best 0.330, 0.330, and 0.438 (five-shot) (see Table 4). Crucially, FoodyLLM achieves these results without examples, underscoring the value of domain-specific fine-tuning over general-purpose LLMs, which struggle even with few-shot prompts (see Table 4). Results are visualized in Figs. 1d and 5 from Appendix E.

4.3.2. Named entity recognition and linking

We also evaluated a two-step NER + NEL setup: the LLM first recognized food entities in text, then linked them to FoodOn, SNOMED CT, and Hansard. Two datasets were used: CafeteriaFCD (a corpus of annotated recipe instructions) (Ispirova et al., 2022) and CafeteriaSA (a corpus of annotated scientific abstracts) (Cenikj et al., 2022b) (see Table 4). Performance was assessed via five-fold cross-validation (Tables 11 and 12 from Appendix E. On CafeteriaFCD, FoodyLLM's NEL scores were slightly lower than on artificial samples, mainly due to reduced recall—possibly from the added NER step or dataset difficulty (future work will clarify). Baselines achieved results similar to artificial NEL in zero-shot, but one- and five-shot prompting improved over prior baselines (Table 10). On CafeteriaSA, FoodyLLM scored lower than on CafeteriaFCD, especially in recall, with baselines also dropping, again likely due to dataset complexity. Across both datasets, FoodyLLM clearly outperforms Gemini 2.0 and non-fine-tuned Llama 3 8B, achieving macro-F1 scores of 0.665–0.835 vs. Gemini's 0.240–0.505 and Llama 3's 0.161–0.417 (five-shot). Notably, general-purpose models showed sharp performance declines on CafeteriaSA, while FoodyLLM remained strong, underscoring the value of domain-specific fine-tuning for food-entity linking. Results are shown in Fig. 1d (and Fig. 5 from Appendix E).

In our discussion, we also consider previously reported NEL outcomes on CafeteriaFCD using BERT and BioBERT (Stojanov et al., 2021). For the combined NER + NEL evaluations, the macro-weighted F1 scores were 0.781 for FoodOn with BERT, 0.789 for Hansard with BioBERT-large, and 0.761 for SNOMED-CT with BERT. On CafeteriaSA, BioBERT achieved macro-weighted F1 scores of 0.430 for FoodOn, 0.660 for Hansard, and 0.580 for SNOMED-CT (Cenikj et al., 2022a).

---

[2] https://world.openfoodfacts.org/.

Table 2
Results by class for the Ingredient dataset (1/2).

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Fat (Green)	Llama 3 8B	0-shot prompt.	0.542 ± 0.021	0.173 ± 0.010	0.262 ± 0.014
	Llama 3 8B	1-shot prompt.	0.545 ± 0.010	0.621 ± 0.009	0.580 ± 0.007
	Llama 3 8B	5-shot prompt.	0.643 ± 0.008	0.562 ± 0.012	0.599 ± 0.007
	Gemini 2.0	0-shot prompt.	0.657 ± 0.009	0.412 ± 0.034	0.506 ± 0.029
	Gemini 2.0	1-shot prompt.	0.826 ± 0.012	0.638 ± 0.016	0.720 ± 0.014
	Gemini 2.0	5-shot prompt.	0.811 ± 0.003	0.710 ± 0.012	0.757 ± 0.007
	FoodyLLM		0.964 ± 0.007	0.953 ± 0.012	0.959 ± 0.003
Fat (Orange)	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.442 ± 0.026	0.151 ± 0.009	0.224 ± 0.013
	Llama 3 8B	5-shot prompt.	0.440 ± 0.006	0.483 ± 0.015	0.460 ± 0.009
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.572 ± 0.019	0.226 ± 0.006	0.324 ± 0.009
	Gemini 2.0	5-shot prompt.	0.555 ± 0.008	0.642 ± 0.017	0.595 ± 0.011
	FoodyLLM		0.919 ± 0.008	0.928 ± 0.005	0.923 ± 0.004
Fat (Red)	Llama 3 8B	0-shot prompt.	0.366 ± 0.009	0.257 ± 0.003	0.302 ± 0.005
	Llama 3 8B	1-shot prompt.	0.429 ± 0.004	0.607 ± 0.015	0.502 ± 0.008
	Llama 3 8B	5-shot prompt.	0.463 ± 0.014	0.473 ± 0.006	0.468 ± 0.008
	Gemini 2.0	0-shot prompt.	0.634 ± 0.011	0.490 ± 0.040	0.552 ± 0.029
	Gemini 2.0	1-shot prompt.	0.634 ± 0.007	0.652 ± 0.010	0.643 ± 0.006
	Gemini 2.0	5-shot prompt.	0.663 ± 0.013	0.618 ± 0.011	0.640 ± 0.009
	FoodyLLM		0.948 ± 0.011	0.949 ± 0.012	0.949 ± 0.002
Salt (Green)	Llama 3 8B	0-shot prompt.	0.615 ± 0.007	0.149 ± 0.004	0.240 ± 0.005
	Llama 3 8B	1-shot prompt.	0.614 ± 0.003	0.834 ± 0.006	0.707 ± 0.004
	Llama 3 8B	5-shot prompt.	0.631 ± 0.010	0.865 ± 0.004	0.730 ± 0.008
	Gemini 2.0	0-shot prompt.	0.761 ± 0.012	0.460 ± 0.032	0.573 ± 0.028
	Gemini 2.0	1-shot prompt.	0.772 ± 0.008	0.807 ± 0.003	0.789 ± 0.003
	Gemini 2.0	5-shot prompt.	0.748 ± 0.009	0.838 ± 0.007	0.791 ± 0.007
	FoodyLLM		0.973 ± 0.007	0.970 ± 0.006	0.971 ± 0.002
Salt (Orange)	Llama 3 8B	0-shot prompt.	0.200 ± 0.447	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.392 ± 0.019	0.130 ± 0.005	0.196 ± 0.008
	Llama 3 8B	5-shot prompt.	0.391 ± 0.006	0.289 ± 0.012	0.332 ± 0.010
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.473 ± 0.017	0.210 ± 0.010	0.290 ± 0.011
	Gemini 2.0	5-shot prompt.	0.459 ± 0.012	0.447 ± 0.007	0.453 ± 0.009
	FoodyLLM		0.892 ± 0.016	0.899 ± 0.013	0.895 ± 0.006
Salt (Red)	Llama 3 8B	0-shot prompt.	0.365 ± 0.012	0.148 ± 0.009	0.211 ± 0.010
	Llama 3 8B	1-shot prompt.	0.475 ± 0.008	0.352 ± 0.013	0.404 ± 0.011
	Llama 3 8B	5-shot prompt.	0.481 ± 0.021	0.244 ± 0.017	0.323 ± 0.019
	Gemini 2.0	0-shot prompt.	0.622 ± 0.018	0.246 ± 0.018	0.353 ± 0.020
	Gemini 2.0	1-shot prompt.	0.490 ± 0.007	0.483 ± 0.023	0.486 ± 0.014
	Gemini 2.0	5-shot prompt.	0.507 ± 0.016	0.389 ± 0.020	0.440 ± 0.018
	FoodyLLM		0.933 ± 0.005	0.930 ± 0.013	0.932 ± 0.005

Although the cross-validation folds differ from those in our study, these results provide a useful benchmark signal. These approaches were first proposed by most of the authors of this paper in their earlier work on NEL, and our current study addresses and overcomes the limitations identified in those initial attempts.

5. FoodyLLM user interface

To lower the barrier for non-technical users, we additionally provide access to a centrally hosted Ollama instance that enables interaction with FoodyLLM through a web-based, ChatGPT-like interface without local installation. To demonstrate practical usability, we provide example application scenarios using the FoodyLLM interface (see Fig. 6 from Appendix F), showing how users can submit recipe text and obtain structured nutritional and semantic outputs, while a formal usability evaluation with domain experts is left for future work. Ollama is an open-source framework for running LLMs locally on macOS, Linux, and Windows, offering a command-line interface and REST API for managing pre-built and custom models without cloud dependency. It supports models such as Llama 3.3, Qwen, DeepSeek-R1, and Gemma, enabling local text generation, embeddings, and AI applications. The current Ollama-based interface should be regarded as a research prototype intended for reproducibility and experimental use rather than as a fully optimized end-user application. Running FoodyLLM locally requires sufficient computational resources to host an 8B-parameter model (typically a modern GPU with at least 16 GB of VRAM, or a high-memory CPU with slower inference), as well as disk space for model weights.

6. Discussion

FoodyLLM's performance is intrinsically linked to the quality, completeness, and temporal validity of the data and knowledge resources used for its training (see Appendix A). Consequently, the model can only provide accurate outputs for foods, nutrients, and concepts that are adequately represented in these underlying datasets and ontologies. Updates to key resources, such as food composition databases or food and health ontologies, are not automatically reflected in the model and require explicit updating or re-training to incorporate new information. This dependency on fixed reference resources is a general limitation of supervised AI models for food and biomedical text analysis, but it can be partially mitigated by integrating FoodyLLM with external knowledge retrieval mechanisms (e.g., retrieval-augmented approaches), which allow the model to access and reason over up-to-date databases and ontologies at inference time without re-training (Drole et al., 2025). Furthermore, the ability of FoodyLLM to generalize to new data sources or application contexts depends on their alignment with the training data; when this alignment is weak, additional domain-specific training (i.e., fine-tuning) is required, whereas when alignment is strong, it indicates that the original training data provide a suitable knowledge basis for transferring nutritional and semantic information to new food applications.

FoodyLLM mitigates recipe parsing inconsistencies through its

Table 3
Results by class for the Ingredient dataset (2/2).

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Saturates (Green)	Llama 3 8B	0-shot prompt.	0.631 ± 0.047	0.096 ± 0.008	0.167 ± 0.014
	Llama 3 8B	1-shot prompt.	0.677 ± 0.006	0.668 ± 0.010	0.673 ± 0.005
	Llama 3 8B	5-shot prompt.	0.718 ± 0.010	0.656 ± 0.012	0.685 ± 0.005
	Gemini 2.0	0-shot prompt.	0.822 ± 0.028	0.122 ± 0.010	0.212 ± 0.016
	Gemini 2.0	1-shot prompt.	0.870 ± 0.005	0.673 ± 0.009	0.758 ± 0.005
	Gemini 2.0	5-shot prompt.	0.820 ± 0.003	0.774 ± 0.006	0.797 ± 0.005
	FoodyLLM		0.961 ± 0.008	0.965 ± 0.007	0.963 ± 0.003
Saturates (Orange)	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.302 ± 0.021	0.147 ± 0.011	0.198 ± 0.014
	Llama 3 8B	5-shot prompt.	0.283 ± 0.011	0.365 ± 0.021	0.319 ± 0.014
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.410 ± 0.014	0.231 ± 0.006	0.295 ± 0.007
	Gemini 2.0	5-shot prompt.	0.373 ± 0.016	0.585 ± 0.015	0.456 ± 0.016
	FoodyLLM		0.862 ± 0.016	0.867 ± 0.013	0.865 ± 0.009
Saturates (Red)	Llama 3 8B	0-shot prompt.	0.431 ± 0.009	0.191 ± 0.009	0.265 ± 0.009
	Llama 3 8B	1-shot prompt.	0.544 ± 0.011	0.617 ± 0.012	0.578 ± 0.008
	Llama 3 8B	5-shot prompt.	0.555 ± 0.005	0.505 ± 0.014	0.529 ± 0.007
	Gemini 2.0	0-shot prompt.	0.772 ± 0.038	0.122 ± 0.011	0.210 ± 0.018
	Gemini 2.0	1-shot prompt.	0.756 ± 0.012	0.574 ± 0.013	0.652 ± 0.010
	Gemini 2.0	5-shot prompt.	0.782 ± 0.007	0.508 ± 0.009	0.615 ± 0.005
	FoodyLLM		0.959 ± 0.002	0.950 ± 0.009	0.955 ± 0.004
Sugar (Green)	Llama 3 8B	0-shot prompt.	0.595 ± 0.021	0.100 ± 0.005	0.172 ± 0.008
	Llama 3 8B	1-shot prompt.	0.732 ± 0.006	0.741 ± 0.005	0.737 ± 0.003
	Llama 3 8B	5-shot prompt.	0.779 ± 0.007	0.746 ± 0.010	0.762 ± 0.007
	Gemini 2.0	0-shot prompt.	0.826 ± 0.006	0.215 ± 0.009	0.341 ± 0.012
	Gemini 2.0	1-shot prompt.	0.861 ± 0.007	0.712 ± 0.005	0.780 ± 0.003
	Gemini 2.0	5-shot prompt.	0.860 ± 0.008	0.735 ± 0.003	0.792 ± 0.004
	FoodyLLM		0.959 ± 0.003	0.955 ± 0.005	0.957 ± 0.002
Sugar (Orange)	Llama 3 8B	0-shot prompt.	0.067 ± 0.149	0.000 ± 0.000	0.000 ± 0.001
	Llama 3 8B	1-shot prompt.	0.317 ± 0.019	0.195 ± 0.010	0.241 ± 0.013
	Llama 3 8B	5-shot prompt.	0.331 ± 0.009	0.387 ± 0.012	0.357 ± 0.009
	Gemini 2.0	0-shot prompt.	0.100 ± 0.224	0.000 ± 0.000	0.000 ± 0.001
	Gemini 2.0	1-shot prompt.	0.425 ± 0.019	0.243 ± 0.014	0.309 ± 0.016
	Gemini 2.0	5-shot prompt.	0.417 ± 0.009	0.541 ± 0.015	0.471 ± 0.010
	FoodyLLM		0.863 ± 0.008	0.872 ± 0.008	0.868 ± 0.002
Sugar (Red)	Llama 3 8B	0-shot prompt.	0.428 ± 0.013	0.188 ± 0.013	0.261 ± 0.014
	Llama 3 8B	1-shot prompt.	0.567 ± 0.010	0.603 ± 0.013	0.585 ± 0.011
	Llama 3 8B	5-shot prompt.	0.606 ± 0.006	0.548 ± 0.004	0.576 ± 0.004
	Gemini 2.0	0-shot prompt.	0.770 ± 0.022	0.220 ± 0.026	0.342 ± 0.033
	Gemini 2.0	1-shot prompt.	0.657 ± 0.008	0.823 ± 0.008	0.730 ± 0.003
	Gemini 2.0	5-shot prompt.	0.709 ± 0.005	0.689 ± 0.011	0.699 ± 0.005
	FoodyLLM		0.957 ± 0.005	0.955 ± 0.001	0.956 ± 0.003

Table 4
Average five-fold cross-validation macro avg. F1 scores across different datasets (standard deviation given in brackets). In cases of CafeteriaFCD and CafeteriaSA, the NEL task is performed after the NER task.

