FoodBench-QA: Overview of the Shared Task on Grounded Food and Nutrition Question Answering

Tome Eftimov[1,2], **Ana Gjorgjevikj**[1], **Matej Martinc**[1], **Gjorgjina Cenikj**[1], **Sašo Džeroski**[1] , **Barbara Koroušić Seljak**[1,3]

[1] Jožef Stefan Institute, Jamova cesta 39, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
[3] Medical Faculty, University of Ljubljana, Vrazov trg 2, 1000 Ljubljana, Slovenia
{tome.eftimov, ana.gjorgjevikj, matej.martinc, gjorgjina.cenikj, saso.dzeroski, barbara.korousic}@ijs.si

Abstract

We present the results of the FoodBench-QA 2026 shared task at the CL4Health workshop, collocated with LREC 2026. FoodBench-QA challenges systems to answer food and nutrition questions using evidence from food composition databases and food-related ontologies. The shared task comprises three main tasks: nutrient estimation from recipe ingredients, evaluated using EU Regulation 1169/2011 tolerance thresholds; FSA traffic-light classification for fat, salt, saturates, and sugars; and food named entity recognition and linking to three ontologies, namely Hansard Taxonomy, FoodOn, and SNOMED CT. We received submissions from five participating teams across all tasks. For nutrient estimation, the best system achieved accuracy rates of 93.57% for protein, 86.50% for sugars, 84.65% for fat, and 86.26% for saturates. For FSA traffic-light prediction, the best macro F1 scores ranged from 0.65 to 0.90 across different nutrient-color combinations. For named entity linking, the best systems achieved macro F1 scores between 60.71% and 80.89% for natural text and 87.75% and 95.75% for artificial NEL datasets, depending on the ontology.

Keywords: question answering, nutrient estimation, food traffic light classification, named-entity linking

1. Introduction

Accurate food and nutrition information is essential for public health, dietary assessment, and personalized nutrition recommendations (Carlsen and Bruggemann, 2022; Agostoni et al., 2021; Guilpart et al., 2022). However, extracting and linking food-related information from text remains challenging due to the complexity of food terminology, the variety of ingredient descriptions, and the need to ground answers in authoritative food composition databases and food-related ontologies (Greenfield and Southgate, 2003; Gjorshoska et al., 2022).

The FoodBench-QA 2026 shared task addresses these challenges by providing a comprehensive evaluation framework for food and nutrition question answering systems. The task is designed to assess systems' abilities to estimate nutritional content (protein, sugars, fat, saturated fat) from recipe ingredients according to European Union (EU) regulatory standards (Bairati, 2017), classify food items according to the UK Food Standards Agency (FSA) traffic-light labeling system (Balcombe et al., 2010), and recognize and link food entities to standardized ontologies including Hansard Taxonomy (Abercrombie and Batista-Navarro, 2018), FoodOn (Dooley et al., 2018), and SNOMED CT (Donnelly et al., 2006).

The challenge builds upon the FoodBench dataset, a curated benchmark of question-answer pairs designed for training and evaluating large language models in food and nutrition domains. This shared task represents the first community evaluation focused specifically on grounded food and nutrition question answering, following the tradition of successful BioNLP shared tasks.

FoodBench-QA 2026 was organized as part of the Third Workshop on Patient-Oriented Language Processing (CL4Health), collocated with LREC 2026: Language Resources and Evaluation Conference in Palma, Mallorca, Spain. The workshop provides a venue for computational linguistics research focused on patients' health and health-related issues concerning the public.

2. Background

To contextualize our approach, we summarize prior work in food and nutrition natural language processing (NLP), relevant ontologies, and applicable EU regulations.

Food and Nutrition NLP. Natural language processing for food and nutrition has gained increasing attention due to its applications in dietary assessment, food safety monitoring, and personalized nutrition (Cenikj et al., 2025). Previous work has addressed various aspects, including ingredient recognition (Stojanov et al., 2021, 2020; Cenikj et al., 2022a, 2020), recipe parsing (Popovski et al., 2019a,b; Ispirova et al., 2022), and nutrient estimation (Ispirova et al., 2024). However, standardized evaluation benchmarks have been lacking in this domain. Such benchmarks are essential in NLP and machine learning because they enable fair, reproducible, and transparent comparison of models

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 223–233
12 May 2026.

using shared datasets and metrics. They help track progress, identify strengths and weaknesses, and support scientific rigor through replicability. At the same time, they must be carefully designed and regularly updated to avoid overfitting and ensure real-world relevance.

Food Ontologies and Datasets. Three major semantic resources are used in this shared task. The **Hansard Taxonomy** is a hierarchical classification system for food products using structured codes such as AG.01.k [Flour], providing standardized categorization for food items (Parliament, 2022). **FoodOn** is a comprehensive ontology that covers food sources, food products, and their production processes, using URIs from the OBO (Open Biological and Biomedical Ontology) namespace and including links to NCBI Taxonomy for biological species (Dooley et al., 2018). **SNOMED CT**, the Systematized Nomenclature of Medicine Clinical Terms, includes food-related concepts relevant for clinical and medical applications, with URIs following the BioOntology format (Donnelly et al., 2006).

EU Regulation 1169/2011. The European Union Regulation 1169/2011 on food information to consumers establishes tolerance thresholds for nutritional declarations (Bairati, 2017). These thresholds define acceptable deviations between declared and actual nutrient values, providing a natural evaluation framework for nutrient estimation systems.

