frontiers | Frontiers in Nutrition

TYPE Original Research
PUBLISHED 13 August 2024
DOI 10.3389/fnut.2024.1429259

OPEN ACCESS

EDITED BY
Tarun Belwal,
Texas A and M University, United States

REVIEWED BY
Frank Lo,
Imperial College London, United Kingdom
Yishuo Zhang,
Changchun University of Chinese Medicine, China
Jiaxin Li,
Changchun University of Chinese Medicine, China, in collaboration with reviewer YZ

*CORRESPONDENCE
Matevž Ogrinc
✉ matevz.ogrinc@ijs.si

RECEIVED 07 May 2024
ACCEPTED 26 July 2024
PUBLISHED 13 August 2024

CITATION
Ogrinc M, Koroušić Seljak B and Eftimov T (2024) Zero-shot evaluation of ChatGPT for food named-entity recognition and linking. *Front. Nutr.* 11:1429259. doi: 10.3389/fnut.2024.1429259

Zero-shot evaluation of ChatGPT for food named-entity recognition and linking

Matevž Ogrinc[1,2]*, Barbara Koroušić Seljak[2] and Tome Eftimov[2]

[1]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia, [2]Department of Computer Systems, Jožef Stefan Institute, Ljubljana, Slovenia

Introduction: Recognizing and extracting key information from textual data plays an important role in intelligent systems by maintaining up-to-date knowledge, reinforcing informed decision-making, question-answering, and more. It is especially apparent in the food domain, where critical information guides the decisions of nutritionists and clinicians. The information extraction process involves two natural language processing tasks named entity recognition—NER and named entity linking—NEL. With the emergence of large language models (LLMs), especially ChatGPT, many areas began incorporating its knowledge to reduce workloads or simplify tasks. In the field of food, however, we noticed an opportunity to involve ChatGPT in NER and NEL.

Methods: To assess ChatGPT's capabilities, we have evaluated its two versions, ChatGPT-3.5 and ChatGPT-4, focusing on their performance across both NER and NEL tasks, emphasizing food-related data. To benchmark our results in the food domain, we also investigated its capabilities in a more broadly investigated biomedical domain. By evaluating its zero-shot capabilities, we were able to ascertain the strengths and weaknesses of the two versions of ChatGPT.

Results: Despite being able to show promising results in NER compared to other models. When tasked with linking entities to their identifiers from semantic models ChatGPT's effectiveness falls drastically.

Discussion: While the integration of ChatGPT holds potential across various fields, it is crucial to approach its use with caution, particularly in relying on its responses for critical decisions in food and bio-medicine.

KEYWORDS

ChatGPT, food data, named-entity recognition, named-entity linking, natural language processing

1 Introduction

Food has always been an important factor in our daily lives. Food can influence our health, mental health, fitness, and other aspects in conjunction with a person's well-being (1), but to understand the intricate relationship between food and healthcare, one needs to dig deep into the vast amount of scientific literature. As such, extracting food information from literature is crucial in ensuring that dietary choices are informed by rigorous research, promoting an accurate understanding of nutritional principles. Evidence-based dietary recommendations from scientific studies empower individuals to make informed choices, fostering a healthier lifestyle supported by robust scientific evidence (2). Yet, manual evaluation of such literature is a daunting task. Additionally, in digital dietary assessment, information on dietary habits is supplied in plain, unstructured text, and by automating food information extraction, we can assist clinicians and dietitians in improving a person's lifestyle and health. Structuring food information from unstructured text sources such as digital dietary assessments, recipes, and scientific literature involves two critical tasks in

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.

FIGURE 1
Food NER example from a recipe text.

natural language processing (NLP): Named Entity Recognition (NER) (3) and named entity linking (4) (NEL). NER is a subtask of information extraction that automatically detects and categorizes entities (one or multiple words) from unstructured text. For instance, in Figure 1, an example of a recipe is presented, where the food entities are highlighted in bold and are automatically extracted by a NER method.

Depending on the methodology, several types of NER methods exist: dictionary-based, rule-based, corpus-based, active learning-based, and deep learning-based. Dictionary-based NERs are dependant on a pre-determined dictionary of the entities of interest (i.e., in our case, food entities) (5); rule-based NER also uses a pred-determined dictionary but in conjunction with rules that describe the characteristics of the entities in the domain of interest (6); corpus-based NERs are dependant on a corpus used to train a supervised machine learning model (7); active learning NERs use semi-supervised learning to train a model and further iteratively improve it using interactions from a user for new training instances (8), and deep learning-based NERs use large amounts of annotated data to train a model using deep neural networks (9). Despite many methodologies, the robustness and accuracy of a NER method is dependent on the amount of resources available for a specific domain.

NEL is the task of linking entities to their unique identifiers describing concepts in a knowledge base [i.e., in most cases, to a semantic model/ontology (10, 11)]. An ontology formally represents knowledge or concepts within a specific domain, detailing the entities, attributes, relationships, and constraints relevant to that domain. It serves as a structured framework for organizing and understanding information, facilitating knowledge sharing, reasoning, and interoperability between different systems and applications. Having a unique identifier helps us collect and comb information for the same entity from multiple sources (e.g., various scientific articles) even if it has a different textual representation (i.e. synonyms). This step is necessary to ensure that the data can interoperate effectively, which is crucial for adhering to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles (12). In Figure 2, a NEL example is presented, where the first discovered entity "cream cheese" from the NER example (see Figure 1) is linked to the SNOMED-CT (13) ontology.

With the emerging development of generative artificial intelligence (AI) (14), particularly large language models (LLMs) [e.g., ChatGPT (15) LLaMA (16) Mistral (17), Gemini (18)], they offer a lot of potential in diverse NLP tasks, including NER and NEL.

