A Domain-Specific Curated Benchmark for Entity and Document-Level Relation Extraction

Marco Martinelli[1,*] **Stefano Marchesin**[1] **Vanessa Bonato**[2] **Giorgio Maria Di Nunzio**[1]
Nicola Ferro[1] **Ornella Irrera**[1] **Laura Menotti**[1] **Federica Vezzani**[2] **Gianmaria Silvello**[1]
[1]Department of Information Engineering, University of Padova
[2]Department of Linguistic and Literary Studies, University of Padova
[*] Corresponding author: martinell2@dei.unipd.it

Abstract

Information Extraction (IE), encompassing Named Entity Recognition (NER), Named Entity Linking (NEL), and Relation Extraction (RE), is critical for transforming the rapidly growing volume of scientific publications into structured, actionable knowledge. This need is especially evident in fast-evolving biomedical fields such as the gut-brain axis, where research investigates complex interactions between the gut microbiota and brain-related disorders. Existing biomedical IE benchmarks, however, are often narrow in scope and rely heavily on distantly supervised or automatically generated annotations, limiting their utility for advancing robust IE methods. We introduce GutBrainIE, a benchmark based on more than 1,600 PubMed abstracts, manually annotated by biomedical and terminological experts with fine-grained entities, concept-level links, and relations. While grounded in the gut-brain axis, the benchmark's rich schema, multiple tasks, and combination of highly curated and weakly supervised data make it broadly applicable to the development and evaluation of biomedical IE systems across domains.[1]

1 Introduction

Recent studies increasingly associate gut microbiota with neurological and psychiatric disorders like Parkinson's, Alzheimer's, Multiple Sclerosis, and mood disorders (Carabotti et al., 2015; Ghaisas et al., 2016; Appleton, 2018; Cryan et al., 2020). PubMed publications on the gut-brain axis more than doubled from 2020 to 2025, increasing from 600 to over 1,500 articles annually. This rapid growth challenges clinicians and researchers to stay updated, identify, and interpret findings in unstructured texts. Natural Language Processing (NLP) methods for Information Extraction (IE) are crucial in tackling this problem by systematically identifying entities, mapping them to structured knowledge bases, and extracting semantic relations. Named Entity Recognition (NER) detects and categorizes mentions of biomedical entities such as diseases, chemicals, or anatomical structures. Named Entity Linking (NEL) disambiguates these mentions by linking them to unique identifiers in external knowledge bases, ensuring semantic consistency across texts. Relation Extraction (RE) identifies and classifies relationships between entities, such as interactions or causal associations. Within RE, two complementary evaluation settings are commonly distinguished. Mention-level RE (M-RE) targets specific pairs of mentions in text, making it sensitive to lexical variation and mention boundaries. Concept-level RE (C-RE) instead operates at the level of linked concepts corresponding to the standard setting adopted by most benchmarks (Detroja et al., 2023).

Despite their importance, implementing these tasks in practice requires substantial amounts of manual annotation, particularly at fine levels of granularity. As a result, existing biomedical corpora are often narrow in scope, restricted to a limited set of entity types or relation predicates, confined to sentence-level annotations, and reliant on distant supervision (Chen et al., 2015; Karp, 2016; Wang et al., 2022). These limitations hinder the development of IE systems capable of handling domains with specialized terminology, diverse concept spaces, and cross-sentence dependencies (Tho et al., 2024; Park et al., 2024). The gut-brain axis literature exacerbates these challenges, combining complex terminology, broad conceptual diversity, and long-range relations (Liu et al., 2021; Hong et al., 2025). To deal with these challenges, we introduce GutBrainIE, a new benchmark for biomedical IE based on more than 1,600 PubMed documents annotated by biomedical and terminological experts, trained laypersons, and distantly supervised methods (Su et al., 2019). It supports the four complementary IE tasks outlined above and is

[1] This work was accepted to the Findings of EACL 2026.

designed both to capture the unique challenges of the gut-brain axis and to serve as a general resource for advancing biomedical IE.

The main contributions of this work are:
(1) The first domain-specific corpus focused on the gut-brain axis with a size consistent with established biomedical corpora (cf. Table 2).
(2) Manual and automatic annotations based on the most comprehensive and fine-grained schema to date in biomedical IE, organized in a stratified structure reflecting different levels of expertise and quality.
(3) Standardized benchmark tasks with competitive baselines, evaluation scripts, and leaderboards, facilitating reproducible comparisons across systems.