3. Task Descriptions

FoodBench-QA 2026 comprises three main tasks, with subtasks based on input modalities (Gjorgjevikj et al., 2026).

Task 1: Nutrient Estimation. The nutrient estimation task requires systems to predict the nutritional content per 100g for recipes based on their ingredients. Two subtasks were defined based on the input provided to systems (Neuhouser et al., 2023; Ispirova et al., 2024). In Task 1.1 (Ingredients Only), systems receive only the list of ingredients with quantities, while in Task 1.2 (Title + Ingredients), systems receive both the recipe title and the complete ingredient list. Four nutrients are evaluated: protein, sugars, fat, and saturated fat (saturates), with all values expressed in grams per 100g.

Task 2: FSA Traffic-Light Classification. The FSA traffic-light task requires systems to classify the nutritional quality of recipes according to the UK Food Standards Agency color-coded system (Balcombe et al., 2010; Kunz et al., 2020; Emrich et al., 2017). Each of four nutrients (fat, salt, saturates, sugars) must be classified as green (low), orange (medium), or red (high). Similar to Task 1, two subtasks were defined: Task 2.1 performs classification based on the ingredient list and their quantities, while Task 2.2 uses the title, ingredients, and their quantities. This formulation results in a 12-class classification problem with four nutrients multiplied by three color categories.

Nutrient	t_i	d_i
Protein, Sugars	≤ 10	≤ 2
	$10 < t_i \leq 40$	$\leq 0.2t_i$
	> 40	≤ 8
Fat	≤ 10	≤ 1.5
	$10 < t_i \leq 40$	$\leq 0.2t_i$
	> 40	≤ 8
Sat. Fat	< 4	≤ 0.8
	≥ 4	$\leq 0.2t_i$

Table 1: Tolerance rules used to evaluate nutrient estimation. t_i denotes the ground truth value, r_i the predicted value, and $d_i = |t_i - r_i|$ the absolute difference.

Task 3: Named-Entity Recognition and Linking. The NER+NEL task requires systems to identify food entities in text and link them to ontology concepts (Eftimov et al., 2017; Popovski et al., 2019a; Cenikj et al., 2020; Stojanov et al., 2021; Cenikj et al., 2022a; Gjorgjevikj et al., 2025). Two subtasks address different data sources. Task 3.1 (NER+NEL) is performed on two corpora: Scientific Abstracts (SA) containing technical food science literature (Cenikj et al., 2022b), and Food Composition Data (FCD) containing recipe descriptions (Ispirova et al., 2022). Task 3.2 (Artificial NEL) focuses on pure entity linking with pre-identified entities. In both subtasks, three ontologies have been considered: FoodOn, SNOMED CT, and Hansard Taxonomy.

4. Evaluation Metrics

Below we summarize the evaluation framework and metrics used to measure system effectiveness.

Task 1: Nutrient Estimation Metrics. Nutrient estimation is evaluated using tolerance thresholds defined in EU Regulation 1169/2011. A prediction is considered within tolerance based on specific rules for each nutrient type (see Table 1). The accuracy for each nutrient is calculated as the proportion of predictions falling within the tolerance threshold.

Task 2: FSA Traffic-Light Metrics. Each category-color combination is treated as a separate class, resulting in 12 classes total. Evaluation metrics include per-class Precision, Recall, and F1-Score.

Task 3: NER+NEL Metrics Entity linking is evaluated using two averaging methods. Macro Average calculates metrics per semantic tag and then averages them, giving equal weight to each semantic tag. Weighted Average weights metrics by the number of ground truth links per semantic tag. For each averaging method, we report Precision, Recall, and F1-Score. Entities with multiple tags are evaluated independently for each entity-link pair (This occurs due to the semantic nature of the tags (parent–child hierarchy) or the use of general food groups when the specific entity is not available).

5. Datasets

This section presents the datasets constructed for each task and summarizes their main statistics. Further details on the dataset construction process, the resources from which the datasets were derived, and the corresponding train–test splits are available in (Gjorgjevikj et al., 2026).

Task 1 and Task 2: Recipe Data. The nutrient estimation and FSA traffic-light datasets are derived from the Recipe1+ dataset (Marın et al., 2021). Table 2 reports the number of QA pairs in the datasets for Tasks 1 and 2.

Task	Training	Test
T1.1 – Ingredients Only	58,567	29,648
T1.2 – Title + Ingredients	58,652	29,649
T2.1 – Ingredients Only	58,585	29,578
T2.2 – Title + Ingredients	58,583	29,671

Table 2: Number of QA pairs in the FoodBench-QA T1 and T2 datasets.

Task 3.1: Natural Text NER+NEL. Two corpora were provided for the NER+NEL task on natural text. The Food Composition Data (FCD) contains 997 recipe descriptions with annotated food entities, split into 514 training recipes (51.6%) and 483 test recipes (48.4%) using a stratified split by recipe categories (Popovski et al., 2019b; Ispirova et al., 2022). The Scientific Abstracts (SA) corpus contains 470 scientific abstracts from food science literature, split into 329 training abstracts (70.0%) and 141 test abstracts (30.0%) using a random 70/30 split due to high category diversity (Cenikj et al., 2022b). Both datasets include annotations for three ontologies: Hansard Taxonomy, FoodOn, and SNOMED CT.