Our contribution: This article delves into a zero-shot evaluation of the capabilities of ChatGPT-3.5 and ChatGPT-4 in the tasks of food NER and NEL, which are crucial for synthesizing data from diverse unstructured sources like academic literature and text in lay language. We explore ChatGPT's abilities, especially its capacity to perform these tasks without prior training (i.e., zero-shot evaluation), by curating a generalized prompt with which ChatGPT is capable of performing both tasks. Our primary goal is the evaluation of the task of food NER for which we utilized two food corpora. Next, we evaluate it on the task of food NEL by linking the recognized food entities to a unique identifier from the SNOMED-CT ontology or the FOODON ontology (19). Further, we perform a secondary evaluation of both models on the task of NER on biomedical domain entities and the task of NEL by linking them to the NCBI (20) ontology and MeSH ontology to compare how resource availability and entity types influence ChatGPT's performance on NER and NEL tasks.

2 Related work

In the domain of food NER and NEL, a lot of work has been done in the last decade to address the lack of annotated resources and NLP methodologies. Before the introduction of the FoodBase annotated corpora (21), most research focused on rule-based NER methods such as drNER (22) for knowledge extraction of evidence-based dietary recommendations and FoodIE (23) which is a rule-based NER method for food information extraction from recipes. Additionally, StandFood (24) has introduced a classification approach focusing on the lexical similarity of food entities that can be used for NEL of food entities to the FoodEx2 database provided by the European Food Safety Authority (EFSA). Following the introduction of FoodBase, further research using ML techniques has been introduced. In 2020 (25) released BuTTER, the first bidirectional LSTM for food NER utilizing the FoodBase annotated corpus to identify food entities. In the following year, FoodNER (26) was released as a fine-tuned bidirectional encoder representation from transformers model for food NER and NEL which can extract and annotate food entities in five different tasks and distinguish food entities on the level of food groups. To visualize and help food experts with the subject of different food standards and interoperability, FoodViz (27) has been developed, which is a web-based framework used to visualize and annotate food entities with semantic tags. Recent studies have also shown results on using deep learning architecture for food NER from recipes (28) and enhancing NER in agriculture by using LLMs (29).

In contrast to the food domain, the biomedical domain has experienced significant advancements over the past two decades, attributed to the large amounts of resources available. This development is predominantly focused on various biomedical entities, including genotypes, phenotypes, diseases, treatments, and drugs. To further advance NLP in this domain, multiple workshops have been organized, notably BioNLP (30), BioCreative (31), i2b2 (32), etc, each emphasizing biomedical data and excluding food entities. Additionally, multiple ontologies have been developed for the biomedical domain, SNOMED-CT, MeSH (33), Disease

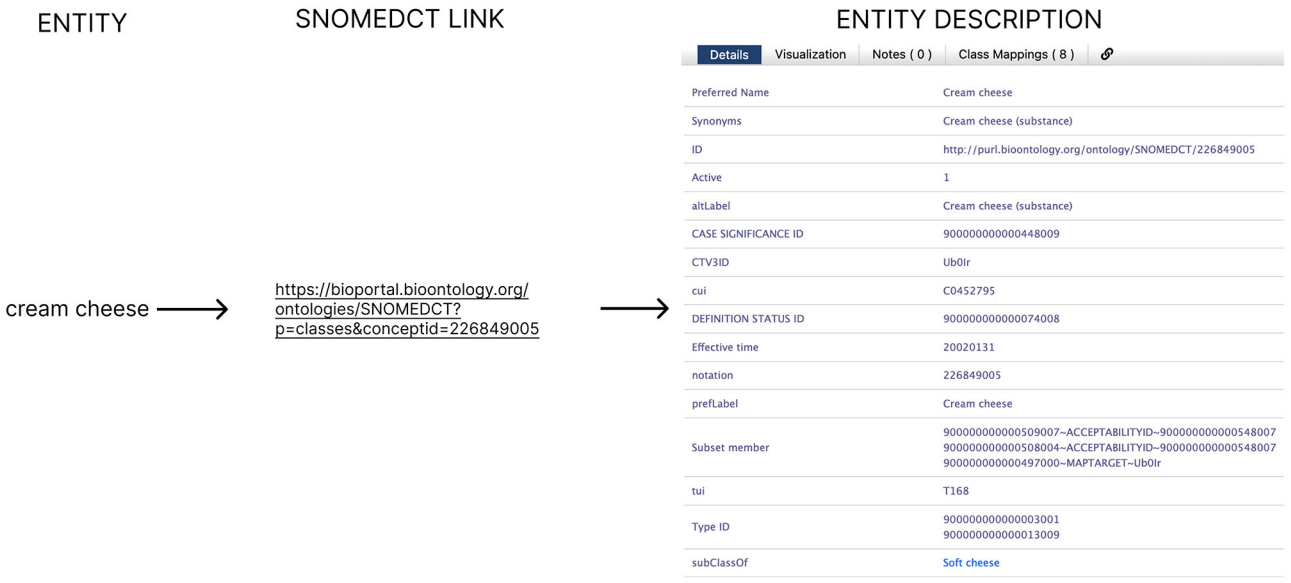

FIGURE 2
NEL example, where "cream cheese" is linked to the SNOMED-CT ontology.

ontology (34), UMLS (35), which facilitate the organization and classification of biomedical entities. These advancements in the field have influenced the evolution of ML models with the most recent iteration of Bert (36), BioBert (37) and BioClinicalBERT (38), three examples of successful models in the field of biomedical NLP. With the introduction of LLMs, such as ChatGPT, researchers began testing its capabilities in the biomedical domain. Studies are focusing on improving LLMs for clinical NER via prompt engineering (39) and fine-tuning ChatGPT on biomedical NLP tasks instead of its zero-shot evaluation (40, 41). In addition, it has been highlighted that ChatGPT is effective in similar clinical NER tasks (42), even in zero-shot settings, despite trailing behind specialized models like BioClinicalBERT.

3 Materials and methods

Our study focuses on evaluating ChatGPT's capabilities in NER and NEL across two pivotal domains, food and biomedical. To achieve this, we utilize specially curated datasets. For the food domain, we used gold standard datasets from the European Food Safety Authority-funded project, encompassing a wide range of food-related data from scientific articles and food consumption data. While for the biomedical domain, we used chemical, disease and species corpora.