The benchmark has been validated through internal experiments with a baseline system and through a large open public evaluation campaign where GutBrainIE served as the reference dataset. The dataset is publicly accessible, with training and development sets released with annotations, and the test set provided without ground-truth labels.[2] For comparability and reproducibility, each task is hosted on Codabench with official evaluation scripts and leaderboards, allowing system predictions on the test set to be evaluated against the hidden ground truth.[3] The rest of the paper is organized as follows: Section 2 details data collection and curation. Section 3 analyzes the GutBrainIE corpus. Section 4 introduces the benchmark tasks and discusses its applications. Section 5 reviews related corpora. Section 6 concludes and outlines future work.

2 Data Collection

The GutBrainIE benchmark features a large-scale biomedical corpus with manual and automatic annotations for entities, concept-level links, and relations across PubMed documents. It covers 13 entity types, including both widely used biomedical categories (e.g., *anatomical location*, *bacteria*, *drug*) and entities specific to the gut-brain axis (e.g., *microbiome* and *dietary supplement*). To account for the frequent occurrence of experimental scenarios in documents, we also introduced specific categories related to medical experiments (e.g., *biomedical technique* and *statistical technique*).

For what concerns RE, GutBrainIE features 17 relation predicates, many of which are overloaded, meaning that the same predicate can link different combinations of entity types depending on context. For example, the predicate *administered* can connect a *chemical*, *drug*, *dietary supplement*, or *food* to either a *human* or *animal* entity. Similarly, the same pair of entity types can be linked through multiple predicates. For instance, a *chemical* can be linked to a *microbiome* entity through either *impact* or *produced by*. This many-to-many design originates 55 unique relation triples. For NEL, GutBrainIE links annotated mentions to 6 standardized biomedical vocabularies and a custom-defined ontology for unmatched mentions.

Given the specialized nature of the gut-brain axis, we relied on an in-house team rather than external crowdworkers. This enabled targeted training, regular feedback, and higher annotation consistency. The team comprised 40 trained master's students in terminography serving as lay annotators and 7 experts, including computer scientists, terminologists specialized in the medical domain, and biomedical specialists with prior experience in evaluation campaigns.

The manual curation of GutBrainIE followed a four-stage workflow to ensure annotation quality and consistency (illustrated in Figure 2). The preparation involved document retrieval from PubMed and the design of the annotation schema and guidelines. The first annotation phase combined NER pre-annotations with expert curation. The second annotation phase refined NER with contributions from both experts and trained lay annotators. Finally, the NEL phase mapped the set of expert-annotated entity mentions to standardized biomedical vocabularies.

The **Preparation Phase** began with the retrieval of domain-specific documents from PubMed using two targeted queries, identified by external biomedical experts: `"gut microbiota" AND "Parkinson"` and `"gut microbiota" AND "mental health"`. Two retrieval rounds were conducted on May 9th and October 31st, 2024, resulting in 1,662 documents. After filtering out duplicates and low-relevance documents from earlier publication years (2013-2020), the final collection included 1,647 unique documents.

Following retrieval, we employed an iterative, brainstorming approach for defining the annotation

[2] The full benchmark, including code and datasets, is publicly available at: https://zenodo.org/records/16845409

[3] https://github.com/MMartinelli-hub/GutBrainIE/blob/main/TestData/codabench_competitions.md

schema and guidelines. Initially, a representative subset of 100 PubMed abstracts was selected and carefully analyzed by a focus group of expert annotators in collaboration with biomedical domain experts and terminologists, leading to the identification of a core set of domain-specific definitions, based on which we drafted a first annotation schema, defining the entities and relations of interest. This schema was further refined by extending the analysis to the full set of retrieved documents, resulting in a finalized structure comprising 13 entity types and 17 fine-grained relation predicates (see Figure 3 and Tables 6-7 in the appendix).