Task 3.2: Artificial NEL. Three artificial entity linking datasets were created with pre-identified food entities. Artificial IR instances were generated to balance the distribution of food entities by identifying underrepresented entities for T3.1 and augmenting them until a minimum threshold of 150 mentions per entity was reached. The generated instances consist only of NEL instruction–response pairs linking sampled entity labels to FoodOn, Hansard, and SNOMED-CT, ensuring that all ontology entities are observed during fine-tuning while preventing duplicate examples to avoid data leakage (more details in (Gjorgjevikj et al., 2025). The FoodOn NEL dataset is the largest with 13,492 entries, split into 9,445 training entries (70.0%) and 4,047 test entries (30.0%), containing 587 unique entities of which 42 are rare entities reserved for zero-shot evaluation. The SNOMEDCT NEL dataset contains 4,445 entries with a medical and clinical terminology focus, split into 3,112 training entries (70.0%) and 1,333 test entries (30.0%), with 310 unique entities including 30 rare entities. The Hansard NEL dataset contains 1,611 entries with the highest entity diversity, split into 915 training entries (56.8%) and 696 test entries (43.2%), featuring 1,082 unique entities of which 568 (52.6%) are rare entities, providing the most challenging zero-shot evaluation scenario.

All artificial NEL datasets employ stratified splits with rare entity protection, ensuring that entities appearing three or fewer times in the full dataset are exclusively placed in the test set for zero-shot evaluation. Table 3 summarizes the complete dataset statistics for the NER+NEL subtask.

6. Participating Systems

Five teams registered for FoodBench-QA 2026, with varying levels of participation across tasks. The best-ranked team (labeled as Team 1) participated in all tasks and submitted the most comprehensive set of runs, achieving the best results across most subtasks. The second-best team (labeled as Team 2) also participated in all tasks with competitive submissions. Team 3 focused on Task 1 (nutrient estimation), while Team 4 and Team 5 participated in the development phase. Participants were allowed to submit multiple runs to explore different configurations. The test phase ran from February 1-9, 2026, with submissions evaluated on the CodaBench platform .

[0] https://www.codabench.org/competitions/12112/

Dataset	Total	Train	Test	Unique Ent.	Rare Ent.
T3.1.1 – SA (Abstracts)	470	329	141	–	–
T3.1.2 – FCD (Recipes)	997	514	483	–	–
T3.2.1 – Artificial FoodOn	13,492	9,445	4,047	587	42
T3.2.2 – Artificial SNOMEDCT	4,445	3,112	1,333	310	30
T3.2.3 – Artificial Hansard	1,611	915	696	1,082	568
Total	21,015	14,315	6,700	–	–

Table 3: Summary of FoodBench-QA 2026 NER+NEL Datasets.

Applied methodologies. **Team 1:** For T1, a rule-based, evidence-grounded pipeline was developed in which heterogeneous ingredient measurements were resolved, quantities were standardized under SI standards (NIST SP 811), and interpretable, recipe-level nutrient values were deterministically estimated using publicly available resources (USDA FoodData Central and Recipe1M+). For T2 and T3, lightweight text-based classification was applied for FSA traffic-light prediction, and rule-based entity recognition was employed together with dictionary-driven ontology linking. **Team 2:** For T1, nutrient estimation and entity linking were framed as retrieval tasks rather than generative tasks, using TF-IDF and k-NN to identify the most similar recipes in the training database. To preserve authentic nutrient profiles, a conditional routing rule was implemented: when a retrieved recipe was nearly identical (over 95% similarity), ensembling was bypassed and the exact nutritional values were directly adopted. For standard matches below this threshold, a squared weighted voting mechanism was applied to ensure that high-confidence neighbors dominated the final prediction. For T2, an identical retrieval-augmented architecture to T1 was utilized. TF-IDF vectorization and weighted k-NN retrieval were applied to predict Food Standards Agency categories based on the closest neighbors. The same conditional routing rule was also applied to bypass ensembling and directly adopt categories from nearly identical matches (over 95% similarity). For T3.1 and 3.2 (NER and NEL), a unified Hybrid Regex-Dictionary approach was employed for both Named Entity Recognition and Linking. Recall was optimized using an automatic pluralization engine, while high precision was ensured through a Longest-Match-First regex strategy to prevent substring errors. The extracted entities were then mapped directly to the FoodOn, SNOMED CT, and Hansard ontologies using a high-precision dictionary, supported by an automated post-processing pipeline for strict formatting compliance. **Team 3:** For T1, multiple modeling strategies were evaluated to estimate recipe-level nutrients from unstructured ingredient lists, ranging from traditional lexical matching to large language models. A TF-IDF representation coupled with Ridge Regression was used as a lightweight baseline, providing moderate estimation accuracy with near-instantaneous inference. To assess deeper semantic modeling, the DeBERTa-v3 encoder was evaluated, although its performance was limited by the scarcity of task-specific training data. In addition, few-shot inference with large language models (e.g., Gemma-3-27B) and a hybrid refinement pipeline combining TF-IDF retrieval with Gemini 2.5 Flash were explored. These approaches leverage pre-trained world knowledge to implicitly perform ingredient disambiguation and unit normalization, achieving higher accuracy under the strict tolerance criteria of EU Regulation 1169/2011, albeit at the cost of increased computational latency. For T2 and T3, lightweight text-based classification was applied for FSA traffic-light prediction, while rule-based entity recognition combined with ontology-based entity linking was used to map extracted food entities to controlled vocabularies. **Baseline:** To assess whether LLM-based solutions show promising signals when trained on the data, a previous study reports results for each task (Gjorgjevikj et al., 2026). However, these results should be interpreted cautiously, as the evaluations were not performed using identical training and test splits.