3.1 Food NER datasets

Our evaluation of the food domain utilizes two corpora from an EFSA-funded project, CAFETERIA. The first corpus is the CafeteriaSA corpus (43), comprised of 500 scientific abstracts, each annotated with food entities leading to a total of 6,407 annotations. These annotations include entities' unique identifiers from various semantic resources, including the Hansard taxonomy (44), FoodOn, and SNOMED-CT terminology. This corpus lays the foundation for extracting and comprehending food information from scientific texts. The second corpus is the CafeteriaFCD corpus (45), which extends the FoodBase corpus, annotating food consumption data (i.e. recipes) with unique identifiers from external resources such as Hansard taxonomy, FoodOn ontology, SNOMED-CT terminology, and the FoodEx2 (46) classification system. The CafeteriaFCD corpora is comprised of 1,000 recipes, each annotated with food entities, leading to a total of 7,429 annotations.

3.2 Biomedical NER and NEL datasets

Our evaluation in the biomedical domain incorporates three distinct corpora from two sources. The BioCreative V challenge (47) and the Linnaeus gold standard corpus (48). The BioCreative V Challenge is the fifth iteration of the Critical Assessment of Information Extraction Systems in Biology challenge. The event evaluates text mining and information extraction systems applied to the biological domain. One of the sub-tasks in the fifth edition of the challenge has been the evaluation of NER methodologies in the context of chemical entities and disease entities within life science literature with each entity linked (NEL) to its MeSH identifier. These two corpora have been further involved in our experiments. In addition, the Linnaeus corpus serves as a gold standard for species entity recognition, offering a comprehensive collection of annotations for species names within biomedical research texts. In contrast to the BioCreative V Challenge, the Linnaeus corpus uses the NCBI identifiers for the NEL task.

3.3 GPT models

In the exploration of ChatGPT's NER and NEL capabilities across the food and biomedical domains, our study employs two advanced iterations of the Generative Pre-trained Transformer models: ChatGPT-3.5 (49) and ChatGPT-4 (50). Each model iteration brings unique strengths to our experimental setup, allowing for an understanding of the evolution and applicability of these AI technologies in handling domain-specific entity recognition and linking tasks.

ChatGPT-3.5 represents an intermediate advancement in OpenAI's lineup of language models. Notably, it has been instrumental in setting benchmarks for language comprehension, context understanding, and the generation of human-like text based on the vast knowledge it has been trained on. Its application in our study serves as an evaluation of its capabilities, especially in processing and analyzing complex domain-specific text, providing critical insights into the limitations and strengths of AI-driven NER and NEL in the context of scientific and food consumption literature and as a benchmark for improvements in subsequent models.

ChatGPT-4, the subsequent iteration, builds upon the foundational successes of its predecessors, offering enhanced understanding and generation capabilities that promise significant advancements in AI's role within NER and NEL tasks. With a broader knowledge base and improved contextual awareness, ChatGPT-4 is designed to surpass the limitations observed in earlier models, providing more accurate entity recognition and linking across the specialized datasets utilized in our study. The introduction of ChatGPT-4 into our experimental workflow allows for a direct comparison of performance metrics with its predecessor in highlighting the progress made in language modeling and its practical implications for food and biomedical NER and NEL tasks.

4 Results

In Figure 3, our experimental flowchart is presented and modeled based on the approach in (51). The design of the prompt plays an important role in receiving suitable responses from ChatGPT. As such, we have created a general prompt to serve as a NER and NEL tool while incorporating essential elements such as contextual background, a clear task directive, and a specific output format constraint. In the following Table 1, we see an example of our prompt for each domain, which we used to extract and link entities.

The responses generated from the prompts proved to be sufficiently detailed for subsequent post-processing and analysis. To assess the performance we utilized the F1 score, a well-established metric combining precision and recall. The F1 metric is used to show the reliability of a model by calculating a score between 0 and 1. A score higher than 0.9 indicates excellent performance. A score between 0.8 and 0.9 is considered good, while a score between 0.5 and 0.8 is average. An F1 score below 0.5 is considered poor performance. The F1 score calculation (Equation 1) is calculated using precision (Equation 2) and recall (Equation 3). TP means true positive or correctly found entities, FP means false positive or entities that have been found but are incorrect, and FN means false negative, entities that have not been found.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

4.1 Food NER and NEL evaluation

To enhance accuracy and minimize ChatGPT's generation of erroneous information in the food domain, we define our initial prompt to align with the specific requirements of the Cafeteria corpora. An example of the text used with the prompt is seen with the response received from ChatGPT-4 in Table 2.

After analyzing the responses, we noticed instances where ChatGPT identified food entities as partial matches, omitting prefixes or suffixes. For example, instead of recognizing PROVOLONE CHEESE in its entirety, ChatGPT identified only CHEESE. Initially categorized as false positives, these partial matches made us reconsider our evaluation criteria. Consequently, we explored whether acknowledging partial matches as correct could enhance the model's performance assessment, shifting from viewing them strictly as errors to potential positives.

4.2 NER

In the NER task, ChatGPT demonstrated good performance across both culinary recipes (CafeteriaFCD) and scientific articles (CafeteriaSA). Figure 4 illustrates the quantity of precisely identified food entities, with an entity deemed accurate only if identified with 100% correctness (not a partial match). The labels FOODON/SNOMED-CT indicate text documents annotated using FOODON or SNOMED-CT identifiers, whereas SA/FCD represents scientific articles or food consumption data. Observations reveal that ChatGPT-3.5 and ChatGPT-4 exhibit greater proficiency in identifying food entities within food consumption data (CafeteriaFCD) than scientific articles (CafeteriaSA).

Moreover, there is a marginally higher performance level in ChatGPT-4 relative to ChatGPT-3.5, which aligns with expectations considering its status as an enhanced model. However, the performance gap between the two versions is surprisingly narrow. Upon reviewing the F1 scores presented in Table 3, it becomes evident that ChatGPT-4 slightly outperforms ChatGPT-3.5 in the accuracy of food entity identification, particularly in culinary recipes over scientific texts. Furthermore, the disparity in performance between ChatGPT-3.5 and ChatGPT-4 is more evident in the context of scientific articles than in food recipes, underscoring the improvements in the latest model's capabilities.