Test Set	LLM	Experiment	Artificial NEL	CafeteriaFCD	CafeteriaSA
FoodOn	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.057 ± 0.002	0.113 ± 0.014	0.045 ± 0.011
	Llama 3 8B	5-shot prompt.	0.099 ± 0.006	0.322 ± 0.003	0.161 ± 0.019
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.102 ± 0.005	0.140 ± 0.018	0.063 ± 0.020
	Gemini 2.0	5-shot prompt.	0.330 ± 0.008	0.394 ± 0.012	0.240 ± 0.009
	FoodyLLM		0.942 ± 0.001	0.823 ± 0.007	0.665 ± 0.029
	Literature baseline	/	/	0.781 (BERT)	0.430 (BioBERT)
Snomed	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.080 ± 0.007	0.096 ± 0.004	0.056 ± 0.010
	Llama 3 8B	5-shot prompt.	0.188 ± 0.006	0.304 ± 0.011	0.222 ± 0.026
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.082 ± 0.004	0.124 ± 0.005	0.078 ± 0.012
	Gemini 2.0	5-shot prompt.	0.330 ± 0.005	0.384 ± 0.019	0.260 ± 0.027
	FoodyLLM		0.975 ± 0.001	0.835 ± 0.017	0.720 ± 0.013
	Literature baseline	/	/	0.761 (BERT)	0.580 (BioBERT)
Hansard	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.121 ± 0.009	0.187 ± 0.017	0.111 ± 0.016
	Llama 3 8B	5-shot prompt.	0.117 ± 0.005	0.417 ± 0.016	0.321 ± 0.028
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.162 ± 0.011	0.221 ± 0.014	0.172 ± 0.016
	Gemini 2.0	5-shot prompt.	0.438 ± 0.011	0.505 ± 0.016	0.441 ± 0.036
	FoodyLLM		0.932 ± 0.004	0.830 ± 0.018	0.735 ± 0.023
	Literature baseline	/	/	0.789 (BioBERT-large)	0.660 (BioBERT)

explicit evaluation on NER and NEL tasks, which assess robust extraction and normalization of ingredient mentions from heterogeneous recipe text, and it partially addresses ingredient quantity ambiguity by learning statistical associations between informal household measures and nutrient values during fine-tuning. However, cultural and language bias in food terminology was not explicitly analyzed in this study and remains an important direction for future work, as the current model primarily reflects the linguistic and cultural distribution of the underlying recipe datasets.

The performance gains obtained through FoodyLLM reflect a fundamental mismatch between the training data of general-purpose LLMs and the requirements of food and nutrition tasks. General-purpose models are predominantly trained on broad web, code, and general scientific text and therefore lack systematic exposure to structured food composition data and recipe–nutrition mappings needed for nutrient estimation, food classification, and semantic interpretation. While domain-specific jargon and the need for numerical precision contribute to this gap, the dominant factor is the limited availability of large-scale, open, and semantically structured food and nutrition datasets compared with biomedical domains that benefit from mature semantic infrastructures (e.g., controlled vocabularies and aligned text–value corpora, Lindberg et al., 1993). As a result, general-purpose LLMs have insufficient coverage of the latent relationships between ingredients, quantities, and nutritional outcomes. Fine-tuning on semantically rich, task-aligned resources such as Recipe1M + enables the model to internalize these domain-specific relations, leading to substantial performance improvements. These findings indicate that the gains arise primarily from exposure to structured and semantically grounded food-related data that is largely absent from generic pre-training corpora, underscoring the importance of open, FAIR food and nutrition datasets for advancing AI-driven applications in this domain.

Although the proposed approach achieves strong performance on recipe-based nutrient estimation tasks, the remaining error rates are still non-negligible from a clinical and public health perspective. Even modest systematic errors may accumulate across meals and days, potentially leading to misleading dietary assessments, especially for populations requiring strict nutritional control (e.g., individuals with diabetes or cardiovascular risk). Consequently, the current system should be interpreted as providing approximate estimates suitable for large-scale, automated analyses such as population-level screening, trend detection, or hypothesis generation, rather than clinically actionable values. Use in clinical or regulatory settings would require substantially lower error rates, rigorous external validation, and alignment with authoritative food composition databases. To this end, in line with European regulatory frameworks such as the AI Act and GDPR, we position the proposed system as a semi-automatic decision-support tool whose outputs require human oversight, thereby limiting risks from automated nutrient misestimation, preserving accountability in public health use, and acknowledging potential cultural bias in dietary recommendations arising from the training data.

Within emerging agentic AI systems, architectures in which multiple specialized AI models act as task-oriented agents that are selected and coordinated according to the application goal, FoodSky and FoodyLLM play complementary roles rather than competing as general-purpose models. FoodSky is suited to recipe and cooking-related tasks, whereas FoodyLLM supports analytical, nutrition-focused, and ontology-driven workflows relevant to food and nutrition science. Accordingly, the novelty of FoodyLLM lies in its nutrition-oriented, ontology-aware, and multi-task design as a specialized analytical agent, rather than in cuisine-specific or generative culinary assistance.

7. Conclusion

In conclusion, FoodyLLM, fine-tuned exclusively on food and nutrition data, clearly outperforms general-purpose LLMs on nutrient estimation, traffic-light labeling, and ontology linking, even in few-shot settings. Its strength lies in combining diverse food ontologies, nutrient databases, and labeling schemes into a unified multi-task dataset, enabling accurate and reliable zero-shot predictions and making it suitable for applications in consumer apps, clinical decision support, and research pipelines where precision is critical. Some limitations remain in the interoperability task, as FoodyLLM can currently recognize only food annotations present in the training data; newly introduced ontology concepts that are absent from the training corpus cannot yet be linked.

While its current scope is limited to three core tasks, the proposed framework is not inherently restricted to macronutrients. Future work will extend FoodyLLM to additional targets, including micronutrients (e.g., sodium, vitamins, and minerals), food additives, allergen information, and food processing indicators such as the NOVA classification, by fine-tuning on appropriately curated and semantically aligned datasets without requiring changes to the underlying model architecture. Additionaly, we will extend FoodyLLM with more culturally and geographically diverse recipe data as soon as suitably licensed public datasets become available, in compliance with EU AI Act and GDPR requirements. In parallel, we will explore retrieval-augmented generation (RAG) and agentic architectures to broaden knowledge coverage and support more complex workflows such as personalized meal planning and dietary monitoring. Together, these directions highlight the broader potential of domain-specific fine-tuning as a pathway toward trustworthy, real-world AI for food and nutrition science.

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors approved the submitted version.

Materials availability

Not applicable.

Code availability

https://github.com/matejMartinc/FoodyLLM.

Author contribution

AG: Writing - original draft, Methodology, Investigation, Conceptualization; MM: Writing - original draft, Methodology, Investigation, Conceptualization, Software, Revision; GC: Writing – review & editing, Investigation, Validation; RS: Writing – review & editing, Investigation, Validation, Software; JD: Writing – review & editing, Investigation, Validation; GI: Writing – review & editing, Investigation, Validation; GM: Writing – review & editing, Investigation, Validation; NO: Writing – review & editing, Investigation, Validation; DT: Writing – review & editing, Investigation, Validation; SD: Writing – review & editing, Conceptualization, Supervision; BKS: Writing - original draft, Conceptualization, Supervision, Revision; TE: Writing - original draft, Visualization, Supervision, Methodology, Investigation, Conceptualization, Revision.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to check and correct grammar. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the support of the Slovenian Research Agency through program grants No. P2-0098, No. P2-0103 and P1-0143, project grants No. J7-70265, No. GC-0001 and No. PR-12393, and bilateral project grant No. BI-US/24-26-081. This work is also funded by the European Union under Grant Agreement 101211695 (HE MSCA-PF AutoLLMSelect), Grant Agreement 101187010 (HE ERA Chair AutoLearn SI), Grant Agreement 101198470 (LLMs4EU), and Grant Agreement 101060712 (HE FishEUTrust). We also acknowledge the support of the EC/EuroHPC JU and the Slovenian Ministry of HESI via the project SLAIF (grant number 101254461).

Appendix A. Methods

This section describes the process of fine-tuning a pre-trained general-purpose LLM to successfully perform tasks from the food and nutrition domain. The section begins with the formulation of the tasks of interest in the ML domain, a description of the data preparation process for each, and evaluation settings. While some of the tasks were used for training purposes only, others were used for both LLM training and evaluation. The section then elaborates on the fine-tuning process of the LLM on the multi-task data. Finally, the experimental choices made during the fine-tuning are summarized.

A.1. Learning Tasks and Data

A.1.1. Assessing Ingredient Nutritional Profiles

In the task of assessing an ingredient's nutritional profile, the LLM was instructed to estimate the nutritional profile, i.e., amount of energy, fat, protein, salt, saturates, and sugars, for a specified household measure of a certain ingredient (e.g., one *cup* of *strawberries, raw*). The goal was to teach the LLM the calculation of nutritional values for different household measures (e.g., ml, gram, pound, tablespoon, cup, glass, drop, pinch, scoop) and quantities for the same ingredient. This task was used only during the LLM fine-tuning. The Recipe1M + dataset was used as a data source (Marın et al., 2021). All recipes were parsed to extract their ingredients, household measures, and nutritional profiles, after which they were analyzed all together to sample a balanced training dataset across ingredients and measurement units. First, all duplicate *ingredient – household measure – quantity* combinations were filtered out, and unique combinations retained. Second, all quantities were converted from string to float so they were comparable for sorting (e.g., the string “1 1/2" was converted to 1.5). The original string was preserved to be used in the prompt. The combinations containing quantities that could not be converted to float were skipped (e.g., ranges not specifying the exact quantity like “1 1/2-2", “1/2-1 1/4", “1/2 to 3/4"). Finally, the sampling was performed by the *ingredient - household measure* combination. For each *in gredient – household measure* combination, the quantities (converted to float) were sorted in ascending order. The smallest quantities by combination were selected (maximum 6 smallest quantities when there are more than 6). The original quantity string, not the one converted to a float, was then used in the prompt. To ensure variability of the instruction prompts, we generated 1000 candidate phrases with GPT-3.5 and randomly sampled one to use with each *ingredient – household measure – quantity* combination. The same was done with the opening phrases of the ground-truth responses accompanying each prompt. The sampling resulted in a training dataset of 9196 instances. The process ensured that the rare *ingredient – household measure* combi nations were part of the dataset, and the frequent *ingredient – household measure* combinations appeared only six times with their lowest quantities (e.g., some of the most frequent include “teaspoon salt, table", “cup wheat flour, white, all-purpose, unenriched", “cup water, bottled, generic"). Two training instances with the same *ingredient-household measure* combination (lemon juice, raw tablespoon), but with different quantities (3/4 in the first and 1 in the second) are provided below. The third and fourth instances refer to different household measures for the same ingredient.

> [INST] Find the nutritional values of 3/4 *tablespoon* lemon juice, raw [/INST]
>
> The nutrient values presented above are: energy - 2.53, fat - 0.03, protein - 0.04, salt - 0.09, saturates - 0.00, sugars - 0.29
>
> [INST] Analyze the nutritional content of 1 *tablespoon* lemon juice, raw [/INST]
>
> These nutrient values detailed are: energy - 3.38, fat - 0.04, protein - 0.05, salt - 0.12, saturates - 0.01, sugars - 0.38
>
> [INST] Ascertain the nutritional profile of 1 1/2 *quart* lemon juice, raw [/INST] These listed nutrient values are: energy - 324.00, fat - 3.54, protein - 5.10, salt - 12.00, saturates - 0.59, sugars - 36.90
>
> [INST] Evaluate the nutrient values of 1 *pound* lemon juice, raw [/INST]
>
> Nutrient values provided above are as follows: energy - 99.79, fat - 1.09, protein - 1.59, salt - 4.54, saturates - 0.18, sugars - 11.43

A.1.2. Assessing Recipe Nutritional Profiles

The next task involves assessing the nutritional profile of recipes given their list of ingredients (task 1) or a concatenation of the recipe title and a list of ingredients (task 2). The LLM was instructed to estimate the nutritional profile (energy, fat, protein, salt, saturates, and sugars) per 100 g of the recipe. These two tasks were used in both fine-tuning and evaluation, so both training and test datasets were sampled. In each of the two tasks, the process started by generating five random numbers to be used as seeds in the sampling of five different training-test datasets. The Recipe1M + dataset was used as a source of recipe titles, ingredient lists, and their nutritional profile per 100 g. The sampling of the recipes was again done by *ingredient – household measure* combination, to ensure a more balanced distribution of such combinations across all recipes in the dataset, as well as to make sure that rare combinations are sampled first. First, for each *ingredient – household measure* combination, all recipes containing that combination were stored in separate lists. The *ingredient – household mea sure* combinations were sorted in ascending order by the number of recipes they appeared in. Therefore, recipes with rare ingredients were sampled first. The count of sampled recipes for each *ingredient – household measure* combination was initialized to zero and increased by one after sampling a recipe containing it. For each *ingredient – household measure* combination, we first checked whether the number of recipes in the sample containing it had already reached the maximum allowed limit (n = 20). If not, a new recipe containing that combination was sampled. Then, for each *ingredient – household measure* combination, the list of corresponding recipes was randomly shuffled. Before adding a recipe to the sample, we checked (1) that it was not already included (based on its ID) and (2) that the Jaccard similarity between its ingredient–household measure list and that of any recipe already in the sample did not exceed 0.75. If it did not, the new recipe was added to the sample. We set the Jaccard similarity threshold to 0.75 as a conservative compromise that filters near-duplicate or highly overlapping ingredient lists while preserving recipes with minor formulation differences, thereby reducing data leakage without excessively shrinking the effective dataset or inducing unnecessary distribution shift between training and test splits. For each *ingredient – household measure* combination present in the sampled recipe, its counter was increased by one. When the counter of recipes containing the currently processed *ingredient – household measure* combination exceeded the allowed number n, the process ended for that *ingredient – household measure* combination. The sampling of the instructions in the prompt and the opening phrases in the ground-truth answer was done with GPT3.5, in the same manner as described in Section A.1.1. The same process was repeated for each of the five random seeds for each of the two tasks. For each random seed, we made sure (1) that the recipes from the training set did not appear in the test set and (2) that the Jaccard similarity of each sampled recipe in the test set to those already sampled in the training and test set did not exceed 0.75. In such a manner, we prevented data leakage. An example of the recipe from the two training datasets (one without and the other with the title in addition to the ingredient list) is given below. For evaluation, each nutrient estimation was mapped from $R \rightarrow \{0, 1\}$ based on the tolerance-based accuracy formulas. In total, we work with approximately 29,410 QA pairs for this task.