7. Results

We present the task-wise results and compare the performance of the submitted systems.

Task 1: Nutrient Estimation Results. Table 4 presents the results for the nutrient estimation task, where accuracy represents the percentage of predictions within EU Regulation 1169/2011 tolerance thresholds.

The best results were achieved by “Team 1”, with accuracy rates exceeding 84% for all nutrients when using title and ingredients (T1.2) (see Figure 1). Adding recipe titles provided marginal improvements over ingredients-only predictions, with the largest gain observed for sugars at 0.06 percentage points.

Task	Team	Protein	Sugars	Fat	Saturates
T1.1	Team 1 (best) (six entries)	**93.44**	**86.44**	**84.46**	**86.18**
	Team 2 (a single entry)	76.35	69.66	67.45	71.55
	Team 3 (two entries)	59.71	43.03	39.21	47.13
T1.2	Team 1 (best) (six entries)	**93.57**	**86.50**	**84.65**	**86.26**
	Team 2 (a single entry)	77.73	71.67	69.28	73.41

Table 4: Task 1 Results: Nutrient Estimation Accuracy (%)

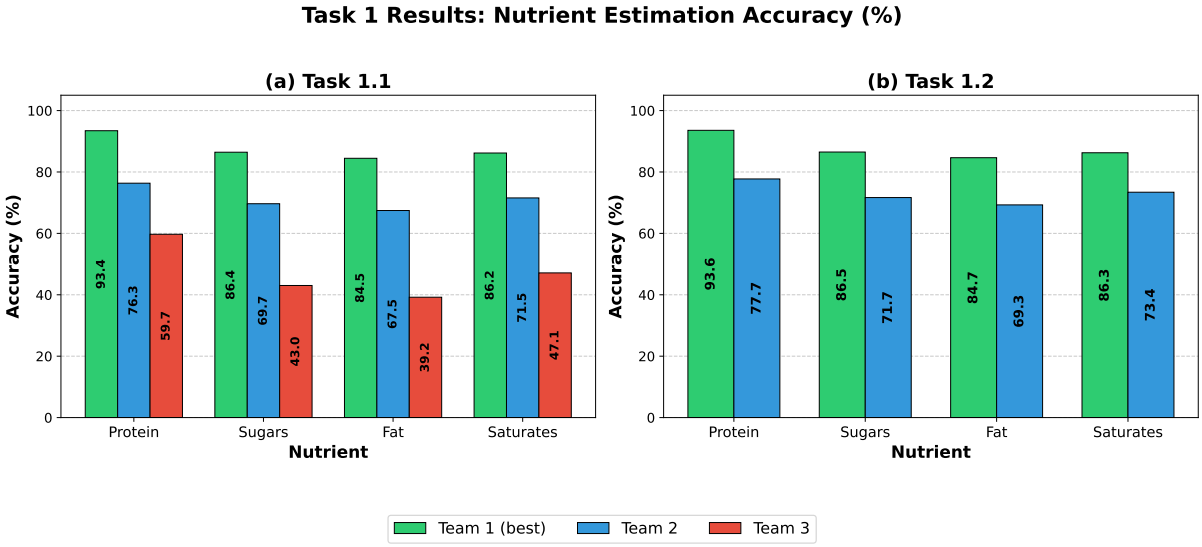

Figure 1: Task 1 results: Nutrient estimation accuracy (%) across participating teams.

Task 2: FSA Traffic-Light Results. Table 5 presents the F1-scores for each nutrient-color combination, showing results for the best-performing submissions (also see Figure 2). The results show that green (low) classifications are generally easier to predict than orange (medium) and red (high) classifications. The orange category consistently shows the lowest F1-scores across all nutrients, suggesting that medium-level classifications are most challenging. Adding recipe titles improved classification performance across most categories. Across both teams, the best overall performance was achieved by Team 2 in Task 2.1 (notably for Saturates and Salt), while Team 1 dominated Task 2.2, consistently obtaining the highest F1-scores across most nutrients and traffic-light classes when title information was included.

Task 3.1: NER+NEL on Natural Text Results Table 6 presents the entity linking results for natural text datasets.

The results demonstrate that SNOMED CT linking achieves the highest F1-scores for both SA and FCD datasets, followed by FoodOn and Hansard. The higher precision on FCD compared to SA suggests that recipe text is easier to process than scientific abstracts. Across both subtasks, clear performance differences emerge between the two teams. On T3.1.1 (Scientific Abstracts), Team 1 consistently outperforms Team 2 in terms of Weighted F1 across all three ontologies, with particularly strong results for SNOMED CT (F1 = 80.89), where it achieves both the highest precision and recall. Team 2 remains competitive in precision, especially for FoodOn and SNOMED CT, but lower recall limits its overall F1 performance on this dataset. In contrast, on T3.1.2 (Food Composition Data), Team 2 demonstrates stronger overall performance, achieving the highest Weighted F1 scores for all three ontologies, including the best overall result (F1 = 80.89 on SNOMED CT). While Team 1 maintains strong recall, particularly for Hansard and SNOMED CT, Team 2's higher precision across ontologies leads to superior overall F1 performance. Notably, SNOMED CT emerges as the most robust ontology across both datasets and teams, consistently yielding the highest scores (see Figure 3).

Task 3.2: Artificial NEL Results. Table 7 presents the entity linking results for the artificial NEL datasets.