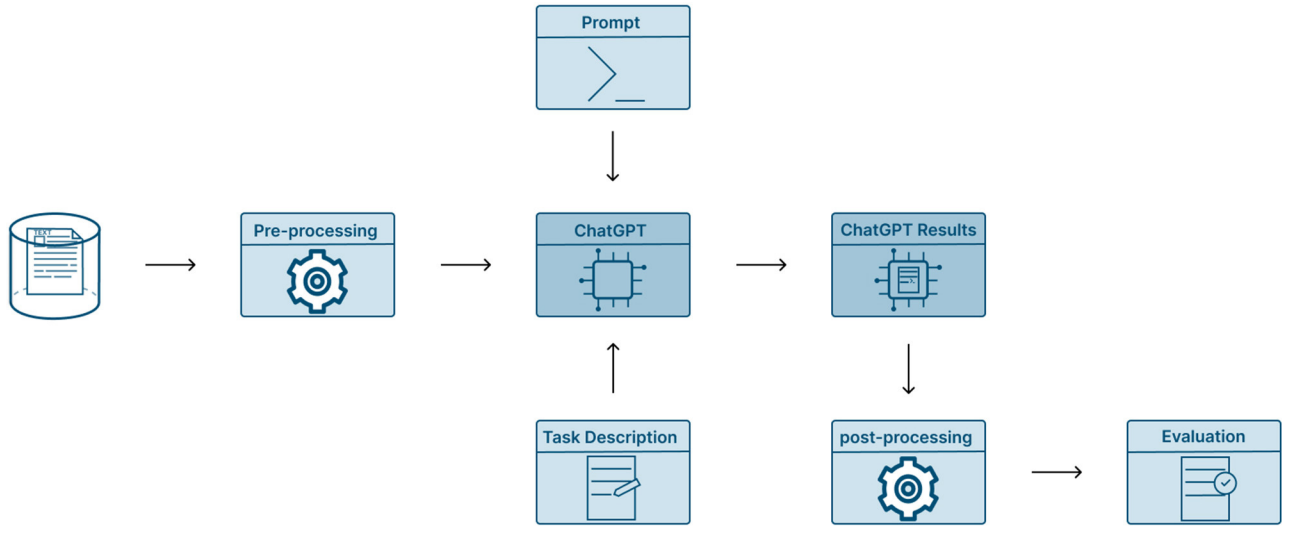

FIGURE 3
Pipeline for ChatGPT evaluation.

TABLE 1 Prompts used to evaluate ChatGPT-3.5 and ChatGPT-4 on different domains.

Domain	Prompt
Food domain	*Extract only food entities and find their correct FOODON/SNOMED-CT ids, display the result only in this format name_of_FOOD|FOODON_id/SNOMEDCT_id without headers and nothing else. Text is:*
Chemical domain	*Extract only chemical entities and find their correct MeSH ids, display the result only in this format name_of_Chemical|MeSH_id without headers and nothing else. Text is:*
Disease domain	*Extract only disease entities and find their correct MeSH ids, display the result only in this format name_of_Disease|MeSH_id without headers and nothing else. Text is:*
Species domain	*Extract only species entities and find their correct NCBI ids, display the result only in this format name_of_Species|NCBI_id without headers and nothing else. Text is:*

TABLE 2 ChatGPT-4 response for food entities.

Example text	"Mix the cream cheese, beef, olives, onion, and Worcestershire sauce together in a bowl until evenly blended."
Domain	ChatGPT-4 response
Food	`"cream cheese|762563006 beef|767623000 olives|722867003 onion|769846004 Worcestershire sauce|771471005"`

In comparison, if we add partial matches to our evaluation, the performance of both models rises, which illustrates that although ChatGPT demonstrates proficiency in food domain NER tasks, its inability to detect prefixes and suffixes, critical elements that significantly impact food identification, may lead to confusion when determining the precise type or brand of food. If we compare our findings with FoodNER and SciFoodNER (52), taking into consideration that both models are fine-tuned to different domain-specific data, FoodNER on food consumption data and SciFoodNER on scientific articles, we notice that both ChatGPT-3.5 and ChatGPT-4 fall behind. In comparison to FoodNER, the F1 score of ChatGPT-4 partials on FOODON FCD came the closest with a difference of 0.104, while in contrast to SciFoodNER, ChatGPT-4 partials on FOODON SA came the closest with a more significant difference of 0.201.

4.3 NEL

For the NEL task, we assessed the performance of both models in accurately associating food entities with their respective SNOMED-CT or FOODON identifiers. Each identifier is a unique code corresponding to a specific food entity. For example, CHEESE is represented with a FOODON identifier of 00001013 and with a SNOMED-CT identifier of 102264005. According to the results, the outcome fell short of our expectations. Apart from ChatGPT-4 successfully linking only two correct identifiers within the FOODON CafeteriaFCD corpus, neither model could associate any food entity with its corresponding identifier. Given these results, calculating the F1 score for this task was deemed unnecessary. Alternatively to ChatGPT's performance, FoodNER models achieved a macro F1 score between 0.733 and 0.789 on food consumption data, while SciFoodNER models achieved a median