> [INST] Gauge the nutritional content per 100 g in a recipe made with these ingredients: 1 cup water, bottled, generic, 1 tablespoon oil, olive, salad or cooking, 34 teaspoon salt, table, 3 cup wheat flours, bread, unenriched, 2 teaspoon sugars, granulated, 1 teaspoon wheat flours, bread, unenriched, 2 teaspoon spices, parsley, dried, 1 teaspoon semolina, unenriched [/INST] Nutritional values for each 100 g: energy - 188.51, fat - 2.32, protein - 5.71, salt - 22.44, saturates - 0.33, sugars - 1.10

> [INST] Discover the nutritional content per 100 g in *Easy Pizza Dough for Bread Machine* with these ingredients: 1 cup water, bottled, generic, 1 tablespoon oil, olive, salad or cooking, 34 teaspoon salt, table, 3 cup wheat flours, bread, unenriched, 2 teaspoon sugars, granulated, 1 teaspoon wheat flours, bread, unenriched, 2 teaspoon spices, parsley, dried, 1 teaspoon semolina, unenriched [/INST] Each 100 g contains the following nutrient values: energy - 188.51, fat - 2.32, protein - 5.71, salt - 22.44, saturates - 0.33, sugars - 1.10

A.1.3. Classifying Recipes by Traffic Light Nutrition Labels

The next task involves assessing the traffic light nutrition labels of recipes given their list of ingredients (task 1) or a concatenation of the recipe title and a list of ingredients (task 2). The LLM was instructed to estimate one of three possible traffic light labels (red, orange, or green) per 100 g of the recipe, by nutrient (fat, salt, saturates, and sugars). These tasks were used in both LLM fine-tuning and evaluation, so for each, both training and test datasets were sampled. The process again started by generating five random numbers to be used as seeds in the sampling of five different training-test datasets. The Recipe1M + dataset was used as a source of recipe titles, ingredient lists, and their traffic light labels per 100 g. The sampling of the recipes was again done by *ingredient – household measure* combination, in the exact same manner as described in Section A.1.2. The goal was again to ensure a more balanced distribution of such combinations across all recipes in the dataset, to make sure that recipes with rare combinations are sampled first, and most importantly, to prevent data leakage between the training and test sets for each random seed. The same recipe as part of the datasets in tasks 1 and 2 is given as an example below. In total, we use approximately 29,410 QA pairs for this task.

> [INST] Verify the FSA traffic light labels per 100 g in a recipe that has these ingredients: 1 cup water, bottled, generic, 1 tablespoon oil, olive, salad or cooking, 34 teaspoon salt, table, 3 cup wheat flours, bread, unenriched, 2 teaspoon sugars, granulated, 1 teaspoon wheat flours, bread, unenriched, 2 teaspoon spices, parsley, dried, 1 teaspoon semolina, unenriched [/INST] Food Standards Agency traffic lights per 100 g presented: fat - green, salt - red, saturates - green, sugars - green

> [INST] Evaluate the FSA traffic light labels per 100 g in *Easy Pizza Dough for Bread Machine* using these ingredients: 1 cup water, bottled, generic, 1 tablespoon oil, olive, salad or cooking, 34 teaspoon salt, table, 3 cup wheat flours, bread, unenriched, 2 teaspoon sugars, granulated, 1 teaspoon wheat flours, bread, unenriched, 2 teaspoon spices, parsley, dried, 1 teaspoon semolina, unenriched [/INST] The fsa traffic lights per 100 g listed are: fat - green, salt - red, saturates - green, sugars - green

A.1.4. Food Named Entity Recognition and Linking

We used two already existing annotated food corpora: CafeteriaFCD and CafeteriaSA. CafeteriaFCD (Ispirova et al., 2022) is a manually curated corpus of food-consumption texts (recipes), annotated with entities from the Hansard taxonomy, FoodOn, and SNOMED-CT. Building on the 1000–recipe FoodBase gold standard, it was automatically annotated via the NCBO Annotator for FoodOn and SNOMED-CT and subsequently refined by domain experts. In total, the corpus comprises 7429 annotations. The length of the recipes ranges from 58 to 2024 characters (mean = 511.8), with each recipe containing at least one annotation. Annotation counts per recipe average 10.7 for Hansard (up to 37), 7.6 for FoodOn (up to 22), and 5.5 for SNOMED-CT (up to

17). CafeteriaSA (Cenikj et al., 2022b) consists of 500 scientific abstracts (SAs) annotated with food-entity mentions across the same three ontologies. It includes 6407 Hansard annotations, 4299 FoodOn annotations, and 3623 SNOMED-CT annotations. Abstract lengths span 236–2510 characters (mean ≈ 1467.6), and the annotation densities average 8.4 mentions per SA for Hansard (1–27 mentions), 4.8 for FoodOn (1–18 mentions), and 3.8 for SNOMED-CT (1–15 mentions).

The fine-tuning corpus was derived by converting the CafeteriaFCD and CafeteriaSA BioC XML corpora (Comeau et al., 2013) - a standard framework for interoperable biomedical data - into sequences of instruction–response (IR) pairs optimized for large language model adaptation. Each BioC in stance (a recipe instruction text or scientific abstract) containing text and food-entity annotations - specified by character offsets, lengths, mention text, and ontology uniform resource identifier (URIs) - was restructured into four IR pairs: one prompting the model to identify all food-entity mentions in the text (NER), and three prompting it to link those mentions to their corresponding entity URIs in the Hansard taxonomy, FoodOn, or SNOMED-CT ontologies (NEL). Every IR pair was wrapped in [INST] … [/INST] tags and immediately followed by the expected response, which consisted of a standard opening phrase plus either a list of detected mention spans (for NER) or a list of mention-URI mappings (for NEL). To prevent data leakage, no recipe or scientific abstract appeared more than once in the fine-tuning set. An example transformation of a CafeteriaFCD BioC instance into its four IR pairs is presented in Listing 1 from the Supplementary Information.

To introduce variability in the instruction prompts, we generated 1000 candidate instructions with GPT-3.5 and randomly assigned one to each BioC-formatted instance, which was then prepended to its recipe or scientific abstract text. Likewise, every model response started with a generic opening phrase - drawn at random from a separate 1000-phrase GPT-3.5 pool - followed by the list of food-entity mentions and their linked ontology URIs. After converting all instances into instruction–response pairs, we evaluated the entity coverage across ontologies using the CafeteriaFCD recipes (chosen for their greater diversity). Entities found to be underrepresented were supplemented with artificially generated IR pairs to ensure a more balanced distribution in the fine-tuning dataset.

Fig. 7 from the Supplementary Information shows the distribution of FoodOn entities in CafeteriaFCD; equivalent analyses for Hansard and SNOMED-CT are omitted for brevity. From these distributions, we established a minimum threshold of 150 mentions per entity[3] and augmented any entity with fewer than 150 occurrences ($k < 150$) by generating artificial NEL-only IR pairs until it reached the threshold. For each underrepresented entity, we sampled $150-k$ mentions. We then pooled all entity labels for each ontology, randomized the list, and randomly drew subsets of 7, 9, and 12 labels. Each set was used to create three IR pairs, instructing the LLM to link those labels to the target ontology. This process produced 13,492 artificial FoodOn NEL pairs, 1611 Hansard NEL pairs, and 4445 SNOMED-CT NEL pairs, ensuring that there are no duplicates to prevent data leakage during cross-validation. Combined with the 1479 original Cafeteria IR sequences (each containing one NER and three NEL instructions), the final fine-tuning dataset comprised 21,027 instances. This design guarantees that FoodyLLM encounters every URI from the three ontologies during training, enabling accurate linking of single- and multi-word mentions without hallucinations, and confines the model to retrieving only from these predefined ontologies.

A.1.5. *Food Synonyms Dataset*

We created a dataset of IR pairs in which 11,463 different ingredients from the Recipe1M + dataset were mapped to their corresponding entries in the USDA Food Composition Database (FCDB). Using this mapping, we generated question–answer tasks to identify synonyms for each ingredient. The mappings were sourced from a previous study that linked ingredients to the USDA FCDB via lexical, nutritional, and semantic similarity measures, and were subsequently validated by three domain experts to establish a gold standard. We used this dataset only in the training phase. An example of an IR pair is presented below:

> [INST] Can you please verify a match for the chosen food concept in the USDA food composition database 1% fat buttermilk?[/INST] As per the USDA database, the appropriate food match is the following food entity: milk, buttermilk, fluid, cultured, reduced fat.

A.1.6. *Household-measure Conversion Dataset*

We also created a dataset of IR pairs for household-measure conversions to help the LLM learn about conversions between different household measures. To do this, we used the conversion vocabulary from the Slovenian FCDB (Korošec et al., 2013) - compiled by food scientists - and generated one IR pair for each conversion. In total, we produced 65,955 IR pairs, which were used exclusively in the training set. An example of an IR pair is shown below:

> [INST] How many g are there in 1 cup, diced of cheese, mozzarella, low sodium? [/INST] The average weight for 1 cup, diced of cheese, mozzarella, low sodium is 132 g.

A.2. *Model Fine-Tuning*

For our experiments, we utilized the instruction-tuned 8B parameter version of Meta's Llama 3 model family (Dubey et al., 2024), specifically the meta-llama/Meta-Llama-3-8B-Instruct checkpoint available on HuggingFace. Llama 3 models are auto-regressive language models employing an optimized transformer architecture (Vaswani et al., 2017). The base 8B model was pre-trained on over 15 trillion tokens from publicly available sources, curated using advanced techniques like heuristic filtering and semantic deduplication, and has a knowledge cutoff of March 2023. The instruction-tuned variant we selected was further optimized for dialogue applications through an alignment process incorporating supervised fine-tuning (SFT), rejection sampling, proximal policy optimization (PPO), and direct preference optimization (DPO). This instruction tuning was performed on over 10 million manually labeled data points, focusing on enhancing reasoning, code generation, and similar capabilities.

Our choice of the Llama 3 8B Instruct model was driven by several factors. Firstly, its open-source nature allows for fine-tuning on specific tasks and domains, a flexibility not offered by proprietary models such as GPT-4, Gemini, or Claude. Secondly, due to limited GPU resources, larger open-source models (e.g., 70B or 400B parameters) were infeasible as they would significantly increase training time and costs. Finally, at the time of our

---

[3] This threshold will be refined in future sensitivity analyses.

experiments, the Llama 3 8B model represented one of the most advanced open-weight large language models available, offering competitive performance on standard benchmarks such as ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), and TruthfulQA (Lin et al., 2022). We also tested Mistral 7B Instruct v0.1, which - at the time - performed comparably to LLaMA 3 in most benchmarks. After running initial experiments with both models, we ultimately chose LLaMA 3 for its superior performance.

To enhance memory efficiency and accelerate the fine-tuning process, we loaded the model in 4-bit precision and employed Low-Rank Adaptation (LoRA) (Hu et al., 2021). This involved freezing all layers of the backbone model and exclusively training low-rank adaptations for the query and value weight matrices. We fine-tuned the model using a supervised fine-tuning (SFT) approach with a next-word prediction objective.

The fine-tuning hyperparameters were set as follows: a learning rate of 2e 4, LoRA intrinsic rank (r) of 16, LoRA alpha of 16, and LoRA dropout of 0.05. The model was trained for one epoch with a batch size of 10 and 10 warm-up steps. Input sequences were processed with a maximum sequence length of 1024 tokens; sequences were padded to the length of the longest sequence in their batch and truncated if they exceeded this maximum. Since the Llama tokenizer lacks a default padding token, we assigned one using a designated reserved special token, specifically "<|reserved_special_token_250|>". This was necessary as initial experiments revealed that using the end-of-sequence (EOS) token for padding - a common alternative - resulted in the model generating repetitions and struggling to properly conclude sequences.

The code and datasets for fine-tuning are available on GitHub.[4] Model weights are available in Hugging Face's model repository.[5]

Table 5
Summary of the data used in the experiments

Task/Dataset	Number of QA pairs	Train/Test
Accessing ingredient nutrition profile	9196	Train
Accessing recipe nutritional profile	29410	Train (19,524)/Test (9,886)
Classifying recipes by traffic light nutrition labels	29410	Train (19,524)/Test (9,886)
Food named-entity recognition and linking	21027	Train (16,822)/Test (4,205)
Food synonyms	11463	Train
Household-measure conversion	65955	Train

Appendix B. Tolerance Levels in Nutrition Content in Foods Besides Food Supplements

To estimate the errors between the true values and the results provided by FoodyLLM, we have applied defined tolerances under international regulations. The European Commission's 2012 guidance, aligned with Regulation EU 1169/2011 (Bairati, 2017), outlines acceptable deviations between declared and actual nutrient values due to natural, production, and storage variations. These tolerances prevent misleading consumers. For our case, we applied the tolerances for standard food products (excluding supplements), as detailed in Tables 6 and 7 These thresholds are used in the final step of our methodology to assess the accuracy of the provided FoodyLLM macronutrient values.

Table 6
Tolerated differences in nutrition content in foods besides food supplements.

Quantity per 100 g/Macronutrient	Protein/Sugar	Fat
<10 g per 100 g	±2 g	±1.5 g
10–40 g per 100 g	±20%	±20%
>40 g per 100 g	±8 g	±8 g

Table 7
Tolerated differences in nutrition content in foods besides food supplements.

Quantity per 100 g/Macronutrient	Saturates
<4 g per 100 g	±0.8 g
≥ 4 g per 100 g	±20%

Appendix C. Assessing Recipe Nutritional Profiles

---

[4] https://github.com/matejMartinc/FoodyLLM.

[5] https://huggingface.co/Matej/FoodyLLM.