After these stages, the expert annotators' team defined a detailed set of annotation guidelines, with the final goal of obtaining high-quality annotations that are consistent through different annotators and documents. Inspired by prior works such as BioRED (Luo et al., 2022), BC5CDR (Li et al., 2016), and BioASQ-QA (Krithara et al., 2023), these guidelines detail the end-to-end annotation process to be followed for each document, including labeling rules, edge case handling, and examples of typical annotations and mistakes.[4] Before starting the manual annotation phases, a hands-on training session was conducted with all expert annotators, during which they jointly annotated a set of abstracts for both entity mentions and relations while reviewing and refining the guidelines to ensure consistent interpretation and resolve any ambiguities early on. Two additional sessions were conducted with biomedical experts, following the same process and providing domain-specific feedback for further guidelines adjustments.

First Annotation Phase. Before starting the actual curation, we adopted an automatic annotation support strategy for entity mentions to reduce the manual effort required from annotators and accelerate the annotation process (Ganchev et al., 2007; Greinacher and Horn, 2018; Mikulová et al., 2023). We began by selecting a representative sample of ten documents and asked two terminologists to manually annotate all entity mentions. We then selected a competitive zero-shot NER model – GLiNER (Zaratiana et al., 2024) – and systematically experimented with different pre-trained checkpoints, temperature values, and post-processing strategies, comparing the model's predicted mentions with the manually annotated ones. Based on these experiments, we selected the *NuNER Zero* checkpoint and adopted a temperature value of 0.8 to obtain a higher recall and minimize the number of entity mentions annotated by terminologists but missed by the model (Bogdanov et al., 2024). To preliminarily assess the impact of these pre-annotations, we asked the same terminologists to annotate ten additional documents using the GLiNER-generated pre-annotations. Without pre-annotations, they needed around 30 minutes per document to fully annotate its entity mentions. With pre-annotations, the time dropped to 10-15 minutes, depending on the document's complexity.

All pre-annotated documents were annotated with *MetaTron*, a free and publicly available web-based annotation platform (Irrera et al., 2024). In this first annotation phase, only expert annotators were involved. Each of them received 20 unique documents along with a shared set of 5 "honeypot documents". The latter were introduced to compute Inter-Annotator Agreement (IAA), enabling direct comparison of annotations across annotators and the assessment of their agreement and consistency. To avoid bias or information leakage, we instructed annotators not to consult with each other until all their annotations were completed (Alonso and Marchionini, 2019).

In computing IAA, two annotations agree only if their text spans, labels (and predicates, in the case of relations) exactly matched. Using Fleiss' κ (Fleiss, 1971) and Mean Pairwise Cohen's κ (Cohen, 1960) as metrics, we observed strong agreement for NER (0.89 and 0.88, respectively), aligned with previous biomedical datasets such as BioRED (Luo et al., 2022) and NLM-Gene (Islamaj et al., 2021). Agreement on RE was lower (0.43 for both metrics) but consistent with prior works involving complex relation annotation, reflecting the semantic difficulty of the gut-brain axis domain (Kim et al., 2013).

To complement the IAA analysis, a subgroup of two expert annotators, selected for domain expertise and prior annotation experience, manually revised all the documents annotated during this phase. This step led to the correction or removal of 204 entities and 135 relations. For entities, roughly one-third of edits were due to incorrect text spans, another third to mislabeling, and the rest to overly generic mentions that violated annotation guidelines. For relations, about one-third were revised for incorrect directionality, another third for wrong

[4] The annotation guidelines are available at: https://zenodo.org/records/16845409/files/GutBrainIE_2025_Annotation_Guidelines.pdf

predicate assignment, and the remainder for insufficient textual support. This review also allowed for the identification of issues in the annotation guidelines, which were refined accordingly. All annotations were then retroactively updated to align with the updated guidelines.

Second Annotation Phase. Prior to the second phase, we fine-tuned GLiNER on the final revised annotations from the first phase to improve the quality of pre-annotations.

In this second phase, each expert annotator labeled 40 new documents, while layperson annotators were introduced and assigned 24 documents each, divided into four batches. In each batch, one honeypot document previously annotated by experts was randomly inserted. Laypeople were required to annotate a minimum of one full batch, ensuring that each of them was labeling at least one honeypot document.

To assess the quality of layperson annotations, we computed Cohen's κ on their annotated honeypot documents against the reference version curated by experts (Cohen, 1960). Agreement for NER was moderate and acceptable (0.50), indicating that with adequate training and supervision non-expert annotators can contribute reliable entity annotations. In contrast, RE agreement was low (0.17), largely due to the annotation of relations against the guidelines.

At the end of this phase, expert annotators conducted a final review meeting to evaluate the second batch of annotated documents, addressing any unresolved issues and making corrections to ensure consistency across the full collection.