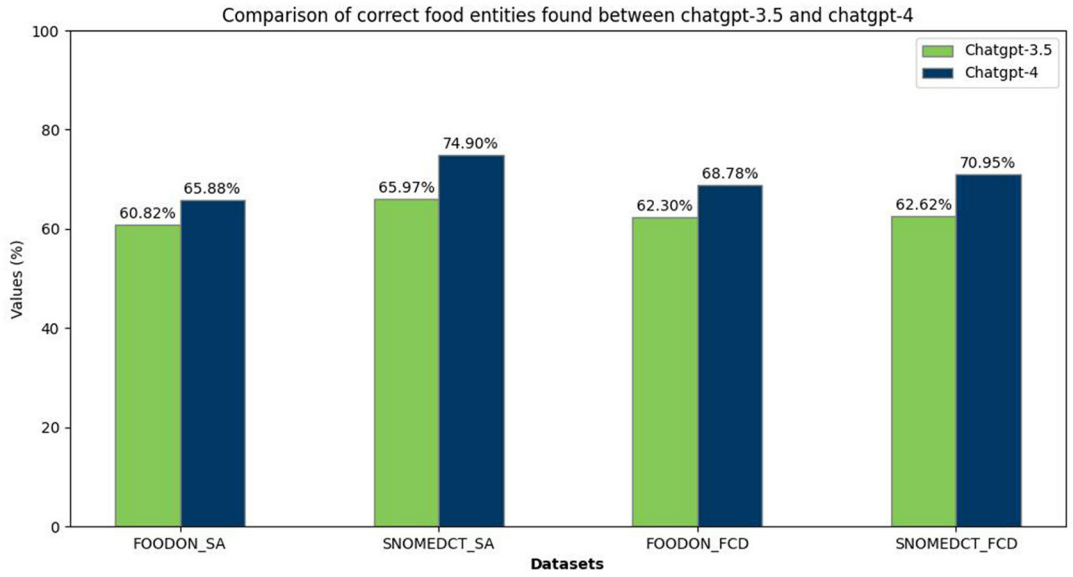

FIGURE 4
Comparison between ChatGPT-3.5 and ChatGPT-4 in finding correct food entities.

TABLE 3 F1 scores for ChatGPT-3.5 and ChatGPT-4 for food domain (NER task).

Model	FOODON SA	SNOMED-CT SA	FOODON FCD	SNOMED-CT FCD
ChatGPT-3.5	0.404	0.363	0.619	0.520
ChatGPT-3.5 partials	0.539	0.463	0.786	0.708
ChatGPT-4	0.533	0.490	0.668	0.562
ChatGPT-4 partials	0.699	0.604	0.839	0.749

macro F1 score of around 0.42. Illustrating the difficulty of NEL in the domain of food.

4.4 Biomedical NER and NEL evaluation

To compare our findings within the food domain, we have extended our analysis to a more established area of study, the biomedical domain. Our examination in this domain drew upon three previously mentioned corpora, two from the BioCreative V challenge and one from the Linnaeus corpus. We adapted our prompt for each corpus by modifying the original to suit the specific corpus focus. For the chemical and disease entities corpora, which use MeSH identifiers, the prompt has been tailored to extract the relevant entities and their corresponding MeSH IDs. For instance in Table 4, we see responses from ChatGPT-4 for the example text.

For the Linnaeus dataset, which catalogs species entities using NCBI identifiers, we used a prompt designed for extracting species entities and their accurate NCBI IDs. An example of the response and text is seen in Table 5.

4.5 NER

For the NER task, the performance outcomes of both ChatGPT-3.5 and ChatGPT-4 are depicted in Figure 5. This figure indicates that the models perform better in identifying chemical and disease entities while struggling with species entities. This disparity could be attributed to the narrower scope of chemical and disease terminology instead of the broad and varied taxonomy of species. The observation shows the challenge in species entity recognition, reflected in both models' performance metrics.

The F1 scores in Table 6 display the same pattern. The task of identifying chemical and disease entities proved to be easier, while recognizing species posed a more significant challenge. Adding to this, our findings from the food domain, the incorporation of partial matches significantly enhanced the performance of both models across the chemical and disease datasets and, to a lesser extent, for the species dataset. In addition, the results correspond to those reported in articles (51, 53) with a little margin of difference.

4.6 NEL

In our analysis of the NEL task within the biomedical domain, we have evaluated the performance of both ChatGPT-3.5 and ChatGPT-4 in mapping chemical, disease, and species entities to their respective MeSH or NCBI identifiers, paralleling our approach in the food domain. The outcomes, as depicted in Figure 6, revealed a marked superiority of ChatGPT-4 over ChatGPT-3.5. Furthermore, accurately linking disease and chemical entities proved challenging for both models, demonstrating the complexity

TABLE 4 ChatGPT-4 response for chemical and disease entities.

Example text	"Famotidine is a histamine H2-receptor antagonist used in inpatient settings for prevention of stress ulcers and is showing increasing popularity because of its low cost"
Domain	ChatGPT-4 response
Disease	`"stress ulcers|D004487"`
Chemical	`"Famotidine|D005242 Histamine H2-Receptor Antagonists|D006632 Stress Ulcers|D013379"`

TABLE 5 ChatGPT-4 response for species entities.

Example text	"including the ribosomal protein S3 from Escherichia coli (1), Mer1p from S.cerevisiae, a meiosis-specific splicing factor (1), MEX-3 from Caenorhabditis elegans, presumably involved in mRNA localization during development (2)"
Domain	ChatGPT-4 response
Species	`"Escherichia coli|562 Caenorhabditis elegans|6239"`