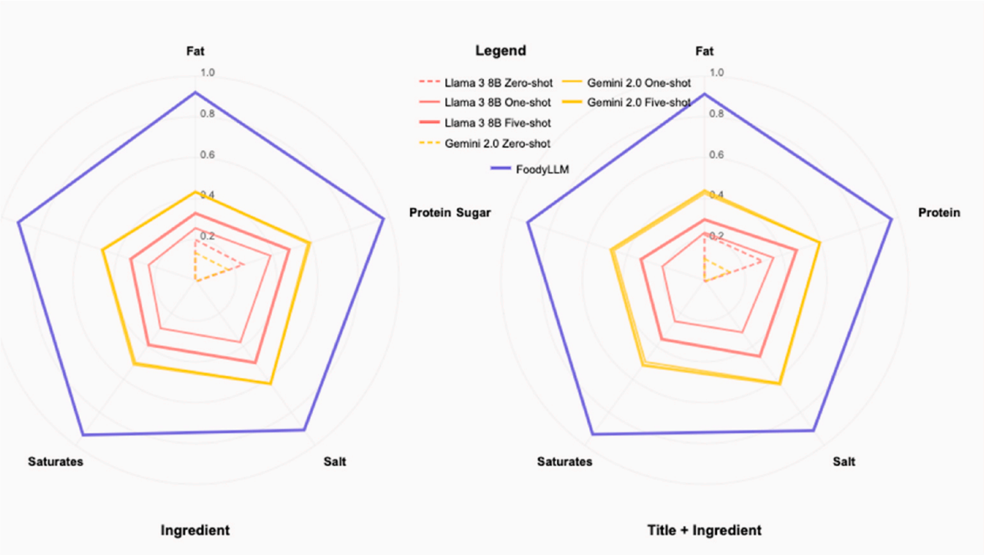


Fig. 2. Nutrient Accuracy Comparison of Zero-, One-, and Five-Shot Baselines versus FoodyLLM in the Ingredient and Title + Ingredient dataset.

Appendix D. Classifying Recipes by Traffic Light Nutrition Labels

Tables 8 and 9 show the food traffic light system results for the Title + Ingredient dataset.

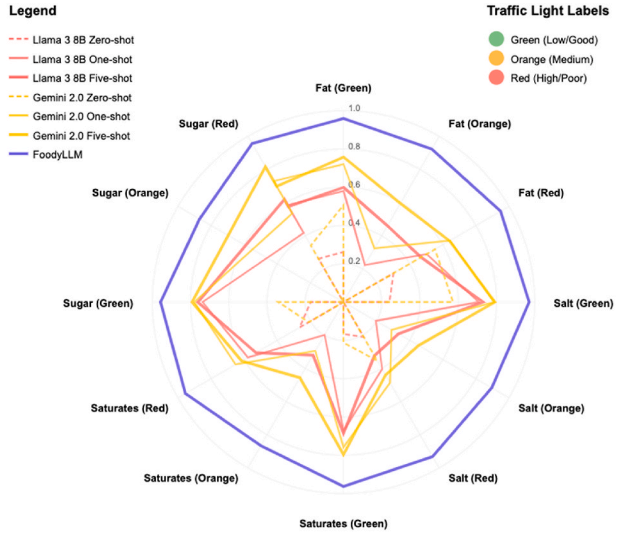


Fig. 3. FSA Macro Average F1 Score Comparison of Zero-, One-, and Five-Shot Base lines versus FoodyLLM in the Ingredient dataset.

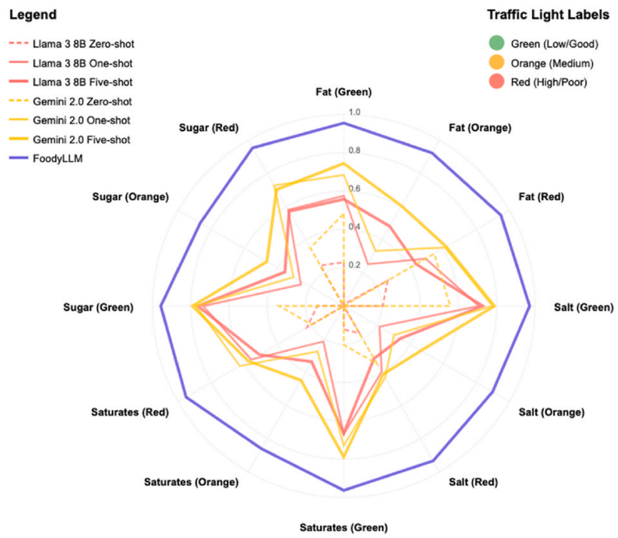


Fig. 4. FSA Macro Average F1 Score Comparison of Zero-, One-, and Five-Shot Base lines versus FoodyLLM in Title + Ingredient dataset.

Table 8
Results by class for the Title + Ingredient dataset (1/2)

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Fat (Green)	Llama 3 8B	0-shot prompt.	0.584 ± 0.042	0.142 ± 0.007	0.229 ± 0.012
	Llama 3 8B	1-shot prompt.	0.573 ± 0.008	0.580 ± 0.013	0.576 ± 0.008
	Llama 3 8B	5-shot prompt.	0.672 ± 0.009	0.478 ± 0.011	0.558 ± 0.010
	Gemini 2.0	0-shot prompt.	0.671 ± 0.017	0.376 ± 0.011	0.482 ± 0.014
	Gemini 2.0	1-shot prompt.	0.835 ± 0.012	0.579 ± 0.009	0.684 ± 0.010
	Gemini 2.0	5-shot prompt.	0.821 ± 0.009	0.682 ± 0.013	0.745 ± 0.010
	FoodyLLM		0.964 ± 0.007	0.948 ± 0.008	0.956 ± 0.003
Fat (Orange)	Llama 3 8B	0-shot prompt.	0.200 ± 0.274	0.000 ± 0.001	0.001 ± 0.001
	Llama 3 8B	1-shot prompt.	0.442 ± 0.011	0.177 ±0.009	0.253 ± 0.010
	Llama 3 8B	5-shot prompt.	0.432 ± 0.019	0.543 ± 0.014	0.481 ± 0.017
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.581 ± 0.024	0.233 ± 0.009	0.332 ± 0.012
	Gemini 2.0	5-shot prompt.	0.548 ± 0.012	0.672 ± 0.005	0.604 ± 0.008
	FoodyLLM		0.918 ± 0.007	0.928 ± 0.007	0.923 ± 0.007
Fat (Red)	Llama 3 8B	0-shot prompt.	0.354 ± 0.010	0.213 ± 0.014	0.265 ± 0.012
	Llama 3 8B	1-shot prompt.	0.413 ± 0.008	0.614 ± 0.012	0.494 ± 0.009
	Llama 3 8B	5-shot prompt.	0.437 ± 0.012	0.438 ± 0.016	0.437 ± 0.014
	Gemini 2.0	0-shot prompt.	0.615 ± 0.008	0.492 ± 0.007	0.547 ± 0.005
	Gemini 2.0	1-shot prompt.	0.633 ± 0.013	0.590 ± 0.010	0.611 ± 0.009
	Gemini 2.0	5-shot prompt.	0.662 ± 0.010	0.582 ± 0.011	0.619 ± 0.009
	FoodyLLM		0.947 ± 0.011	0.952 ± 0.013	0.949 ± 0.005
Salt (Green)	Llama 3 8B	0-shot prompt.	0.640 ± 0.023	0.121 ± 0.003	0.204 ± 0.003
	Llama 3 8B	1-shot prompt.	0.617 ± 0.010	0.828 ± 0.009	0.707 ± 0.010
	Llama 3 8B	5-shot prompt.	0.630 ± 0.007	0.859 ± 0.005	0.727 ± 0.004
	Gemini 2.0	0-shot prompt.	0.774 ± 0.007	0.435 ± 0.013	0.557 ± 0.011
	Gemini 2.0	1-shot prompt.	0.761 ± 0.003	0.796 ± 0.007	0.778 ± 0.003
	Gemini 2.0	5-shot prompt.	0.734 ± 0.003	0.850 ± 0.009	0.788 ± 0.004
	FoodyLLM		0.976 ± 0.006	0.968 ± 0.004	0.972 ± 0.002
Salt (Orange)	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.396 ± 0.021	0.149 ± 0.011	0.216 ± 0.014
	Llama 3 8B	5-shot prompt.	0.390 ± 0.016	0.303 ± 0.013	0.341 ± 0.014
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.459 ± 0.022	0.226 ± 0.012	0.303 ± 0.015
	Gemini 2.0	5-shot prompt.	0.457 ± 0.011	0.475 ± 0.012	0.466 ± 0.011
	FoodyLLM		0.893 ± 0.003	0.903 ± 0.007	0.898 ± 0.003
Salt (Red)	Llama 3 8B	0-shot prompt.	0.384 ± 0.020	0.104 ± 0.010	0.164 ± 0.014
	Llama 3 8B	1-shot prompt.	0.478 ± 0.016	0.340 ± 0.011	0.397 ± 0.013
	Llama 3 8B	5-shot prompt.	0.491 ± 0.015	0.234 ± 0.015	0.317 ± 0.016
	Gemini 2.0	0-shot prompt.	0.641 ± 0.015	0.239 ± 0.008	0.348 ± 0.009
	Gemini 2.0	1-shot prompt.	0.524 ± 0.017	0.369 ± 0.004	0.433 ± 0.005
	Gemini 2.0	5-shot prompt.	0.545 ± 0.009	0.331 ± 0.012	0.412 ± 0.011
	FoodyLLM		0.934 ± 0.011	0.936 ± 0.012	0.935 ± 0.003

Table 9
Results by class for the Title + Ingredient dataset (2/2)

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Saturates (Green)	Llama 3 8B	0-shot prompt.	0.704 ± 0.037	0.067 ± 0.006	0.122 ± 0.010
	Llama 3 8B	1-shot prompt.	0.692 ± 0.004	0.647 ± 0.010	0.669 ± 0.006
	Llama 3 8B	5-shot prompt.	0.713 ± 0.011	0.623 ± 0.007	0.665 ± 0.008
	Gemini 2.0	0-shot prompt.	0.857 ± 0.031	0.115 ± 0.008	0.203 ± 0.012
	Gemini 2.0	1-shot prompt.	0.871 ± 0.003	0.627 ± 0.007	0.729 ± 0.004
	Gemini 2.0	5-shot prompt.	0.820 ± 0.006	0.759 ± 0.008	0.788 ± 0.007
	FoodyLLM		0.963 ± 0.006	0.963 ± 0.007	0.963 ± 0.002
Saturates (Orange)	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.297 ± 0.014	0.168 ± 0.008	0.214 ± 0.009
	Llama 3 8B	5-shot prompt.	0.287 ± 0.015	0.406 ± 0.015	0.336 ± 0.015
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.384 ± 0.027	0.213 ± 0.014	0.274 ± 0.018
	Gemini 2.0	5-shot prompt.	0.360 ± 0.012	0.593 ± 0.011	0.448 ± 0.012
	FoodyLLM		0.862 ± 0.008	0.860 ± 0.011	0.861 ± 0.003
Saturates (Red)	Llama 3 8B	0-shot prompt.	0.431 ± 0.020	0.148 ± 0.014	0.221 ± 0.018
	Llama 3 8B	1-shot prompt.	0.523 ± 0.009	0.601 ± 0.013	0.559 ± 0.010
	Llama 3 8B	5-shot prompt.	0.548 ± 0.012	0.476 ± 0.007	0.509 ± 0.008
	Gemini 2.0	0-shot prompt.	0.728 ± 0.034	0.112 ± 0.008	0.194 ± 0.013
	Gemini 2.0	1-shot prompt.	0.762 ± 0.020	0.532 ± 0.009	0.627 ± 0.013
	Gemini 2.0	5-shot prompt.	0.768 ± 0.009	0.461 ± 0.007	0.576 ± 0.005
	FoodyLLM		0.951 ± 0.006	0.952 ± 0.006	0.952 ± 0.002
Sugar (Green)	Llama 3 8B	0-shot prompt.	0.640 ± 0.034	0.077 ± 0.010	0.139 ± 0.017
	Llama 3 8B	1-shot prompt.	0.754 ± 0.006	0.717 ± 0.013	0.735 ± 0.007
	Llama 3 8B	5-shot prompt.	0.794 ± 0.004	0.734 ± 0.005	0.763 ± 0.004
	Gemini 2.0	0-shot prompt.	0.849 ± 0.009	0.210 ± 0.002	0.337 ± 0.003
	Gemini 2.0	1-shot prompt.	0.873 ± 0.003	0.693 ± 0.010	0.773 ± 0.006
	Gemini 2.0	5-shot prompt.	0.874 ± 0.004	0.723 ± 0.006	0.791 ± 0.003
	FoodyLLM		0.957 ± 0.006	0.957 ± 0.005	0.957 ± 0.002
Sugar (Orange)	Llama 3 8B	0-shot prompt.	0.567 ± 0.365	0.001 ± 0.001	0.002 ± 0.001
	Llama 3 8B	1-shot prompt.	0.329 ± 0.003	0.213 ± 0.013	0.259 ± 0.011
	Llama 3 8B	5-shot prompt.	0.330 ± 0.014	0.384 ± 0.012	0.355 ± 0.013
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.431 ± 0.020	0.233 ± 0.007	0.303 ± 0.010
	Gemini 2.0	5-shot prompt.	0.415 ± 0.007	0.525 ± 0.014	0.464 ± 0.006
	FoodyLLM		0.864 ± 0.008	0.870 ± 0.011	0.867 ± 0.004
Sugar (Red)	Llama 3 8B	0-shot prompt.	0.459 ± 0.026	0.167 ± 0.007	0.245 ± 0.011
	Llama 3 8B	1-shot prompt.	0.553 ± 0.007	0.613 ± 0.013	0.581 ± 0.008
	Llama 3 8B	5-shot prompt.	0.585 ± 0.008	0.557 ± 0.012	0.571 ± 0.009
	Gemini 2.0	0-shot prompt.	0.737 ± 0.016	0.229 ± 0.013	0.350 ± 0.017
	Gemini 2.0	1-shot prompt.	0.655 ± 0.011	0.820 ± 0.012	0.728 ± 0.008
	Gemini 2.0	5-shot prompt.	0.690 ± 0.013	0.716 ± 0.014	0.703 ± 0.008
	FoodyLLM		0.958 ± 0.008	0.951 ± 0.003	0.954 ± 0.003

Table 10
Average five-fold cross-validation NEL results (standard deviation given in brack ets).