The artificial NEL task shows significantly higher performance than natural text NER+NEL (see Figure 4). In Task 3.2 (Artificial NEL), Team 1 clearly

Figure 2: Task 2 results: FSA traffic-light classification F1-scores across nutrients and teams.

Nutrient	Team	Green	Orange	Red
Task 2.1 (Ingredients Only)				
Fat	Team 1	**0.828**	0.709	**0.791**
Salt	Team 1	**0.875**	0.657	0.758
Saturates	Team 1	0.851	0.577	0.802
Sugars	Team 1	0.860	0.616	0.806
Fat	Team 2	**0.828**	**0.750**	0.764
Salt	Team 2	0.874	**0.743**	**0.768**
Saturates	Team 2	**0.858**	**0.707**	**0.814**
Sugars	Team 2	**0.867**	**0.656**	0.693
Task 2.2 (Title + Ingredients)				
Fat	Team 1	**0.856**	**0.756**	**0.818**
Salt	Team 1	**0.898**	0.718	**0.794**
Saturates	Team 1	**0.877**	0.647	**0.831**
Sugars	Team 1	**0.880**	**0.687**	**0.847**
Fat	Team 2	0.822	0.749	0.765
Salt	Team 2	0.866	**0.730**	0.764
Saturates	Team 2	0.857	**0.707**	0.812
Sugars	Team 2	0.866	0.662	0.692

Table 5: Task 2 Results: FSA Traffic-Light F1-Scores

outperforms Team 2 across all ontologies and evaluation metrics. Team 1 achieves substantially higher recall, particularly for FoodOn (94.80) and SNOMED CT (98.90), which translates into consistently superior F1 scores. The largest performance gap is observed for Hansard, where Team 1 reaches an F1 of 87.75 compared to 56.20 for Team 2. Although Team 2 remains competitive in precision—especially for FoodOn and SNOMED CT—the significantly lower recall limits its overall effectiveness.

8. Analysis of Results

The following analysis highlights major trends and insights derived from the experimental results.

Task Difficulty Analysis: The results reveal a clear hierarchy of task difficulty. Artificial NEL proved to be the easiest task, with weighted F1-scores reaching up to 95.75%, benefiting from pre-identified entities that eliminate the need for entity recognition. Nutrient estimation achieved accuracy rates up to 93.57%, leveraging well-defined tolerance thresholds. FSA traffic-light classification showed F1-scores ranging from 0.61 to 0.90,

Dataset	Ontology	Team	Precision	Recall	F1
T3.1.1 – Scientific Abstracts (SA)					
SA	Hansard	Team 1	**54.11**	**71.51**	**60.71**
SA	Hansard	Team 2	47.99	38.38	41.13
SA	FoodOn	Team 1	74.53	**71.48**	**71.38**
SA	FoodOn	Team 2	**75.16**	54.23	61.43
SA	SNOMEDCT	Team 1	**83.90**	**84.34**	**80.89**
SA	SNOMEDCT	Team 2	**83.90**	69.40	74.32
T3.1.2 – Food Composition Data (FCD)					
FCD	Hansard	Team 1	74.35	**77.06**	75.09
FCD	Hansard	Team 2	**81.04**	75.60	**77.75**
FCD	FoodOn	Team 1	91.15	**70.39**	77.92
FCD	FoodOn	Team 2	**94.67**	69.98	**79.01**
FCD	SNOMEDCT	Team 1	91.30	**72.65**	79.42
FCD	SNOMEDCT	Team 2	**95.10**	72.51	**80.89**

Table 6: Task 3.1 Weighted Results: NER+NEL on Natural Text (Best Submissions)