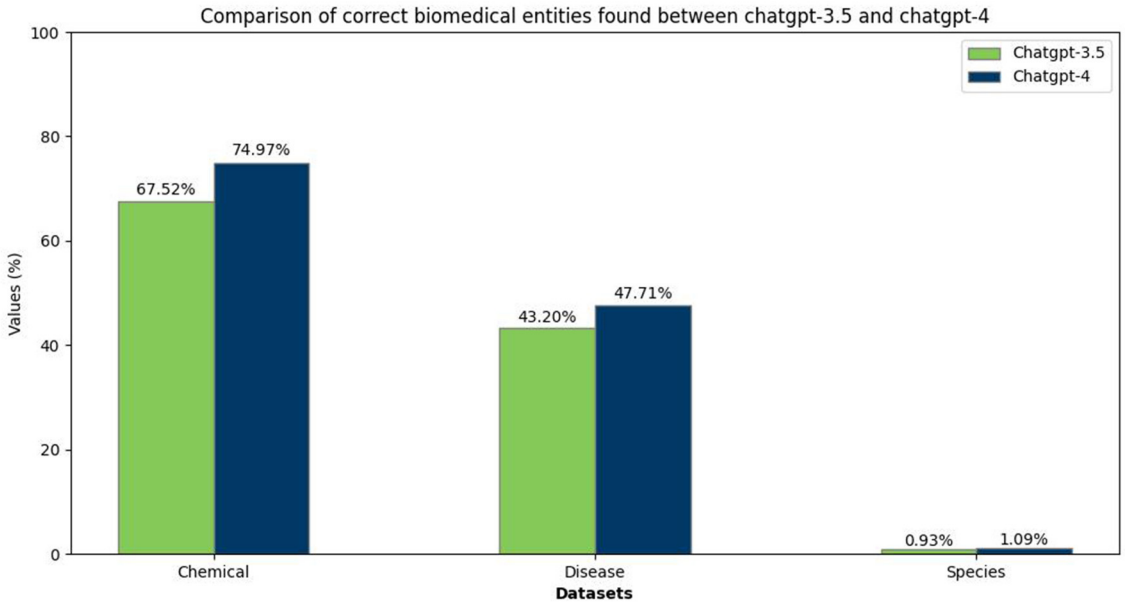

FIGURE 5
Comparison between ChatGPT-3.5 and ChatGPT-4 in finding biomedical entities.

of named entity linking despite the well-documented and recognized datasets.

The F1 scores in the accompanying Table 7 confirm ChatGPT-4 enhanced proficiency in linking entities across all categories compared to ChatGPT-3.5. Additionally, ChatGPT-4 exhibits a notable advancement in identifying chemical and disease entities. However, the overall performance, particularly in species entity linking, remains sub-optimal, highlighting areas for future model refinement and research focus.

TABLE 6 F1 scores for ChatGPT-3.5 and ChatGPT-4 for the biomedical domain (NER task).

Model	Chemical	Disease	Species
ChatGPT-3.5	0.578	0.404	0.016
ChatGPT-3.5 partials	0.646	0.535	0.021
ChatGPT-4	0.698	0.514	0.021
ChatGPT-4 partials	0.772	0.692	0.023

5 Discussion

We observe that in food NER, ChatGPT models perform comparably to FoodNER and BuTTER, which are fine-tuned on specific datasets. This suggests that GPT models can be applied to food NER tasks but incur financial costs. Conversely, developing specialized food NER models like FoodNER and BuTTER involves extensive manual data annotation, leading to a time-consuming process. In addition, ChatGPT models are not effective for food NEL without fine-tuning.

The tasks of NER and NEL are essential for making data interoperable, which is especially apparent in the food domain and is a crucial part of accomplishing FAIR principles. Additionally,

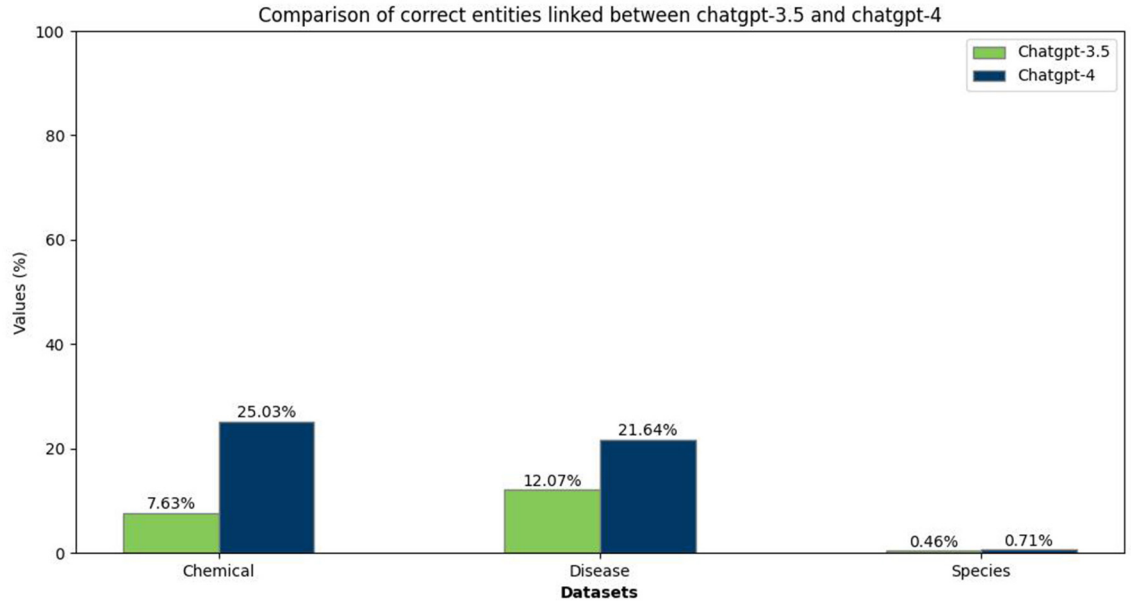

FIGURE 6
Comparison between ChatGPT-3.5 and ChatGPT-4 in linking biomedical entities to identifiers.

TABLE 7 F1 scores for ChatGPT-3.5 and ChatGPT-4 for biomedical domain (NEL task).

Model	Chemical	Disease	Species
ChatGPT-3.5	0.065	0.113	0.008
ChatGPT-4	0.233	0.233	0.0139

these tasks open up the ability to link food data to biomedical data, where data normalization is necessary and can be achieved by performing the tasks presented in this study. By facilitating the linkage of food data to biomedical data, these tasks empower professionals to make well-informed decisions regarding patient care, dietary recommendations, and overall health management. The normalization of data achieved through NER and NEL offers valuable insights that can directly impact clinical and nutritional assessments, fostering more precise and personalized interventions. While ChatGPT-4 yields promising results in the task of NER, its shortcomings in the task of NEL leave much room for improvement before becoming a reliable method for accomplishing both tasks.