Test Set	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
FoodOn	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.199 ± 0.007	0.035 ± 0.002	0.057 ± 0.002
	Llama 3 8B	5-shot prompt.	0.362 ± 0.007	0.059 ± 0.004	0.099 ± 0.006
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.580 ± 0.020	0.059 ± 0.003	0.102 ± 0.005
	Gemini 2.0	5-shot prompt.	0.680 ± 0.006	0.231 ± 0.007	0.330 ± 0.008
	FoodyLLM		0.951 ± 0.001	0.936 ± 0.001	0.942 ± 0.001
Snomed	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.209 ± 0.015	0.052 ± 0.005	0.080 ± 0.007
	Llama 3 8B	5-shot prompt.	0.416 ± 0.012	0.127 ± 0.004	0.188 ± 0.006
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.558 ± 0.010	0.045 ± 0.002	0.082 ± 0.004
	Gemini 2.0	5-shot prompt.	0.759 ± 0.010	0.224 ± 0.004	0.330 ± 0.005
	FoodyLLM		0.980 ± 0.002	0.970 ± 0.002	0.975 ± 0.001
Hansard	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.159 ± 0.014	0.107 ± 0.007	0.121 ± 0.009
	Llama 3 8B	5-shot prompt.	0.404 ± 0.016	0.071 ± 0.003	0.117 ± 0.005
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.309 ± 0.020	0.121 ± 0.008	0.162 ± 0.011
	Gemini 2.0	5-shot prompt.	0.545 ± 0.013	0.389 ± 0.010	0.438 ± 0.011
	FoodyLLM		0.943 ± 0.005	0.926 ± 0.007	0.932 ± 0.004

Appendix E. Making food data interoperable

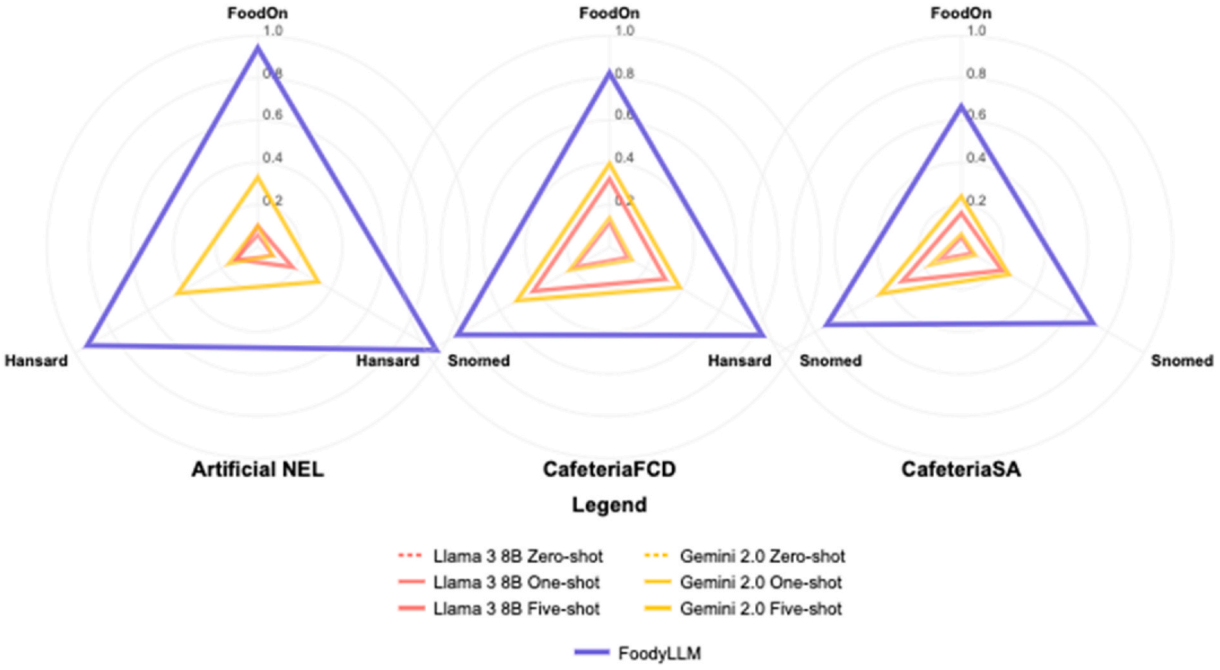


Fig. 5. Macro-Averaged F1 Score Comparison of Zero-, One-, and Five-Shot Baselines versus FoodyLLM on NEL evaluation in artificially generated NEL samples, CafeteriaFCD, and CafeteriaSA.

Table 11
Average five-fold cross-validation NEL results on CafeteriaFCD dataset (standard deviation given in brackets). The NEL task is performed after an NER task.

Test Set	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
FoodOn	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.329 ± 0.022	0.072 ± 0.009	0.113 ± 0.014
	Llama 3 8B	5-shot prompt.	0.525 ± 0.016	0.249 ± 0.004	0.322 ± 0.003
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.534 ± 0.031	0.086 ± 0.013	0.140 ± 0.018
	Gemini 2.0	5-shot prompt.	0.690 ± 0.025	0.298 ± 0.010	0.394 ± 0.012
	FoodyLLM		0.923 ± 0.006	0.764 ± 0.012	0.823 ± 0.007
Snomed	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.276 ± 0.013	0.061 ± 0.003	0.096 ± 0.004
	Llama 3 8B	5-shot prompt.	0.484 ± 0.018	0.235 ± 0.011	0.304 ± 0.011
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.507 ± 0.038	0.073 ± 0.003	0.124 ± 0.005
	Gemini 2.0	5-shot prompt.	0.701 ± 0.036	0.283 ± 0.015	0.384 ± 0.019
	FoodyLLM		0.946 ± 0.009	0.767 ± 0.021	0.835 ± 0.017
Hansard	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.314 ± 0.014	0.142 ± 0.016	0.187 ± 0.017
	Llama 3 8B	5-shot prompt.	0.522 ± 0.021	0.369 ± 0.014	0.417 ± 0.016
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.410 ± 0.025	0.164 ± 0.011	0.221 ± 0.014
	Gemini 2.0	5-shot prompt.	0.623 ± 0.017	0.447 ± 0.016	0.505 ± 0.016
	FoodyLLM		0.907 ± 0.012	0.774 ± 0.021	0.830 ± 0.018

Table 12
Average five-fold cross-validation NEL results on CafeteriaSA dataset (standard deviation given in brackets). The NEL task is performed after an NER task.

Test Set	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
FoodOn	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.192 ± 0.026	0.028 ± 0.007	0.045 ± 0.011
	Llama 3 8B	5-shot prompt.	0.347 ± 0.025	0.114 ± 0.015	0.161 ± 0.019
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.225 ± 0.076	0.039 ± 0.012	0.063 ± 0.020
	Gemini 2.0	5-shot prompt.	0.493 ± 0.040	0.171 ± 0.008	0.240 ± 0.009
	FoodyLLM		0.804 ± 0.020	0.597 ± 0.039	0.665 ± 0.029
Snomed	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000

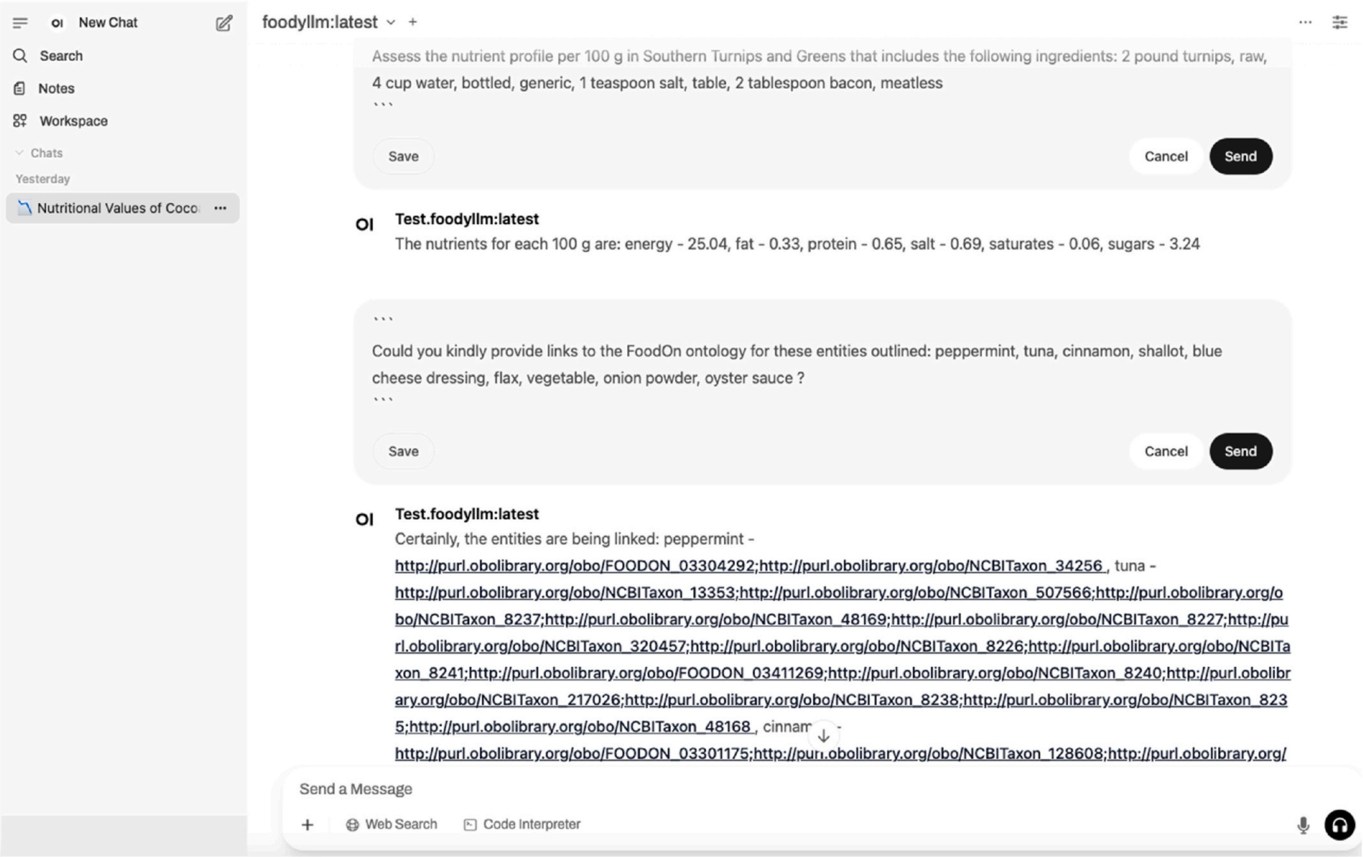
(*continued on next page*)

Table 12 (*continued*)

Test Set	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
	Llama 3 8B	1-shot prompt.	0.212 ± 0.023	0.034 ± 0.006	0.056 ± 0.010
	Llama 3 8B	5-shot prompt.	0.392 ± 0.041	0.167 ± 0.020	0.222 ± 0.026
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.305 ± 0.044	0.046 ± 0.008	0.078 ± 0.012
	Gemini 2.0	5-shot prompt.	0.526 ± 0.017	0.185 ± 0.024	0.260 ± 0.027
	FoodyLLM		0.851 ± 0.023	0.646 ± 0.012	0.720 ± 0.013
Hansard	Llama 3 8B	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Llama 3 8B	1-shot prompt.	0.269 ± 0.063	0.077 ± 0.010	0.111 ± 0.016
	Llama 3 8B	5-shot prompt.	0.474 ± 0.058	0.269 ± 0.025	0.321 ±0.028
	Gemini 2.0	0-shot prompt.	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Gemini 2.0	1-shot prompt.	0.401 ± 0.048	0.121 ± 0.011	0.172 ± 0.016
	Gemini 2.0	5-shot prompt.	0.623 ± 0.060	0.380 ± 0.029	0.441 ± 0.036
	FoodyLLM		0.873 ± 0.018	0.656 ± 0.026	0.735 ± 0.023

Appendix F. FoodyLLM User Interface

An example of a conversation of a user with the FoodyLLM through the Ollama interface is presented in Fig. 6.



Assess the nutrient profile per 100 g in Southern Turnips and Greens that includes the following ingredients: 2 pound turnips, raw, 4 cup water, bottled, generic, 1 teaspoon salt, table, 2 tablespoon bacon, meatless
```

**Test.foodyllm:latest**
The nutrients for each 100 g are: energy - 25.04, fat - 0.33, protein - 0.65, salt - 0.69, saturates - 0.06, sugars - 3.24

```
Could you kindly provide links to the FoodOn ontology for these entities outlined: peppermint, tuna, cinnamon, shallot, blue cheese dressing, flax, vegetable, onion powder, oyster sauce ?
```

Test.foodyllm:latest
Certainly, the entities are being linked: peppermint - http://purl.obolibrary.org/obo/FOODON_03304292;http://purl.obolibrary.org/obo/NCBITaxon_34256 , tuna - http://purl.obolibrary.org/obo/NCBITaxon_13353;http://purl.obolibrary.org/obo/NCBITaxon_507566;http://purl.obolibrary.org/obo/NCBITaxon_8237;http://purl.obolibrary.org/obo/NCBITaxon_48169;http://purl.obolibrary.org/obo/NCBITaxon_8227;http://purl.obolibrary.org/obo/NCBITaxon_320457;http://purl.obolibrary.org/obo/NCBITaxon_8226;http://purl.obolibrary.org/obo/NCBITaxon_8241;http://purl.obolibrary.org/obo/FOODON_03411269;http://purl.obolibrary.org/obo/NCBITaxon_8240;http://purl.obolibrary.org/obo/NCBITaxon_217026;http://purl.obolibrary.org/obo/NCBITaxon_8238;http://purl.obolibrary.org/obo/NCBITaxon_8235;http://purl.obolibrary.org/obo/NCBITaxon_48168 , cinnam -
http://purl.obolibrary.org/obo/FOODON_03301175;http://purl.obolibrary.org/obo/NCBITaxon_128608;http://purl.obolibrary.org/

Fig. 6. An Ollama interface showing a conversation of a user with the FoodyLLM model.

Appendix G. An example of conversion between a BioC instance and an IR pair
```

```
<document >
<id >0 recipe1006 </id >
<infon key =" full_text ">
Mix the cream cheese , beef , olives , onion , and Worcestershire
sauce together in a bowl until evenly blended . Keeping the
mixture in the bowl , scrape it into a semi - ball shape . Cover ,
and refrigerate until firm , at least 2 hours . Place a large
sheet of waxed paper on a flat surface . Sprinkle with walnuts .
Roll the cheese ball in the walnuts until completely covered .
Transfer the cheese ball to a serving plate , or rewrap with waxed
paper and refrigerate until needed .
</ infon >
<annotation id="1">
<infon key =" semantic_tags ">
http: // purl . obolibrary . org / obo / FOODON_03301889 ;
http: // purl . obolibrary . org / obo / FOODON_00001013
</ infon >
<location offset ="3" length ="12"/ >
<text > CREAM CHEESE </ text >
</ annotation >
<annotation id="2">
<infon key =" semantic_tags ">
http: // purl . obolibrary . org / obo / FOODON_03301704 ;
http: // purl . obolibrary . org / obo / NCBITaxon_4679
</ infon >
<location offset ="10" length ="5"/ >
<text > ONION </ text >
</ annotation >
<annotation id="3">
<infon key =" semantic_tags ">
http: // purl . obolibrary . org / obo / FOODON_03305003 ;
http: // purl . obolibrary . org / obo / FOODON_03311146
</ infon >
<location offset ="13" length ="20"/ >
<text > WORCESTERSHIRE SAUCE </ text >
</ annotation >
<annotation id="4">
<infon key =" semantic_tags ">
http: // purl . obolibrary . org / obo / NCBITaxon_16718
</ infon >
<location offset ="63" length ="7"/ >
<text > WALNUTS </ text >
</ annotation >
<! -- ... additional annotations ... -->
<infon key =" category "> Appetizers and snacks </ infon >
</ document >
```

Listing 1: BioC-formatted recipe from CafeteriaFCD mapped to FoodOn ontology.