Named Entity Linking Phase. We performed entity linking on expert annotations only, as laypeople and automatic mentions often lack the precision required for reliable concept mapping.

We applied normalization based on heuristics derived from manual analysis of expert-curated annotations. This involved removing HTML tags and special symbols, using regular-expression substitutions to handle terminological variants and expand common acronyms into their full forms. Then, we implemented a three-stage linking approach. For each annotated mention at first we attempted exact string matching against the reference biomedical vocabularies defined for the entity's label. Resources were queried sequentially in a predefined priority order for each entity type, aiming to prioritize linkages to the largest and most established biomedical vocabularies. If no exact match was found, we computed the embeddings with BiomedBERT and then calculated the cosine similarity between the mention and all possible candidates in the reference vocabularies (Gu et al., 2021). Only candidates with similarity scores above a dynamic cut-off were retained, computed using the central limit theorem to approximate the distribution of similarity scores as normal and ensure statistical reliability (Kwak and Kim, 2017). If both previous steps failed, we assumed the mention was not in the reference vocabularies. In such cases, we first attempted to map the mention to an existing entry in our custom ontology. If no match was found, we added a new concept to our ontology by prompting an LLM (`LLaMA3-8B-8192` in this case) with the sentence containing the mention (highlighted with inline markers) to generate a pertinent definition (see Figure 5 for an example). In all stages, if multiple candidate links were returned, we manually selected the most appropriate one. Finally, all generated linkages and individuals created in our custom ontology were manually reviewed and verified by experts to ensure consistency and resolve ambiguous mappings. The accuracy of the NEL annotations was estimated at 0.915 ± 0.0473 using a sampling-based evaluation framework with statistical guarantees (Martinelli et al., 2026).

Automatically-Annotated Data Collection. After finalizing the manual annotations, we automatically annotated the remaining unlabelled documents from the original retrieval. For NER, we fine-tuned again the GLiNER model using the full set of expert-annotated data. For RE, we introduced ATLOP, a model leveraging adaptive thresholding and localized context pooling to effectively capture long-tail relations (Zhou et al., 2021). This feature is critical for the GUTBRAINIE corpus, where various low-frequency relations play a significant role in accurately representing biomedical interactions. ATLOP was trained with expert annotations as primary supervision and student annotations as weak supervision. Although no manual revision was performed on this data, we applied a post-processing step to ensure adherence to the same schema and guidelines as the human-curated data.

Collection Overview. Once all the documents from the original retrieval were annotated, either manually or automatically, we organized them into four quality-based folds reflecting the reliability of the annotations and the level of expertise of the

Table 1: GutBrainIE dataset statistics.

Collection	No. Docs	No. Ents	Ents/Doc	No. Rels	Rels/Doc
Silver	499	15,275	30.61	10,616	21.27
Gold	208	5,192	24.96	1,994	9.59
Platinum	111	3,638	32.77	1,455	13.11
Dev	40	1,117	27.93	623	15.58
Test	40	1,237	30.92	777	19.42
Manual	898	26,459	29.46	15,465	17.22
Automatic (Bronze)	749	21,420	28.51	8,533	10.90
Overall	1,647	47,879	29.03	23,998	14.35

annotators involved: 1. **Platinum**: highest-quality annotations produced by experts during the first manual annotation phase and internally reviewed by a dedicated subgroup. 2. **Gold**: high-quality annotations created by experts during the second manual annotation phase. 3. **Silver**: annotations of intermediate quality produced by trained layperson annotators under expert supervision. 4. **Bronze**: automatically generated annotations.

The training set includes all four quality folds. As for the Development and Test sets, they contain only expert annotations (Platinum and Gold).

Each annotation is tagged with an anonymized annotator ID, with expert annotators identified as `expert[1-7]` and automatic annotations as `automatic`. For layperson annotators, we clustered them into two groups (A and B) based on annotation quality: Group A achieved the highest overlap with experts ($\geq 65\%$ for entities and $\geq 40\%$ for relations), whereas Group B showed lower agreement. Accordingly, student annotators are identified as `student[A/B]`, depending on the reliability cluster to which they have been assigned.