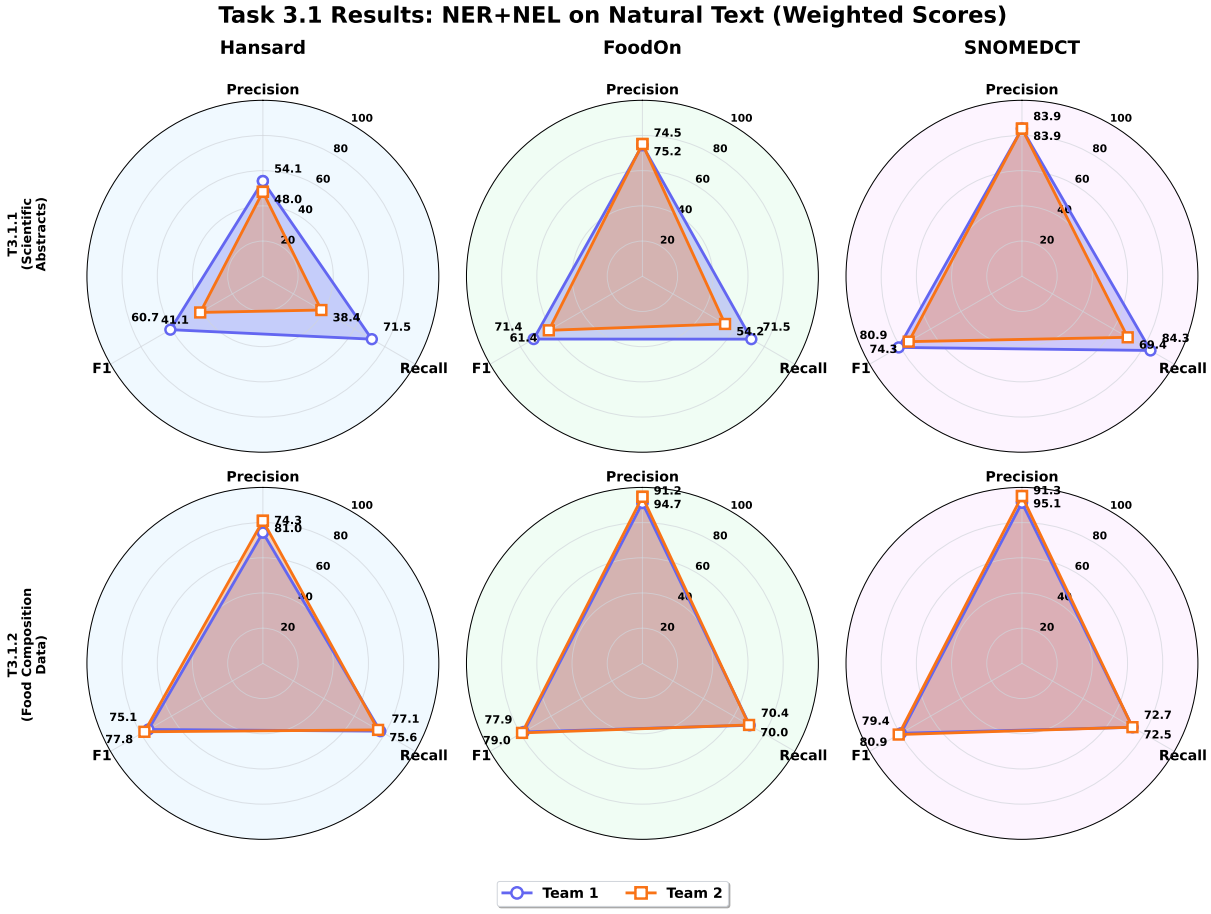

Figure 3: Performance evaluation of named entity recognition and linking on natural text: A comparison across Scientific Abstracts (T3.1.1) and Food Composition Data (T3.1.2).

with orange classifications proving most challenging across all nutrients. Natural text NER+NEL emerged as the hardest task with F1-scores up to 80.89%, as it requires both accurate entity recognition and correct linking to ontology concepts.

Impact of Recipe Titles: Comparing the ingredients-only subtasks (T1.1 and T2.1) with the title plus ingredients subtasks (T1.2 and T2.2) reveals the impact of additional contextual information. Nutrient estimation showed marginal

Dataset	Ontology	Team	Precision	Recall	F1
Task 3 – Natural Text Evaluation					
Artificial text	Hansard	Team 1	**90.53**	**85.89**	**87.75**
Artificial text	Hansard	Team 2	75.48	45.86	56.20
Artificial text	FoodOn	Team 1	86.35	**94.80**	**90.04**
Artificial text	FoodOn	Team 2	**92.49**	64.56	74.71
Artificial text	SNOMED CT	Team 1	93.07	**98.90**	**95.75**
Artificial text	SNOMED CT	Team 2	**93.22**	72.01	80.68

Table 7: Task 3.2 Weighted Results: Artificial NEL (Best Submissions).

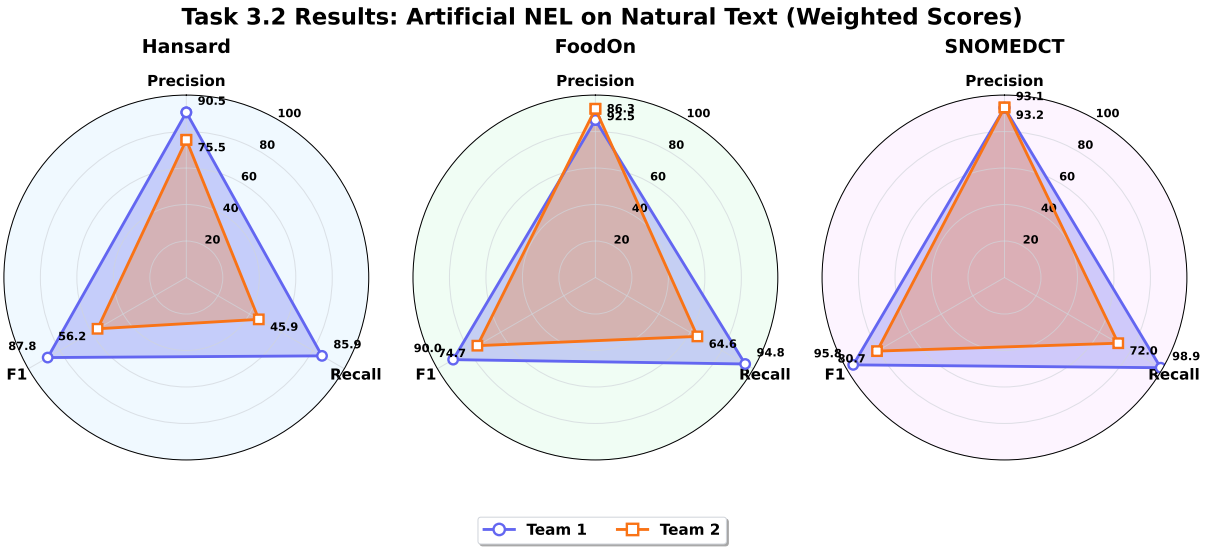

Figure 4: Performance analysis of named entity linking on artificially generated text using Hansard, FoodOn, and SNOMED CT ontologies.

improvements of approximately 0.13 percentage points for protein and 0.06 percentage points for sugars. FSA traffic-light classification demonstrated more substantial improvements, particularly for red classifications. Recipe titles provide contextual information that helps disambiguate ingredient interpretations and improve overall prediction accuracy.