[INST] Find all food entities present in the text: Mix the cream cheese, beef, olives, onion, and Worcestershire sauce together in a bowl until evenly blended. Keeping the mixture in the bowl, scrape it into a semi-ball shape. Cover and refrigerate until firm, at least 2 hours. Place a large sheet of waxed paper on a flat surface. Sprinkle with walnuts. Roll the cheese ball in the walnuts until completely covered. Transfer the cheese ball to a serving plate, or rewrap with waxed paper and refrigerate until needed. [/INST] Certainly, the entities connected with food are outlined as follows: cream cheese, beef, olives, onion, Worcester shire sauce, walnuts, cheese ball.

[INST] Is it within your capabilities to link the extracted food entities to a Hansard taxonomy? [/INST] Certainly, the entities are associated properly: cream cheese - AG.01.e [Dairy produce]; AG.01.e.02 [Cheese]; AG.01.n [Dishes and prepared food]; AG.01.n.18 [Preserve], beef - AG.01.d.03 [Beef], olives - AG.01.h.01.e [Fruit containing stone], onion - AG.01.h.02.e [Onion/leek/garlic], Worcestershire sauce - AG.01.h [Fruit and vegeta bles]; AG.01.l.04 [Sauce/dressing], walnuts - AG.01.h.01.f [Nut], cheese ball - AG.01.e.02 [Cheese]; AG.01.n.18 [Preserve].

[INST] Might I trouble you to connect the extracted food entities to a FoodOn ontology, if possible? [/INST] Definitely, the entities are linked suitably: cream cheese - FOODON-03301889; FOODON-00001013, onion - FOODON-03301704; NCBITaxon-4679, worcestershire sauce - FOODON-03305003; FOODON-03311146, walnuts - NCBITaxon-16718, cheese - FOODON-00001013.

[INST] Might it be possible for you to link them to an SNOMEDCT ontology, if needed? [/INST] Absolutely, the entities are related as such: cream cheese - SNOMEDCT-226849005; SNOMEDCT 255621006; SNOMEDCT-102264005, beef - SNOMEDCT-226916002, olives - SNOMEDCT-227436000, onion - SNOMEDCT-735047000, worcestershire sauce - SNOMEDCT-443701000124100; SNOMEDCT-227519005, cheese - SNOMEDCT-102264005.

Appendix H. Distribution of FoodOn Semantic Tags Across CafeteriaFCD

Fig. 7 shows the distribution of FoodOn entities in CafeteriaFCD;

equivalent analyses for Hansard and SNOMED-CT are omitted for brevity.

Appendix I. Baseline FSA Synonym Evaluation

To ensure a fair baseline evaluation, we evaluated the impact of considering synonyms of the FSA traffic light labels that appear in the baseline results. The considered synonyms include ‘low’ for ‘green’, ‘high’ for ‘red’, and ‘amber’, ‘yellow’, ‘medium’, and ‘moderate’ for ‘orange’. We did not consider these synonyms for the fine-tuned LLM because we expect it to output only the ‘orange’ label. The baseline results considering synonyms are given in Tables 13–16. The results indicate that prior to its fine-tuning, the LLM outputs several different labels for the ’orange’ FSA traffic light.

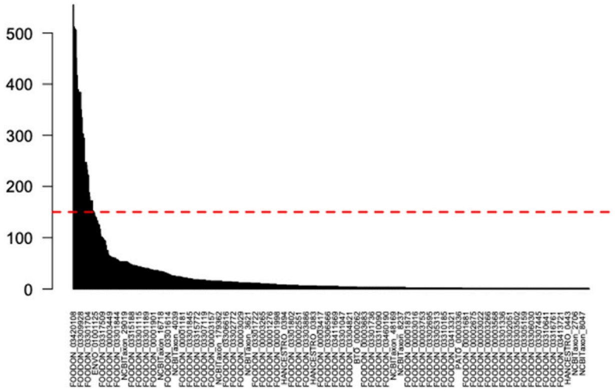


Fig. 7. Distribution of FoodON entities in CafeteriaFCD. Only a subset of entity names is displayed on the x-axis. The red line represents the threshold of 150.

Table 13
Results by class for the Ingredient dataset (1/2)

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Fat (Green)	Llama 3 8B	0-shot prompt.	0.559 ± 0.022	0.168 ± 0.009	0.258 ± 0.013
	Llama 3 8B	1-shot prompt.	0.545 ± 0.010	0.621 ± 0.009	0.581 ± 0.007
	Llama 3 8B	5-shot prompt.	0.643 ± 0.008	0.562 ± 0.012	0.599 ± 0.007
	Gemini 2.0	0-shot prompt.	0.662 ± 0.007	0.409 ± 0.032	0.505 ± 0.027
	Gemini 2.0	1-shot prompt.	0.826 ± 0.012	0.639 ± 0.016	0.720 ± 0.014
	Gemini 2.0	5-shot prompt.	0.811 ± 0.003	0.710 ± 0.012	0.757 ± 0.007
Fat (Orange)	Llama 3 8B	0-shot prompt.	0.424 ± 0.031	0.068 ± 0.008	0.118 ± 0.012
	Llama 3 8B	1-shot prompt.	0.445 ± 0.022	0.193 ± 0.011	0.269 ± 0.014
	Llama 3 8B	5-shot prompt.	0.440 ± 0.006	0.483 ± 0.015	0.460 ± 0.009
	Gemini 2.0	0-shot prompt.	0.597 ± 0.015	0.281 ± 0.015	0.382 ± 0.016
	Gemini 2.0	1-shot prompt.	0.536 ± 0.009	0.636 ± 0.012	0.582 ± 0.009
	Gemini 2.0	5-shot prompt.	0.555 ± 0.007	0.647 ± 0.017	0.597 ± 0.010
Fat (Red)	Llama 3 8B	0-shot prompt.	0.372 ± 0.010	0.260 ± 0.007	0.306 ± 0.008
	Llama 3 8B	1-shot prompt.	0.429 ± 0.005	0.607 ± 0.015	0.503 ± 0.008
	Llama 3 8B	5-shot prompt.	0.463 ± 0.014	0.473 ± 0.006	0.468 ± 0.008
	Gemini 2.0	0-shot prompt.	0.629 ± 0.011	0.431 ± 0.031	0.511 ± 0.024
	Gemini 2.0	1-shot prompt.	0.634 ± 0.007	0.652 ± 0.010	0.643 ± 0.006
	Gemini 2.0	5-shot prompt.	0.663 ± 0.013	0.618 ± 0.011	0.640 ± 0.009
Salt (Green)	Llama 3 8B	0-shot prompt.	0.629 ± 0.008	0.148 ± 0.004	0.240 ± 0.004
	Llama 3 8B	1-shot prompt.	0.614 ± 0.003	0.834 ± 0.006	0.707 ± 0.004
	Llama 3 8B	5-shot prompt.	0.631 ± 0.010	0.865 ± 0.004	0.730 ± 0.008
	Gemini 2.0	0-shot prompt.	0.769 ± 0.010	0.464 ± 0.032	0.578 ± 0.028
	Gemini 2.0	1-shot prompt.	0.772 ± 0.008	0.808 ± 0.003	0.789 ± 0.004
	Gemini 2.0	5-shot prompt.	0.748 ± 0.009	0.838 ± 0.007	0.791 ± 0.007
Salt (Orange)	Llama 3 8B	0-shot prompt.	0.375 ± 0.041	0.051 ± 0.008	0.089 ± 0.014
	Llama 3 8B	1-shot prompt.	0.395 ± 0.021	0.146 ± 0.007	0.213 ± 0.010
	Llama 3 8B	5-shot prompt.	0.391 ± 0.006	0.289 ± 0.012	0.332 ± 0.010
	Gemini 2.0	0-shot prompt.	0.528 ± 0.016	0.329 ± 0.023	0.405 ± 0.022
	Gemini 2.0	1-shot prompt.	0.467 ± 0.011	0.434 ± 0.009	0.450 ± 0.009
	Gemini 2.0	5-shot prompt.	0.459 ± 0.012	0.448 ± 0.007	0.453 ± 0.009
Salt (Red)	Llama 3 8B	0-shot prompt.	0.363 ± 0.011	0.148 ± 0.011	0.210 ± 0.011
	Llama 3 8B	1-shot prompt.	0.475 ± 0.008	0.352 ± 0.013	0.404 ± 0.011
	Llama 3 8B	5-shot prompt.	0.481 ± 0.021	0.244 ± 0.017	0.323 ± 0.019
	Gemini 2.0	0-shot prompt.	0.620 ± 0.021	0.249 ± 0.020	0.355 ± 0.023
	Gemini 2.0	1-shot prompt.	0.489 ± 0.006	0.483 ± 0.022	0.486 ± 0.014
	Gemini 2.0	5-shot prompt.	0.507 ± 0.016	0.389 ± 0.021	0.441 ± 0.019

Appendix J. Computational Resources and Budget

All models were trained on one NVIDIA A100 GPU, and the total computational budget for all training and testing experiments was roughly 500 GPU hours. We minimized the training costs by quantizing the model to 4- bits and by fine-tuning it by using the LoRA procedure, during which most model weights are frozen. The inference time of the model depends on the hardware and the inference setting (e.g., are the input queries grouped into batches, maximum context window, padding). In our setting, we used a batch of 16 for inference, and the model was easily capable of generating roughly 80 tokens per second per query on one NVIDIA A100 GPU. The inference time could be further optimized as our work is still in the research stage.

Table 14
Results by class for the Ingredient dataset (2/2)

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Saturates (Green)	Llama 3 8B	0-shot prompt.	0.635 ± 0.048	0.093 ± 0.007	0.161 ± 0.013
	Llama 3 8B	1-shot prompt.	0.677 ± 0.006	0.669 ± 0.010	0.673 ± 0.005
	Llama 3 8B	5-shot prompt.	0.718 ± 0.010	0.656 ± 0.012	0.685 ± 0.005
	Gemini 2.0	0-shot prompt.	0.824 ± 0.027	0.123 ± 0.010	0.214 ± 0.016
	Gemini 2.0	1-shot prompt.	0.869 ± 0.005	0.673 ± 0.009	0.759 ± 0.005
	Gemini 2.0	5-shot prompt.	0.820 ± 0.003	0.774 ± 0.006	0.797 ± 0.005
Saturates (Orange)	Llama 3 8B	0-shot prompt.	0.249 ± 0.033	0.032 ± 0.006	0.057 ± 0.010
	Llama 3 8B	1-shot prompt.	0.299 ± 0.024	0.182 ± 0.017	0.227 ± 0.019
	Llama 3 8B	5-shot prompt.	0.283 ± 0.011	0.365 ± 0.021	0.319 ± 0.014
	Gemini 2.0	0-shot prompt.	0.522 ± 0.013	0.083 ± 0.004	0.142 ± 0.007
	Gemini 2.0	1-shot prompt.	0.371 ± 0.006	0.660 ± 0.013	0.475 ± 0.006
	Gemini 2.0	5-shot prompt.	0.374 ± 0.016	0.600 ± 0.016	0.461 ± 0.017
Saturates (Red)	Llama 3 8B	0-shot prompt.	0.436 ± 0.009	0.195 ± 0.008	0.269 ± 0.008
	Llama 3 8B	1-shot prompt.	0.544 ± 0.011	0.616 ± 0.012	0.578 ± 0.008
	Llama 3 8B	5-shot prompt.	0.555 ± 0.005	0.505 ± 0.014	0.529 ± 0.007
	Gemini 2.0	0-shot prompt.	0.754 ± 0.040	0.123 ± 0.013	0.212 ± 0.020
	Gemini 2.0	1-shot prompt.	0.756 ± 0.012	0.574 ± 0.013	0.652 ± 0.010
	Gemini 2.0	5-shot prompt.	0.782 ± 0.007	0.508 ± 0.009	0.615 ± 0.005
Sugar (Green)	Llama 3 8B	0-shot prompt.	0.599 ± 0.025	0.094 ± 0.005	0.163 ± 0.009
	Llama 3 8B	1-shot prompt.	0.732 ± 0.006	0.741 ± 0.004	0.737 ± 0.003
	Llama 3 8B	5-shot prompt.	0.779 ± 0.007	0.746 ± 0.010	0.762 ± 0.007
	Gemini 2.0	0-shot prompt.	0.828 ± 0.008	0.215 ± 0.010	0.341 ± 0.014
	Gemini 2.0	1-shot prompt.	0.861 ± 0.008	0.714 ± 0.005	0.781 ± 0.003
	Gemini 2.0	5-shot prompt.	0.860 ± 0.008	0.735 ± 0.003	0.792 ± 0.004
Sugar (Orange)	Llama 3 8B	0-shot prompt.	0.266 ± 0.046	0.056 ± 0.008	0.093 ± 0.013
	Llama 3 8B	1-shot prompt.	0.325 ± 0.017	0.225 ± 0.010	0.266 ± 0.012
	Llama 3 8B	5-shot prompt.	0.331 ± 0.010	0.387 ± 0.012	0.357 ± 0.009
	Gemini 2.0	0-shot prompt.	0.495 ± 0.025	0.227 ± 0.006	0.311 ± 0.007
	Gemini 2.0	1-shot prompt.	0.421 ± 0.006	0.423 ± 0.012	0.422 ± 0.008
	Gemini 2.0	5-shot prompt.	0.417 ± 0.009	0.543 ± 0.015	0.472 ± 0.010
Sugar (Red)	Llama 3 8B	0-shot prompt.	0.431 ± 0.012	0.191 ± 0.013	0.265 ± 0.014
	Llama 3 8B	1-shot prompt.	0.567 ± 0.011	0.604 ± 0.013	0.585 ± 0.011
	Llama 3 8B	5-shot prompt.	0.606 ± 0.006	0.548 ± 0.004	0.576 ± 0.004
	Gemini 2.0	0-shot prompt.	0.765 ± 0.023	0.223 ± 0.025	0.345 ± 0.032
	Gemini 2.0	1-shot prompt.	0.657 ± 0.008	0.823 ± 0.008	0.730 ± 0.003
	Gemini 2.0	5-shot prompt.	0.709 ± 0.005	0.689 ± 0.011	0.699 ± 0.005