This tiered annotation quality, along with annotator metadata, allows models to be trained or tuned by weighting or filtering annotations differently based on their reliability, supporting training strategies such as reliability-aware loss weighting (Lin et al., 2019; Ibrahim et al., 2020; Guo and Yang, 2024) and selective denoising (Ghosh et al., 2023; Alomar et al., 2023). Table 1 shows the distribution of documents and annotations per fold.

3 Data Analysis

To better contextualize our corpus features and strengths, we compare it against a selection of widely used general-purpose and biomedical IE datasets. We excluded most distantly supervised datasets and focused on manually annotated ones that provide reliable labels and are directly comparable to ours (Amin et al., 2020, 2022).

Data Size. The GutBrainIE corpus consists of 1,647 documents, of which 898 are manually annotated. This positions it among the largest manually curated biomedical corpora, with a size comparable to BioRED (600 documents) and NCBI disease (793). In terms of length, GutBrainIE documents average 235 words, which is aligned with widely used biomedical and general-purpose datasets such as BioRED (240 words), JNLPBA (240), NCBI Disease (227), and DocRED (210) (Luo et al., 2022; Collier et al., 2004; Doğan et al., 2014; Yao et al., 2019). Table 2 reports overall statistics for GutBrainIE and several representative IE datasets. We report aggregated statistics for both the full collection and the manually annotated subset.

Annotated Entities and Relations. GutBrainIE contains an average of 29 entity mentions and 15 relations annotated per document, comparable to BioRED (34 mentions, 11 relations) and exceeding other general-purpose corpora like DocRED (26 mentions, 11 relations) and HacRED (11 mentions, 6 relations), as well as biomedical datasets such as JNLPBA (25 mentions, 15 relations), NCBI Disease (9 mentions), and CHEMDNER (8 mentions) (Luo et al., 2022; Cheng et al., 2021; Collier et al., 2004; Doğan et al., 2014; Krallinger et al., 2015). Nevertheless, while the average number of entity mentions and relations per document in GutBrainIE is comparable to other widely used corpora, our dataset offers greater granularity, more than doubling the number of entity and relation types compared to BioRED (7 and 8, respectively), which is the most fine-grained manually curated biomedical corpus to date (Luo et al., 2022).

Linkage Statistics. As stated in Section 2, all entity mentions in the Gold and Platinum folds have been linked to standardized medical vocabularies, resulting in a total of 1,819 unique URIs assigned across 11,184 annotated mentions.

Each entity type is linked, on average, to concepts from four different vocabularies, with preference given to authoritative resources in the field. In particular, we prioritized mappings to UMLS, a widely used metathesaurus, to maximize interoperability with other biomedical datasets (Bodenreider, 2004). Reference vocabularies associated with each entity type are reported in Table 8.

Table 2: Comparison of GutBrainIE with the main IE datasets. Reported annotation counts include both entities and relations for datasets featuring RE. and the "*Task(s)*,, column indicates the tasks supported by each dataset. Rows highlighted in red correspond to biomedical IE datasets. while those in green indicate general-purpose IE datasets.

Dataset	Year	Task(s)	No. Docs	No. Entity Mentions	No. Entity Types	No. Relations	No. Relation Predicates.	No. Annotations
GutBrainIE (Manual)	2025	NER. NEL. RE	898	26,459 (29.46 per doc)	13	15,465 (17.22 per doc)	17	41,924 (46.69 per doc)
GutBrainIE (Manual+Automatic)	2025	NER. NEL. RE	1,647	47,879 (29.07 per doc)	13	23,998 (14.57 per doc)	17	71,877 (43.64 per doc)
JNLPBA (Collier et al., 2004)	2004	NER	2404	59,963 (24.94 per doc)	5	–	–	59,963 (24.94 per doc)
EU-ADR (Van Mulligen et al., 2012)	2012	NER. RE	300	7,011 (23.37 per doc)	3	2,436 (8.12 per doc)	3	9,447 (31.49 per doc)
BioNLP-CG (Pyysalo et al., 2013)	2013	NER. RE	600	21,683 (36.14 per doc)	4	917 (1.53 per doc)	4	22,600 (37.67 per doc)
NCBI Disease (Doğan et al., 2014)	2014	NER. NEL	793	6,892 (8.69 per doc)	1	–	–	6,892 (8.69 per doc)
CHEMDNER (Krallinger et al., 2015)	2015	NER	10,000	84,355 (8.44 per doc)	1	–	–	84,355 (8.44 per doc)
BC5CDR (Li et al., 2016)	2016	NER. NEL. RE	1500	12,850 (8.57 per doc)	2	3,116 (2.08 per doc)	2	15,966 (10.64 per doc)
NLM-Gene (Islamaj et al., 2021)	2021	NER. NEL	550	15,553 (28.28 per doc)	1	–	–	15,553 (28.28 per doc)
BioRED (Luo et al., 2022)	2022	NER. NEL. RE	600	20,419 (34.03 per doc)	6	6,503 (10.84 per doc)	8	26,922 (44.87 per doc)
DocRED (Manual) (Yao et al., 2019)	2019	NER. NEL. RE	5,053	98,533 (19.50 per doc)	6	56,798 (11.24 per doc)	96	155,331 (30.74 per doc)
HacRED (Cheng et al., 2021)	2021	NER. NEL. RE	9,231	98,772 (10.70 per doc)	10	56,798 (6.15 per doc)	26	155,570 (16.85 per doc)
Re-DocRED (Tan et al., 2022)	2022	NER. NEL. RE	4,053	78,628 (19.40 per doc)	6	120,664 (29.77 per doc)	96	199,292 (49.17 per doc)