6 Conclusion

In our study, we evaluated the performance of ChatGPT versions 3.5 and 4 on NER and NEL tasks within food and biomedical domains. After testing multiple prompt designs, we found a general prompt that was the most effective and least costly in retrieving results from both ChatGPT-3.5 and ChatGPT-4. ChatGPT-4 showed a slight edge over ChatGPT-3.5, especially in identifying entities in food consumption data (FCD) versus scientific articles (SA). Incorporating partial matches significantly improved both models' performance, suggesting a refined approach to entity recognition. Unfortunately, the performance of ChatGPT on NEL in the food domain highlights ChatGPT's lack of information for this particular task. In the biomedical domain, similar performance trends were observed, with ChatGPT-4 outperforming ChatGPT-3.5 in mapping entities to MeSH or NCBI identifiers. Despite yielding better results in linking disease and chemical entities, NEL proves a difficult challenge for both models. Additionally, ChatGPT's performance in species entity recognition from the Linnaeus dataset aligns closer to the results from the food domain. The comparison to specialized models discussed in related literature for the biomedical domain (54) and for the food domain (25) indicates specialized models' superiority in specific NER tasks. Yet, compared to models not trained with the same dataset or out-of-corpus (OOC), ChatGPT's performance aligns more closely, showcasing its potential adaptability (i.e., fine-tuning), which should be considered for future work.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

MO: Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft.

BK: Conceptualization, Writing – review & editing. TE: Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research work was financially supported by the Slovenian Research Agency under programme P2-0098, the European Union's Horizon 2020 research and innovation programme [grant agreement 101005259] (COMFOCUS), and FoodMarketMap project selected as an innovator within the FOODITY project that has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101086105.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

1. Mozaffarian D, Aspry KE, Garfield K, Kris-Etherton P, Seligman H, Velarde GP, et al. "Food Is Medicine" strategies for nutrition security and cardiometabolic health equity. *J Am Coll Cardiol.* (2024) 83:843–64. doi: 10.1016/j.jacc.2023.12.023

2. Tanna N. The impact of dietary guidelines for americans on dietary intake and obesity rates; 2024. In: *Copyright - Database Copyright ProQuest LLC; ProQuest Does Not Claim Copyright in the Individual Underlying Works*. Available at: https://www.proquest.com/dissertations-theses/impact-dietary-guidelines-americans-on-intake/docview/2919908020/se-2 (accessed February 12, 2024).

3. Nadeau D, Sekine S. A survey of named entity recognition and classification. *Lingvisticae Investigationes.* (2007) 30:3–26. doi: 10.1075/li.30.1.03nad

4. Shen W, Wang J, Han J. Entity linking with a knowledge base: issues, techniques, and solutions. *IEEE Trans Knowl Data Eng.* (2015) 27:443–60. doi: 10.1109/TKDE.2014.2327028

5. Zhou X, Zhang X, Hu X. MaxMatcher: Biological concept extraction using approximate dictionary lookup. In: *Pacific RIM International Conference on Artificial Intelligence*. Cham: Springer (2006). p. 1145–1149.

6. Soomro PD, Kumar S, Shaikh AA, Raj H, Raj H. Bio-NER: biomedical named entity recognition using rule-based and statistical learners. *Int J Adv Comp Sci Appl.* (2017) 8:20. doi: 10.14569/IJACSA.2017.081220

7. Kim JD, Tsujii J. Corpus-based approach to biological entity recognition. In: *Text data mining SIG (ISMB2002)*. Oxford: Oxford University Press (2002).

8. Shen Y, Yun H, Lipton ZC, Kronrod Y, Anandkumar A. Deep active learning for named entity recognition. *arXiv* [preprint] arXiv:170705928. (2017). doi: 10.18653/v1/W17-2630

9. Li J, Sun A, Han J, Li C. A Survey on deep learning for named entity recognition. *IEEE Trans Knowl Data Eng.* (2022) 34:50–70. doi: 10.1109/TKDE.2020.2981314

10. Peckham J, Maryanski F. Semantic data models. *ACM Comput Surv.* (1988) 20:153–189. doi: 10.1145/62061.62062

11. Maedche A, Staab S. Ontology learning for the semantic web. *IEEE Intell Syst.* (2001) 16:72–9. doi: 10.1109/5254.920602

12. Guizzardi G. Ontology, ontologies and the I of fair. *Data Intellig.* (2020) 2:181–91. doi: 10.1162/dint_a_00040

13. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth Studies in health technology and informatics. *Stud Health Technol Inform.* (2006) 121:279.

14. Byrne MD. Generative artificial intelligence and ChatGPT. *J PeriAnesthesia Nurs.* (2023) 38:519–22. doi: 10.1016/j.jopan.2023.04.001

15. Post OB. *Introducing ChatGPT*. (2022). Available at: https://openai.com/blog/chatgpt (accessed May 10, 2023).

16. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. *arXiv* [preprint] arXiv:230213971. (2023). doi: 10.48550/arXiv.2302.13971

17. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas Ddl, et al. Mistral 7B. *arXiv* [preprint] arXiv:231006825. (2023). doi: 10.48550/arXiv.2310.06825

18. Pichai S. Introducing Gemini: our largest and most capable AI model. In: *Google*. (2023). Available at: https://blog.google/technology/ai/google-gemini-ai/#sundar-note (accessed April 6, 2024).

19. Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, Lange MC, et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Sci Food.* (2018) 2:23. doi: 10.1038/s41538-018-0032-6

20. Federhen S. The NCBI taxonomy database. *Nucleic Acids Res.* (2012) 40:D136–43. doi: 10.1093/nar/gkr1178

21. Popovski G, Seljak BK, Eftimov T. FoodBase corpus: a new resource of annotated food entities. *Database.* (2019) 2019:baz121. doi: 10.1093/database/baz121

22. Eftimov T, Koroušić Seljak B, Korošec P. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS ONE.* (2017) 12:e0179488. doi: 10.1371/journal.pone.0179488

23. Popovski G, Kochev S, Seljak BK, Eftimov T. FoodIE: a rule-based named-entity recognition method for food information extraction. In: *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - ICPRAM*. Setbal: SciTePress (2019). p. 915–922.

24. Eftimov T, Koroec P, Korou Seljak B. StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. *Nutrients.* (2017) 9:542. doi: 10.3390/nu9060542

25. Cenikj G, Popovski G, Stojanov R, Seljak BK, Eftimov T. BuTTER: BidirecTional LSTM for food named-entity recognition. In: *2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA: IEEE (2020). p. 3550–3556.