Ontology-Specific Performance: Across both natural text (scientific abstracts and recipe description) and artificial NEL tasks, consistent patterns emerged in ontology-specific performance. SNOMED CT consistently achieved the highest F1-scores, likely due to more standardized medical terminology that facilitates accurate entity linking. FoodOn showed strong performance with good coverage of food items across diverse categories.

Zero-Shot Evaluation: The artificial NEL datasets were specifically designed with rare entities (appearing three or fewer times) exclusively in the test set to enable zero-shot evaluation. The FoodOn dataset contains 42 rare entities, SNOMED CT contains 30 rare entities, and Hansard contains 568 rare entities representing 52.6% of its unique entities. The Hansard dataset provides the most challenging zero-shot scenario due to its high proportion of rare entities, testing systems' ability to generalize to previously unseen food terminology.

9. Conclusions and Future Work

FoodBench-QA 2026 established the first community benchmark for grounded food and nutrition question answering, attracting five teams and providing insights into the capabilities and limitations of current food NLP systems. It introduced a multi-task evaluation framework for nutrient estimation, FSA classification, and entity linking, with datasets, evaluation scripts, and submissions publicly available on CodaBench`https://www.codabench.org/competitions/12112/` and Zenodo at `https://doi.org/10.5281/zenodo.17798877`.

10. Acknowledgments

The authors acknowledge the support of the Slovenian Research Agency through program grants No. P2-0098, No. and P2-0103, project grants No. J7-70265, No. GC-0001 and No. PR-12393. 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), Grant Agreement 101060712 (HE FishEUTrust), and Grant Agreement 101214398 (ELLIOT). We also acknowledge the support of the EC/EuroHPC JU and the Slovenian Ministry of HESI via the project SLAIF (grant number 101254461). We thank all participating teams for their contributions to FoodBench-QA 2026. We also thank the CL4Health workshop organizers for hosting the shared task. We acknowledge the participating teams based on the Codabench platform: Team 1 (rbls-lab), Team 2 (raviranjan), and Team 3 (andybox111).

11. Ethics Statement and Broader Impact

Intended Use and Societal Impact. The FoodBench-QA shared task aims to advance research in grounded food and nutrition question answering, with potential applications in dietary assessment, public health monitoring, and personalized nutrition support. By promoting standardized evaluation benchmarks and transparent comparison of methods, the task contributes to the development of reliable AI systems for food and health-related information processing. However, the presented systems are intended for research purposes and should not be used as substitutes for professional medical or nutritional advice.

Risks and Misuse. Automated nutrient estimation and food classification systems may produce inaccurate predictions, particularly in cases of incomplete ingredient information, ambiguous food descriptions, or unseen entities. Incorrect nutritional estimates or classifications could potentially mislead users if deployed without appropriate validation or human oversight. We therefore emphasize that outputs should be interpreted with caution and require expert verification in real-world health or clinical applications.

Data Sources and Privacy. The datasets used in the shared task are derived from publicly available resources, including recipe datasets, scientific abstracts, and food ontologies. These datasets do not contain personally identifiable information, and no personal or sensitive user data were collected or processed. The annotation and dataset construction processes followed standard research practices to ensure responsible data handling.

Bias and Representation. The datasets primarily reflect food terminology, recipes, and nutritional standards derived from specific cultural and regulatory contexts, including European Union nutrition regulations and English-language resources. As a result, the benchmark may not fully represent global dietary practices, regional cuisines, or multilingual food descriptions. Systems trained or evaluated on this benchmark may therefore exhibit reduced performance for underrepresented food concepts or cultural contexts.

Reproducibility and Transparency. To promote transparency and responsible research, all datasets, evaluation scripts, and submission results are publicly released through CodaBench and Zenodo. This enables independent validation, comparison of methods, and further research on trustworthy food and nutrition AI systems.

12. Limitations

The FoodBench-QA shared task has several limitations that should be considered when interpreting the results.

First, participation was limited to a small number of teams, and not all teams submitted results for every task or subtask. This restricts the diversity of methodological approaches and limits the generalizability of performance comparisons.

Second, the benchmark focuses on structured recipe data and curated text corpora, which may not fully capture the complexity and variability of real-world food descriptions, informal language, or multilingual settings.

Third, nutrient estimation is evaluated using tolerance thresholds defined by EU Regulation 1169/2011, which provide a regulatory evaluation framework but may not reflect all practical requirements for dietary assessment or clinical decision-making.

Fourth, the artificial named entity linking datasets simplify the problem by providing pre-identified entities, which do not fully reflect the challenges of end-to-end information extraction from natural text.

Finally, although the benchmark covers multiple ontologies, it remains limited to three semantic resources and does not account for the full diversity of food-related knowledge representations.

Future work should address these limitations by expanding dataset diversity, increasing participation, incorporating multilingual and culturally di-

verse food data, and evaluating systems in real-world deployment scenarios.

13. Bibliographical References

Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018. A sentiment-labelled corpus of hansard parliamentary debate speeches. In *International Language Resource and Evaluation Conference 2018: ParlaCLARIN Workshop*, pages 43–47. Clarin.

C Agostoni, Stefania Boccia, S Banni, PM Mannucci, and A Astrup. 2021. Sustainable and personalized nutrition: From earth health to public health. *European Journal of Internal Medicine*, 86:12–16.