4 Benchmark Setup and Validation

The GutBrainIE benchmark natively supports four tasks: NER, NEL, M-RE, and C-RE.

The NER task requires identifying and classifying entity mentions into 13 predefined entity types. Each mention is represented as a tuple including: location (title or abstract), character offsets (start and end positions), text span, and entity label. A prediction is considered correct only if it exactly matches a ground truth entry in all these fields.

The NEL task extends NER by additionally requiring the linkage to a URI from one of the reference vocabularies. Here, predicted entities must match ground truth annotations in all NER fields plus URI to be considered correct.

The M-RE task involves detecting and classifying relations between pairs of entity mentions. Each relation is expressed as a 5-tuple: `(subject text span, subject label, relation predicate, object text span, object label)`. Predicted relations are considered correct only if they fully match a ground truth tuple.

The C-RE task mirrors M-RE but evaluates predictions at the concept level, with relations represented as 5-tuples: `(subject URI, subject label, predicate, object URI, object label)`. For example, a predicted relation involving "*Parkinson*" would match a ground truth relation comprising "*Parkinson's disease*" if both mentions are linked to the same URI. The correctness of predicted relations is assessed as in M-RE.

Benchmark Validation. To assess the complexity and utility of the GutBrainIE benchmark, we conducted internal experiments across all four proposed benchmark tasks, adapting the automatic annotation system to serve as a baseline system. Moreover, GutBrainIE has been externally validated by featuring the NER and M-RE tasks in an international evaluation campaign, which attracted 17 participating teams. Internal and external results have been analyzed to assess performance variations across tasks and their relative difficulty, and to validate the intended progression in complexity. Finally, we estimated human performance by comparing non-expert annotations on shared honeypot documents with our baseline system.

All tasks are evaluated using standard IE metrics of Precision (P), Recall (R), and F1-score (F_1), with both macro- and micro-averaging. Let TP_ℓ, FP_ℓ, and FN_ℓ denote the number of True Positives, False Positives, and False Negatives for label ℓ. The label set $\mathcal{L}$ is defined as the set of 13 entity types for the NER and NEL tasks, and the 55 possible relation triples `(subject label, predicate, object label)` for the M-RE and C-RE tasks. Before evaluation, duplicate predictions are removed and overlapping ones are merged, keeping the one with the longest text span. Micro-averaged F_1 is used as the reference metric for all tasks, as it better captures the balanced effectiveness of IE systems and accounts for class imbalance.

Baseline System. To provide reference performances, we derived a baseline system from the one used for pre-annotating documents. To perform NER we adopted GLiNER fine-tuned from the `NuNERZero` checkpoint on the Platinum, Gold, and Silver collections (Zaratiana et al., 2024; Bogdanov et al., 2024). At inference time, we applied

Table 3: Baseline results across all four GutBrainIE supported tasks.