26. Stojanov R, Popovski G, Cenikj G. Koroušić Seljak B, Eftimov T. A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *J Med Intern Res.* (2021) 23:e28229. doi: 10.2196/28229

27. Stojanov R, Popovski G, Jofce N, Trajanov D, Seljak BK, Eftimov T. FoodViz: visualization of food entities linked across different standards. In: Nicosia G, Ojha V, La Malfa E, Jansen G, Sciacca V, Pardalos P, et al., editors. *Machine Learning, Optimization, and Data Science*. Cham: Springer International Publishing (2020). p. 28–38. doi: 10.1007/978-3-030-64580-9_4

28. Agarwal A, Kapuriya J, Agrawal S, Konam AV, Goel M, Gupta R, et al. Deep learning based named entity recognition models for recipes. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Paris: ELRA Language Resources Association (2024). p. 4542–4554.

29. Chebbi A, Kniesel G, Abdennadher N, Dimarzo G. Enhancing named entity recognition for agricultural commodity monitoring with large language models. In: *Proceedings of the 4th Workshop on Machine Learning and Systems*. New York, NY: Association for Computing Machinery (2024). p. 208–213.

30. Demner-fushman D, Ananiadou S, Cohen K. *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Toronto, Canada: Association for Computational Linguistics (2023). Available at: https://aclanthology.org/2023.bionlp-1.0 (accessed April 2, 2024).

31. Arighi C, Krallinger M, Leitner F. *Biocreative*. (2023). Available at: https://biocreative.bioinformatics.udel.edu/ (accessed April 2, 2024).

32. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc*. (2013) 20:806–13. doi: 10.1136/amiajnl-2013-001628

33. Rogers FB. *Medical Subject Headings (MESH). Mixed Function Oxigenases D*. Chicago, IL: Medical Library Association (2006). p. 8.

34. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. (2019) 47:D955–62. doi: 10.1093/nar/gky1032

35. Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. *Yearb Med Inform*. (1993) 2:41–51. doi: 10.1055/s-0038-1637976

36. Özdil U, Arslan B, Taşar DE, Polat G, Ozan Ş. Ad text classification with bidirectional encoder representations. In: *2021 6th International Conference on Computer Science and Engineering (UBMK)*. Ankara: IEEE (2021). p. 169–173. doi: 10.1109/UBMK52708.2021.9558966

37. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. (2020) 36:1234–40. doi: 10.1093/bioinformatics/btz682

38. Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis: Association for Computational Linguistics (2019). p. 72–78. Available at: https://www.aclweb.org/anthology/W19-1909 (accessed April 4, 2024).

39. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. (2024) 2024:ocad259. doi: 10.1093/jamia/ocad259

40. Bousselham H, Nfaoul EH, Mourhir A. Fine-tuning GPT on biomedical NLP tasks: an empirical evaluation. In: *2024 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*. Kolkata: IEEE (2024). p. 1–6.

41. García-Barragán Á, González Calatayud A, Solarte-Pabón O, Provencio M, Menasalvas E, Robles V. GPT for medical entity recognition in Spanish. In: *Multimedia Tools and Applications*. Springer Nature (2024) p. 1–20.

42. Hu Y, Ameer I, Zuo X, Peng X, Zhou Y, Li Z, et al. Improving large language models for clinical named entity recognition via prompt engineering. *arXiv* [preprint] arXiv:230316416. (2023). doi: 10.48550/arXiv.2303.16416

43. Cenikj G, Valenčič E, Ispirova G, Ogrinc M, Stojanov R, Korošec P, et al. CafeteriaSA corpus: scientific abstracts annotated across different food semantic resources. *Database*. (2022) 2022:baac107. doi: 10.1093/database/baac107

44. Alexander M, Davies M, Dallachy F. *The Hansard Corpus 1803-2005*. The University of Glasgow, Scotland (2015).

45. Ispirova G, Cenikj G, Ogrinc M. Valenčič E, Stojanov R, Korošec P, et al. Cafeteriafcd corpus: food consumption data annotated with regard to different food semantic resources. *Foods*. (2022) 11:2684. doi: 10.3390/foods11172684

46. EFSA. *The Food Classification and Description System FoodEx 2 (revision 2)*. Hoboken, NJ: Wiley Online Library. (2015).

47. Krallinger M, Pérez-Pérez M, Pérez-Rodríguez G, Blanco-Míguez A, Fdez-Riverola F, Capella-Gutierrez S, et al. *The Biocreative v. 5 Evaluation Workshop: Tasks, Organization, Sessions and Topics*. (2017).

48. Gerner M, Nenadic G, Bergman CM, LINNAEUS. a species name identification system for biomedical literature. *BMC Bioinformatics*. (2010) 11:1–17. doi: 10.1186/1471-2105-11-85

49. OpenAI. *OpenAI's GPT-3.5 Turbo*. (2023). Available at: https://openai.com/ (accessed February 7, 2024).

50. Achiam J, Adler S, Agarwal S, Ahmad L, et al. *GPT-4 Technical Report*. (2023).

51. Chen Q, Sun H, Liu H, Jiang Y, Ran T, Jin X, et al. A Comprehensive Benchmark Study on Biomedical Text Generation and Mining with ChatGPT. *bioRxiv*. (2023). Available at: https://www.biorxiv.org/content/early/2023/04/20/2023.04.19.537463 (accessed April 25, 2024).

52. Cenikj G, Petelin G, Korou Seljak B, Eftimov T. SciFoodNER: food named entity recognition for scientific text. In: *2022 IEEE International Conference on Big Data (Big Data)*. Osaka: IEEE (2022). p. 4065–4073.

53. Tian S, Jin Q, Yeganova L, Lai PT, Zhu Q, Chen X, et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Brief Bioinformat*. (2024) 25:bbad493. doi: 10.1093/bib/bbad493

54. Giorgi JM, Bader GD. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*. (2020) 36:280–6. doi: 10.1093/bioinformatics/btz504