Table 15
Results by class for the Title + Ingredient dataset (1/2)

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Fat (Green)	Llama 3 8B	0-shot prompt.	0.603 ± 0.040	0.139 ± 0.008	0.226 ± 0.011
	Llama 3 8B	1-shot prompt.	0.573 ± 0.008	0.580 ± 0.013	0.576 ± 0.008
	Llama 3 8B	5-shot prompt.	0.672 ± 0.009	0.478 ± 0.011	0.558 ± 0.010
	Gemini 2.0	0-shot prompt.	0.673 ± 0.016	0.373 ± 0.013	0.480 ± 0.015
	Gemini 2.0	1-shot prompt.	0.836 ± 0.012	0.581 ± 0.009	0.685 ± 0.010
	Gemini 2.0	5-shot prompt.	0.821 ± 0.009	0.682 ± 0.013	0.745 ± 0.010
Fat (Orange)	Llama 3 8B	0-shot prompt.	0.409 ± 0.027	0.067 ± 0.007	0.116 ± 0.011
	Llama 3 8B	1-shot prompt.	0.447 ± 0.011	0.228 ± 0.009	0.302 ± 0.011
	Llama 3 8B	5-shot prompt.	0.432 ± 0.019	0.543 ± 0.014	0.481 ± 0.017
	Gemini 2.0	0-shot prompt.	0.584 ± 0.027	0.270 ± 0.014	0.369 ± 0.018
	Gemini 2.0	1-shot prompt.	0.518 ± 0.010	0.684 ± 0.012	0.589 ± 0.010
	Gemini 2.0	5-shot prompt.	0.547 ± 0.013	0.680 ± 0.007	0.606 ± 0.010
Fat (Red)	Llama 3 8B	0-shot prompt.	0.357 ± 0.014	0.211 ± 0.016	0.265 ± 0.016
	Llama 3 8B	1-shot prompt.	0.414 ± 0.008	0.614 ± 0.012	0.494 ± 0.009
	Llama 3 8B	5-shot prompt.	0.437 ± 0.012	0.438 ± 0.016	0.437 ± 0.014
	Gemini 2.0	0-shot prompt.	0.608 ± 0.011	0.433 ± 0.009	0.506 ± 0.008
	Gemini 2.0	1-shot prompt.	0.633 ± 0.013	0.590 ± 0.010	0.611 ± 0.009
	Gemini 2.0	5-shot prompt.	0.662 ± 0.010	0.582 ± 0.011	0.619 ± 0.009
Salt (Green)	Llama 3 8B	0-shot prompt.	0.651 ± 0.024	0.122 ± 0.002	0.205 ± 0.003

(*continued on next page*)

Table 15 (*continued*)

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
	Llama 3 8B	1-shot prompt.	0.617 ± 0.010	0.828 ± 0.009	0.707 ± 0.010
	Llama 3 8B	5-shot prompt.	0.630 ± 0.007	0.859 ± 0.005	0.727 ± 0.004
	Gemini 2.0	0-shot prompt.	0.778 ± 0.006	0.438 ± 0.013	0.560 ± 0.011
	Gemini 2.0	1-shot prompt.	0.761 ± 0.003	0.798 ± 0.007	0.779 ± 0.003
	Gemini 2.0	5-shot prompt.	0.734 ± 0.003	0.851 ± 0.008	0.788 ± 0.004
Salt (Orange)	Llama 3 8B	0-shot prompt.	0.394 ± 0.025	0.052 ± 0.001	0.091 ± 0.003
	Llama 3 8B	1-shot prompt.	0.395 ± 0.018	0.169 ± 0.008	0.237 ± 0.011
	Llama 3 8B	5-shot prompt.	0.390 ± 0.016	0.303 ± 0.016	0.341 ± 0.014
	Gemini 2.0	0-shot prompt.	0.506 ± 0.014	0.331 ± 0.013	0.400 ± 0.012
	Gemini 2.0	1-shot prompt.	0.446 ± 0.010	0.511 ± 0.010	0.476 ± 0.008
	Gemini 2.0	5-shot prompt.	0.457 ± 0.011	0.476 ± 0.012	0.466 ± 0.011
Salt (Red)	Llama 3 8B	0-shot prompt.	0.387 ± 0.017	0.108 ± 0.007	0.169 ± 0.010
	Llama 3 8B	1-shot prompt.	0.478 ± 0.016	0.339 ± 0.011	0.397 ± 0.013
	Llama 3 8B	5-shot prompt.	0.491 ± 0.015	0.234 ± 0.015	0.317 ± 0.016
	Gemini 2.0	0-shot prompt.	0.632 ± 0.020	0.239 ± 0.009	0.347 ± 0.010
	Gemini 2.0	1-shot prompt.	0.524 ± 0.018	0.370 ± 0.004	0.433 ± 0.005
	Gemini 2.0	5-shot prompt.	0.545 ± 0.010	0.332 ± 0.012	0.413 ± 0.011

Table 16
Results by class for the Title + Ingredient dataset (2/2)

Class	LLM	Experiment	Macro Avg. Prec.	Macro Avg. Rec.	Macro Avg. F1
Saturates (Green)	Llama 3 8B	0-shot prompt.	0.708 ± 0.038	0.065 ± 0.006	0.118 ± 0.011
	Llama 3 8B	1-shot prompt.	0.692 ± 0.004	0.647 ± 0.010	0.669 ± 0.006
	Llama 3 8B	5-shot prompt.	0.713 ± 0.011	0.623 ± 0.007	0.665 ± 0.008
	Gemini 2.0	0-shot prompt.	0.854 ± 0.033	0.116 ± 0.008	0.204 ± 0.013
	Gemini 2.0	1-shot prompt.	0.871 ± 0.003	0.629 ± 0.007	0.731 ± 0.004
	Gemini 2.0	5-shot prompt.	0.820 ± 0.006	0.759 ± 0.008	0.788 ± 0.007
Saturates (Orange)	Llama 3 8B	0-shot prompt.	0.258 ± 0.032	0.033 ± 0.003	0.058 ± 0.006
	Llama 3 8B	1-shot prompt.	0.307 ± 0.011	0.215 ± 0.007	0.253 ± 0.007
	Llama 3 8B	5-shot prompt.	0.287 ± 0.015	0.406 ± 0.015	0.336 ± 0.014
	Gemini 2.0	0-shot prompt.	0.519 ± 0.035	0.086 ± 0.008	0.148 ± 0.013
	Gemini 2.0	1-shot prompt.	0.355 ± 0.006	0.689 ± 0.018	0.469 ± 0.008
	Gemini 2.0	5-shot prompt.	0.360 ± 0.014	0.610 ± 0.011	0.452 ± 0.013
Saturates (Red)	Llama 3 8B	0-shot prompt.	0.437 ± 0.018	0.151 ± 0.013	0.225 ± 0.017
	Llama 3 8B	1-shot prompt.	0.523 ± 0.009	0.601 ± 0.013	0.559 ± 0.010
	Llama 3 8B	5-shot prompt.	0.548 ± 0.012	0.476 ± 0.007	0.509 ± 0.008
	Gemini 2.0	0-shot prompt.	0.716 ± 0.029	0.114 ± 0.008	0.197 ± 0.012
	Gemini 2.0	1-shot prompt.	0.762 ± 0.020	0.532 ± 0.009	0.627 ± 0.013
	Gemini 2.0	5-shot prompt.	0.768 ± 0.009	0.461 ± 0.007	0.576 ± 0.005
Sugar (Green)	Llama 3 8B	0-shot prompt.	0.657 ± 0.025	0.077 ± 0.010	0.137 ± 0.015
	Llama 3 8B	1-shot prompt.	0.753 ± 0.007	0.717 ± 0.013	0.735 ± 0.007
	Llama 3 8B	5-shot prompt.	0.794 ± 0.004	0.734 ± 0.005	0.763 ± 0.004
	Gemini 2.0	0-shot prompt.	0.843 ± 0.007	0.212 ± 0.003	0.338 ± 0.004
	Gemini 2.0	1-shot prompt.	0.872 ±0.003	0.698 ± 0.009	0.776 ± 0.006
	Gemini 2.0	5-shot prompt.	0.874 ± 0.004	0.723 ± 0.006	0.791 ± 0.003
Sugar (Orange)	Llama 3 8B	0-shot prompt.	0.285 ± 0.023	0.061 ± 0.006	0.101 ± 0.010
	Llama 3 8B	1-shot prompt.	0.332 ± 0.007	0.246 ± 0.015	0.282 ± 0.012
	Llama 3 8B	5-shot prompt.	0.330 ± 0.014	0.384 ± 0.012	0.355 ± 0.013
	Gemini 2.0	0-shot prompt.	0.474 ± 0.016	0.215 ± 0.015	0.296 ± 0.017
	Gemini 2.0	1-shot prompt.	0.417 ± 0.013	0.438 ± 0.011	0.427 ± 0.011
	Gemini 2.0	5-shot prompt.	0.416 ± 0.007	0.528 ± 0.015	0.466 ± 0.007
Sugar (Red)	Llama 3 8B	0-shot prompt.	0.462 ± 0.025	0.170 ± 0.006	0.249 ± 0.009
	Llama 3 8B	1-shot prompt.	0.553 ± 0.007	0.613 ± 0.013	0.582 ± 0.009
	Llama 3 8B	5-shot prompt.	0.585 ± 0.008	0.557 ± 0.012	0.571 ± 0.009
	Gemini 2.0	0-shot prompt.	0.732 ± 0.020	0.233 ± 0.014	0.354 ± 0.018
	Gemini 2.0	1-shot prompt.	0.654 ± 0.011	0.820 ± 0.011	0.728 ± 0.007
	Gemini 2.0	5-shot prompt.	0.690 ± 0.013	0.716 ± 0.014	0.703 ± 0.008

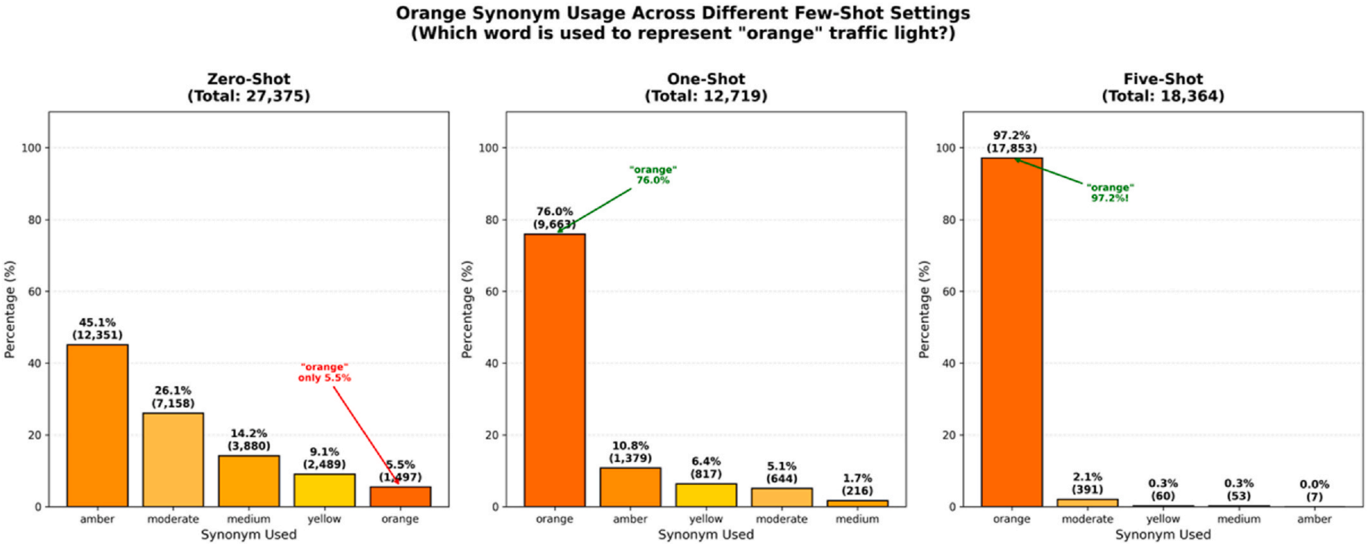


Fig. 8. Distribution of synonyms used by the Llama baseline model to represent the "orange" traffic light label across all four macronutrient classes (fat, salt, saturates, sugars) aggregated over all five cross-validation folds.

Data availability

https://huggingface.co/Matej/FoodyLLM.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt 4 technical report. arXiv preprint arXiv:2303.08774.

Agostoni, C., Boccia, S., Banni, S., Mannucci, P., Astrup, A., 2021. Sus tainable and personalized nutrition: from Earth health to public health. Eur. J. Intern. Med. 86, 12–16.