	Macro-avg			Micro-avg		
Task	P	R	F_1	P	R	F_1
NER	0.69	0.71	0.70	0.76	0.82	0.79
NEL	0.40	0.41	0.39	0.50	0.54	0.52
M-RE	0.35	0.18	0.21	0.50	0.25	0.33
C-RE	0.30	0.18	0.20	0.38	0.20	0.27

a 0.6 confidence threshold and merged adjacent and overlapping spans. For RE, we used ATLOP, giving the document text and the entities predicted by GLiNER as inputs (Zhou et al., 2021). ATLOP is trained using the Platinum, Gold, and Silver collections as primary supervision, while the Bronze collection is used as distantly supervised data. Inferred relations are processed to retain only triples defined in the annotation schema. The NEL module links GLiNER-predicted entities by applying the same three-stage hierarchical approach described above. Baseline results are reported in Table 3.

External Validation. To exhaustively validate the GutBrainIE benchmark, the NER and M-RE tasks have been offered within an international evaluation campaign. Participants were provided with the dataset, annotation guidelines, and the baseline system. Teams were free to employ any model architecture, training strategy, or external resource.

A total of 17 distinct teams participated: 15 in NER and 12 in M-RE, submitting 101 and 95 runs respectively. Considering micro-averaged F_1 as the reference metric, 8 teams outperformed the baseline on NER (38 runs) and 5 on M-RE (24 runs). These results highlight both the competitiveness of the baseline system and the clear margin for improvement (see Tables 4 and 5 in the annex for team scores). For full details on approaches and results, see (Martinelli et al., 2025).

Model Performance. The results obtained by the baseline and the teams participating in the campaign provide valuable insights into the strengths and limitations of current NLP methods across biomedical IE tasks of increasing complexity.

For NER, most teams relied on supervised fine-tuning of pre-trained biomedical transformers, including PubMedBERT, BioBERT, BioLinkBERT, and BioMedELECTRA (Gu et al., 2021; Lee et al., 2019; Yasunaga et al., 2022; Alrowili and Shanker, 2021). Several participants also adopted GLiNER, the same architecture used in the baseline, but with different checkpoints and fine-tuning strategies (Zaratiana et al., 2024). Many submissions further improved performance through ensemble methods, either by combining different models or multiple instances of the same model trained with different configurations, seeds, or data splits. Indeed, while most systems were trained exclusively on the manually curated Platinum, Gold, and Silver collections, a few teams also included the Bronze collection, using reweighting or filtering strategies to mitigate its noise. However, these attempts led to inconsistent improvements, suggesting that the benefits of using automatically annotated data in high-granularity IE tasks remain limited.

M-RE results reflected the inherent high complexity of this task. Most teams approached M-RE as a supervised classification problem over entity mention pairs predicted by an upstream NER module. To address class imbalances and long-tail relations, participants employed negative sampling, class-weighted loss functions, and filtering heuristics, and a few explored more advanced architectures such as query-based encoders and hypergraph models (Feng et al., 2019).

Nevertheless, the baseline system achieved competitive results, ranking close to the median across both subtasks, indicating the difficulty of achieving substantial improvements on our benchmark. In contrast, prompt-based and zero-shot LLM approaches used by a few teams to perform end-to-end NER and RE performed substantially worse, indicating the current limitations of LLMs for domain-specific and fine-grained IE tasks.

Overall, these results demonstrate that IE systems face numerous challenges in achieving robust performance on the GutBrainIE benchmark, as participants who achieved substantial improvements over the baseline did so by adopting sophisticated architectures, advanced training strategies, and computationally intensive solutions. This underscores that our benchmark offers ample opportunities for methodological research and innovation.

Human Performance. To estimate human-level performance on the benchmark tasks proposed in the evaluation campaign, we evaluated layperson annotations on the shared honeypot documents. Each set of annotations by a student was treated as an individual system submission and evaluated using the same script and metrics applied to participant test runs, with the final annotated version of each honeypot document used as ground truth. To establish a fair comparison, we re-trained our baseline system leaving out the honeypot documents to prevent data leakage, then ran inference on the honeypot set and evaluated its predictions.

For the NER task, all laypeople achieved Precision (P), Recall (R), and F_1 scores above 0.40, with average scores of 0.79 P, 0.77 R, and 0.77 F_1. Although lower, results were still robust for M-RE, where, on average, laypeople scored higher on P (0.61) and slightly lower on R (0.52) and consequently on F_1 (0.53). The baseline system achieved a micro-averaged P, R, and F_1 of 0.83 for NER, and 0.44 P, 0.31 R, and 0.37 F_1 for M-RE. These results indicate that, while NER can be effectively tackled by automatic systems with performance comparable to non-expert annotators, RE remains significantly more complex. In this task, layperson annotators consistently outperformed the baseline across all metrics, highlighting the semantic and contextual difficulty of our benchmark.