Lorenzo Bairati. 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(1).

Kelvin Balcombe, Iain Fraser, and Salvatore Di Falco. 2010. Traffic lights and food choice: A choice experiment examining the relationship between nutritional food labels and price. *Food policy*, 35(3):211–220.

Lars Carlsen and Rainer Bruggemann. 2022. The 17 united nations' sustainable development goals: A status by 2020. *International Journal of Sustainable Development & World Ecology*, 29(3):219–229.

Gjorgjina Cenikj, Mauro Dragoni, Tome Eftimov, Barbara KOROU ŠI C SELJAK, Agnieszka Ławrynowicz, Fnu Mohbat, Oshani Seneviratne, Yoko Yamakata, and Mohammed J Zaki. 2025. Neurosymbolic methods for food computing. In *Handbook on Neurosymbolic AI and Knowledge Graphs*, pages 1019–1056. IOS Press.

Gjorgjina Cenikj, Gašper Petelin, Barbara Koroušić Seljak, and Tome Eftimov. 2022a. Scifoodner: Food named entity recognition for scientific text. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 4065–4073. IEEE.

Gjorgjina Cenikj, Gorjan Popovski, Riste Stojanov, Barbara Koroušić Seljak, and Tome Eftimov. 2020. Butter: Bidirectional lstm for food named-entity recognition. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3550–3556. IEEE.

Gjorgjina Cenikj, Eva Valenčič, Gordana Ispirova, Matevž Ogrinc, Riste Stojanov, Peter Korošec, Ermanno Cavalli, Barbara Koroušić Seljak, and Tome Eftimov. 2022b. Cafeteriasa corpus: scientific abstracts annotated across different food semantic resources. *Database*, 2022:baac107.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Damion M Dooley, Emma J Griffiths, Gurinder S Gosal, Pier L Buttigieg, Robert Hoehndorf, Matthew C Lange, Lynn M Schriml, Fiona SL Brinkman, and William WL Hsiao. 2018. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2(1):23.

Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488.

Teri E Emrich, Ying Qi, Wendy Y Lou, and Mary R L'Abbe. 2017. Traffic-light labels could reduce population intakes of calories, total fat, saturated fat, and sodium. *PloS one*, 12(2):e0171188.

Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2025. Foodsem: Large language model specialized in food named-entity linking. In *International Conference on Discovery Science*, pages 395–410. Springer.

Ana Gjorgjevikj, Matej Martinc, Gjorgjina Cenikj, Riste Stojanov, Jan Drole, Gordana Ispirova, Giulia Menichetti, Nives Ogrinc, Dimitar Trajanov, Sašo Džeroski, Barbara Koroušić Seljak, and Tome Eftimov. 2026. Large language models in food and nutrition science: Opportunities, challenges, and the case of foodyllm. *Current Research in Food Science*. In press (accepted February 13, 2026).

Ivana Gjorshoska, Tome Eftimov, and Dimitar Trajanov. 2022. Missing value imputation in food composition data with denoising autoencoders. *Journal of Food Composition and Analysis*, 112:104638.

Heather Greenfield and David AT Southgate. 2003. *Food composition data: production, management, and use*. Food & Agriculture Org.

Nicolas Guilpart, Toshichika Iizumi, and David Makowski. 2022. Data-driven projections suggest large opportunities to improve europe's soybean self-sufficiency under climate change. *Nature Food*, 3(4):255–265.

Gordana Ispirova, Gjorgjina Cenikj, Matevž Ogrinc, Eva Valenčič, Riste Stojanov, Peter Korošec, Ermanno Cavalli, Barbara Koroušić Seljak, and Tome Eftimov. 2022. Cafeteriafcd corpus: food consumption data annotated with regard to different food semantic resources. *Foods*, 11(17):2684.

Gordana Ispirova, Tome Eftimov, Sašo Džeroski, and Barbara Koroušić Seljak. 2024. Msgen: Measuring generalization of nutrient value prediction across different recipe datasets. *Expert Systems with Applications*, 237:121507.

Sonja Kunz, Simona Haasova, Jannik Rieß, and Arnd Florack. 2020. Beyond healthiness: the impact of traffic light labels on taste expectations and purchase intentions. *Foods*, 9(2):134.

Javier Marın, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2021. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.

Marian L Neuhouser, Ross L Prentice, Lesley F Tinker, and Johanna W Lampe. 2023. Enhancing capacity for food and nutrient intake assessment in population sciences research. *Annual review of public health*, 44(1):37–54.

THIRTEENTH Parliament. 2022. Hansard. *URL https://hansard. parliament. uk*.

Gorjan Popovski, Stefan Kochev, Barbara Korousic-Seljak, and Tome Eftimov. 2019a. Foodie: A rule-based named-entity recognition method for food information extraction. *ICPRAM*, 12:915.

Gorjan Popovski, Barbara Koroušić Seljak, and Tome Eftimov. 2019b. Foodbase corpus: a new resource of annotated food entities. *Database*, 2019:baz121.

Riste Stojanov, Gorjan Popovski, Gjorgjina Cenikj, Barbara Koroušić Seljak, and Tome Eftimov. 2021. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of Medical Internet Research*, 23(8):e28229.

Riste Stojanov, Gorjan Popovski, Nasi Jofce, Dimitar Trajanov, Barbara Koroušić Seljak, and Tome Eftimov. 2020. Foodviz: Visualization of food entities linked across different standards. In *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part II 6*, pages 28–38. Springer.