Arora, N.K., Mishra, I., 2023. Responsible consumption and production: a roadmap to sustainable development. Environ. Sustain. 6, 1–6.

Bairati, L., 2017. The food consumer's right to information on product country of origin: trends and outlook, beyond EU regulation 1169/2011. Journal of European Consumer and Market Law 6.

Balcombe, K., Fraser, I., Di Falco, S., 2010. Traffic lights and food choice: a choice experiment examining the relationship between nutritional food labels and price. Food Policy 35, 211–220.

Bermingham, K.M., Linenberg, I., Polidori, L., Asnicar, F., Arrè, A., Wolf, J., Badri, F., Bernard, H., Capdevila, J., Bulsiewicz, W.J., et al., 2024. Effects of a personalized nutrition program on cardiometabolic health: a randomized controlled trial. Nat. Med. 30, 1888–1897.

Blesh, J., Hoey, L., Jones, A.D., Friedmann, H., Perfecto, I., 2019. Development pathways toward “zero hunger”. World Dev. 118, 1–14.

Brega, A.G., Stotz, S.A., Moore, K.R., McNulty, M.C., Jiang, L., 2025. Reliability and validity of diabetes nutrition self-efficacy and behavior measures for the “what can i eat” diabetes nutrition education program for American indian and alaska native adults with type 2 diabetes. J. Acad. Nutr. Diet. 125, 78–89.

Brinkley, S., Gallo-Franco, J.J., Vázquez-Manjarrez, N., Chaura, J., Quartey, N.K., Toulabi, S.B., Odenkirk, M.T., Jermendi, E., Laporte, M.A., Lut terodt, H.E., et al., 2025. The state of food composition databases: data attributes and fair data harmonization in the era of digital innovation. Front. Nutr. 12, 1552367.

Carlsen, L., Bruggemann, R., 2022. The 17 united nations' sustainable development goals: a status by 2020. Int. J. Sustain. Dev. World Ecol. 29, 219–229.

Cenikj, G., Eftimov, T., Seljak, B.K., 2023a. Foodis: a food-disease relation mining pipeline. Artif. Intell. Med. 142, 102586.

Cenikj, G., Petelin, G., Seljak, B.K., Eftimov, T., 2022a. Scifoodner: food named entity recognition for scientific text. In: 2022 IEEE International Conference on Big Data (Big Data). IEEE, pp. 4065–4073.

Cenikj, G., Strojnik, L., Angelski, R., Ogrinc, N., Koroušić Seljak, B., Efti mov, T., 2023b. From language models to large-scale food and biomedical knowledge graphs. Sci. Rep. 13, 7815.

Cenikj, G., Valenčič, E., Ispirova, G., Ogrinc, M., Stojanov, R., Korošec, P., Cavalli, E., Seljak, B.K., Eftimov, T., 2022b. Cafeteriasa corpus: scientific abstracts annotated across different food semantic resources. Database 2022 baac107.

Chen, Z., Cano, A.H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., et al., 2023. Meditron 70b: scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O., 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457.

Comeau, D.C., Islamaj Doğan, R., Ciccarese, P., Cohen, K.B., Krallinger, M., Leitner, F., Lu, Z., Peng, Y., Rinaldi, F., Torii, M., et al., 2013. Bioc: a Minimalist Approach to Interoperability for Biomedical Text Processing. Database 2013.

Donnelly, K., et al., 2006. Snomed-ct: the advanced terminology and coding system for ehealth. Stud. Health Technol. Inf. 121, 279.

Dooley, D.M., Griffiths, E.J., Gosal, G.S., Buttigieg, P.L., Hoehndorf, R., Lange, M.C., Schriml, L.M., Brinkman, F.S., Hsiao, W.W., 2018. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food 2, 23.

Drole, J., Gjorgjevikj, A., Korušić Seljak, B., Eftimov, T., 2025. Beyond fine-tuning: robust food entity linking under ontology drift with FoodOntoRAG. In: Proceedings of the 2025 IEEE International Conference on Big Data (Big Data). IEEE (in press).

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Eftimov, T., Koroušić Seljak, B., Korošec, P., 2017. A rule-based named entity recognition method for knowledge extraction of evidence-based dietary recommendations. PLoS One 12, e0179488.

Emrich, T.E., Qi, Y., Lou, W.Y., L'Abbe, M.R., 2017. Traffic-light labels could reduce population intakes of calories, total fat, saturated fat, and sodium. PLoS One 12, e0171188.

Fang, Y., Deng, X., Liu, K., Zhang, N., Qian, J., Yang, P., Fan, X., Chen, H., 2025. A multi-modal ai copilot for single-cell analysis with instruction following. arXiv preprint arXiv:2501.08187.

Fathallah, N., Staab, S., Algergawy, A., 2024. Llms4life: large language models for ontology learning in life sciences. arXiv preprint arXiv:2412.02035.

Floridi, L., Chiriatti, M., 2020. Gpt-3: its nature, scope, limits, and consequences. Minds Mach. 30, 681–694.

Gjorshoska, I., Eftimov, T., Trajanov, D., 2022. Missing value imputation in food composition data with denoising autoencoders. J. Food Compos. Anal. 112, 104638.

Greenfield, H., Southgate, D.A., 2003. Food Composition Data: Production, Management, and Use. Food & Agriculture Org.

Guilpart, N., Iizumi, T., Makowski, D., 2022. Data-driven projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change. Nat. Food 3, 255–265.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J., 2021. Measuring massive multitask language understanding. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=d7KBjmI3GmQ.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2021. Lora: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Ispirova, G., Cenikj, G., Ogrinc, M., Valenčič, E., Stojanov, R., Korošec, P., Cavalli, E., Koroušić Seljak, B., Eftimov, T., 2022. Cafeteriafcd corpus: food consumption data annotated with regard to different food semantic resources. Foods 11, 2684.

Ispirova, G., Eftimov, T., Džeroski, S., Seljak, B.K., 2024. Msgen: measuring generalization of nutrient value prediction across different recipe datasets. Expert Syst. Appl. 237, 121507.

Ispirova, G., Eftimov, T., Seljak, B.K., 2020. Evaluating missing value imputation methods for food composition databases. Food Chem. Toxicol. 141, 111368.

Jain, S.M., 2022. Hugging face. In: Introduction to Transformers for NLP: with the Hugging Face Library and Models to Solve Problems. Springer, pp. 51–67.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al., 2021. Highly accurate protein structure prediction with alphafold. nature 596, 583–589.

Kim, Y., Xu, X., McDuff, D., Breazeal, C., Park, H.W., 2024. Health llm: large language models for health prediction via wearable sensor data. Proceedings of Machine Learning Research 248, 522–539.

Korošec, M., Golob, T., Bertoncelj, J., Stibilj, V., Seljak, B.K., 2013. The Slovenian food composition database. Food Chem. 140, 495–499.

Kunz, S., Haasova, S., Rieß, J., Florack, A., 2020. Beyond healthiness: the impact of traffic light labels on taste expectations and purchase intentions. Foods 9, 134.

Leclercq, C., Allemand, P., Balcerzak, A., Branca, F., Sousa, R., Lartey, A., Lipp, M., Quadros, V., Verger, P., 2019. Fao/who gift (Global Individual Food Consumption Data Tool): a global repository for harmonised individual quantitative food consumption studies. Proc. Nutr. Soc. 78, 484–495.

Lin, J., Wang, T., Lin, W., 2024. Becoming healthier without paying more? Experimental evidence from the impact of multiple traffic lights on Chinese college students. Nutrients 16, 2124.

Lin, S., Hilton, J., Evans, O., 2022. TruthfulQA: measuring how models mimic human falsehoods. In: Muresan, S., Nakov, P., Villavicencio, A. (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp. 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229. URL: https://aclanthology.org/2022.acl-long.229/.

Lindberg, D.A., Humphreys, B.L., McCray, A.T., 1993. The unified medical language system. Yearbook of medical informatics 2 (1), 41–51.

Luo, H., Liu, Y., Zhang, R., Wang, J., Sun, G., Niyato, D., et al., 2025a. Toward edge general intelligence with multiple-large language model (Multi-LLM): architecture, trust, and orchestration. IEEE Trans. Cogn. Commun. Netw.

Luo, H., Sun, G., Liu, Y., Zhao, D., Niyato, D., Yu, H., Dustdar, S., 2025b. A weighted byzantine fault tolerance consensus driven trusted multiple large Language models network. IEEE Trans. Cogn. Commun. Netw.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., Zhang, Y., 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint arXiv:2308.08747.

Luo, Y., Zhang, J., Fan, S., Yang, K., Hong, M., Wu, Y., Qiao, M., Nie, Z., 2024. Biomedgpt: an open multimodal large language model for biomedicine. IEEE J. Biomed. Health Inform.

Ma, P., Tsai, S., He, Y., Jia, X., Zhen, D., Yu, N., Wang, Q., Ahuja, J.K., Wei, C.I., 2024. Large language models in food science: innovations, applications, and future. Trends Food Sci. Technol., 104488

Marın, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A., 2021. Recipe1m+: a dataset for learning cross-modal embeddings for cooking recipes and food images. IEEE Trans. Pattern Anal. Mach. Intell. 43, 187–203.

Merchant, A.T., Dehghan, M., 2006. Food composition database development for between-country comparisons. Nutr. J. 5, 1–8.

MosaicML, 2023. Dolly-hhrlhf dataset. https://huggingface.co/datasets/mosaicml/dolly_hhrlhf.

Neuhouser, M.L., Prentice, R.L., Tinker, L.F., Lampe, J.W., 2023. Enhancing capacity for food and nutrient intake assessment in population sciences research. Annu. Rev. Publ. Health 44, 37–54.

Näslund-Koch, C., Kjeldsen, E.W., Vedel-Krogh, S., Bojesen, S.E., Skov, L., 2024. Adherence to general national dietary guidelines and risk of psoriasis: results from a general population study of 105,332 individuals. Clin. Exp. Dermatol. llae091.

Ogrinc, M., Koroušić Seljak, B., Eftimov, T., 2024. Zero-shot evaluation of chatgpt for food named-entity recognition and linking. Front. Nutr. 11, 1429259.

Parliament, T., 2022. Hansard. URL. https://hansard.parliament.uk.

Popovski, G., Kochev, S., Korousic-Seljak, B., Eftimov, T., 2019a. Foodie: a rule-based named-entity recognition method for food information extraction. ICPRAM 12, 915.

Popovski, G., Korousic-Seljak, B., Eftimov, T., 2019b. Foodontomap: linking food concepts across different food ontologies. In: KEOD, pp. 195–202.

Popovski, G., Seljak, B.K., Eftimov, T., 2019c. Foodbase Corpus: a New Resource of Annotated Food Entities. Database 2019, baz121.

Porter, R., Diehl, A., Pastel, B., Hinnefeld, J.H., Nerenberg, L., Maung, P., Kerbrat, S., Hanson, G., Astorino, T., Tarsa, S.J., 2024. Llmd: a large language model for interpreting longitudinal medical records. arXiv preprint arXiv:2410.12860.

Ravandi, B., Ispirova, G., Sebek, M., Mehler, P., Barabási, A.L., Menichetti, G., 2025. Prevalence of processed foods in major us grocery stores. Nat. Food 1–13.

Sacks, G., Rayner, M., Swinburn, B., 2009. Impact of front-of-pack 'traffic light'nutrition labelling on consumer food purchases in the UK. Health Promot. Int. 24, 344–352.

Saldanha, M.F., Dos Santos, R.R., Jansen, A.K., 2024. Energy and nutrient intake in older adults with healthy aging during the coronavirus disease 19 pandemic–a cohort study. Clinical Nutrition ESPEN 59, 181–187.

Sarasty, O., Carpio, C., Cabrera, T., 2023. Effect of the traffic-light system on nutrition labeling in processed food products in the ecuadorian population. Q Open 3 qoad018.

Schanes, K., Dobernig, K., Gözet, B., 2018. Food waste matters systematic review of household food waste practices and their policy implications. J. Clean. Prod. 182, 978–991.

Shanahan, M., 2024. Talking about large language models. Commun. ACM 67, 68–79.

Song, L., Segal, E., Xing, E., 2024. Toward ai-driven digital organism: multiscale foundation models for predicting, simulating and programming biology at all levels. arXiv preprint arXiv:2412.06993.

Sonnenberg, L., Gelsomin, E., Levy, D.E., Riis, J., Barraclough, S., Thorndike, A.N., 2013. A traffic light food labeling intervention increases consumer awareness of health and healthy choices at the point-of-purchase. Prev. Med. 57, 253–257.

Stojanov, R., Popovski, G., Cenikj, G., Koroušić Seljak, B., Eftimov, T., 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: algorithm development and validation. J. Med. Internet Res. 23, e28229.

T Kennedy, E., Bowman, S.A., T Spence, J., Freedman, M., King, J., 2001. Popular diets: correlation to health, nutrition, and obesity. J. Am. Diet. Assoc. 101, 411–420.

Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W., 2023. Large language models in medicine. Nat. Med. 29, 1930–1940.

Top, J., Janssen, S., Boogaard, H., Knapen, R., Şimşek-Şenel, G., 2022. Cultivating fair principles for agri-food data. Comput. Electron. Agric. 196, 106909.

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al., 2024. Towards generalist biomedical ai. Nejm Ai 1. AIoa2300138.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., Choi, Y., 2019. Hellaswag: can a machine really finish your sentence?. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Zhang, Q., Ding, K., Lv, T., Wang, X., Yin, Q., Zhang, Y., Yu, J., Wang, Y., Li, X., Xiang, Z., et al., 2025. Scientific large language models: a survey on biological & chemical domains. ACM Comput. Surv. 57, 1–38.

Zhao, F.F., He, H.J., Liang, J.J., Cen, J., Wang, Y., Lin, H., Chen, F., Li, T.P., Yang, J.F., Chen, L., et al., 2025. Benchmarking the performance of large language models in uveitis: a comparative analysis of chatgpt-3.5, chatgpt-4.0, google gemini, and anthropic claude3. Eye 39, 1132–1137.

Zhou, P., Min, W., Fu, C., Jin, Y., Huang, M., Li, X., et al., 2025. FoodSky: a food-oriented large language model that can pass the chef and dietetic examinations. Patterns 6 (5).