Discussion. The experimental results confirm that GUTBRAINIE is a robust and well-designed benchmark, presenting significant challenges for current IE systems, especially in RE tasks, which demand methodological advancements and refined IE approaches. External validation shows that established BioNLP methods are highly effective across tasks. In contrast, emerging methods based on LLMs are still immature and do not match the performance of supervised systems in specialized domains, indicating that their potential remains largely unexploited for fine-grained IE. Moreover, we found out that models trained on smaller subsets of expert-annotated data consistently outperformed those trained on larger datasets that included the automatically annotated noisier portion of the corpus, reinforcing the importance of high-quality annotations in IE and validating the tiered quality structure of the GUTBRAINIE corpus.

5 Related Work

A variety of datasets have been developed to support IE in the biomedical domain, particularly for NER, NEL, and RE. Early biomedical corpora such as JNLPBA (Collier et al., 2004), EU-ADR (Van Mulligen et al., 2012), BioNLP-CG (Pyysalo et al., 2013), and BC5CDR (Li et al., 2016) focused on a small number of entity types, including genes, diseases, and chemicals, and helped establish benchmarks for single-type NER systems. However, systems trained on one or a few entity types often fail to generalize across broader arrays of biomedical concepts. Moreover, various studies demonstrated that in BioNLP multi-type NER models can perform comparably or better than single-type models, showing greater capabilities in leveraging contextual information and handling ambiguities (Crichton et al., 2017; Wang et al., 2019). For what concerns RE, the high cost of manual annotations led most biomedical IE datasets to rely on distant supervision for inferring relations, inevitably introducing noise and incorrectness (Karp, 2016; Amin et al., 2020, 2022). Recognizing that gap, BioRED introduced a fine-grained dataset covering six biomedical entity types and eight relation predicates, with manually curated NER, NEL, and document-level RE annotations (Luo et al., 2022).

Compared to existing resources, GUTBRAINIE introduces a large domain-specific IE benchmark with manual annotations for entities, concept-level linkages, and relations divided into a multi-tiered quality structure. To our knowledge, GUTBRAINIE provides the most fine-grained annotation schema to date for both entities and relations in the biomedical domain.

6 Conclusions and Future Work

In this work we presented GUTBRAINIE, a comprehensive IE benchmark focusing on the emerging biomedical research area of the gut-brain axis. GUTBRAINIE provides a large domain-specific dataset manually curated, supporting four well-defined tasks of increasing complexity (NER, NEL, M-RE, and C-RE), each accompanied by standardized evaluation measures and a competitive baseline system. To demonstrate its impact and practical utility, we featured two of its tasks (NER and M-RE) as part of an international evaluation campaign, which attracted 17 participating teams with nearly 200 system submissions. Experimental results indicate that current NER systems are effective even in specialized domains, while RE remains a significantly more challenging task in domains requiring deep contextual and semantic understanding. GUTBRAINIE has been developed to support IE research at large, proposing a reliable and challenging evaluation framework for settings characterized by domain specificity, limited training data, and complex terminology. In future work, we plan to extend the GUTBRAINIE corpus by annotating documents related to additional neurodegenerative diseases (e.g., Alzheimer's, Multiple Sclerosis) and to further enhance the quality of existing data by manually revising the Silver and Bronze folds.

Limitations

While GutBrainIE is a novel and high-quality benchmark, it has a few limitations. First, from the analysis of layperson annotations conducted to cluster them into the two reliability groups (cf. Section 2), we observed that the Silver collection includes annotations that are not fully consistent with those in the Platinum and Gold collections. Moreover, the automatically generated annotations of the Bronze collection exhibit notable noise. As observed in experiments, incorporating these lower-quality folds directly into training might degrade model performance.

Concerning the annotation workflow, it was conducted in separate batches, which may have introduced inconsistencies in annotations. Future annotation cycles could leverage active learning techniques with continuous model-in-the-loop feedback to enhance consistency and reduce manual effort.

Finally, given that the involvement of biomedical domain experts played a critical role in the development of the conceptual schema and annotation guidelines, future annotation cycles could further benefit from their integration as external reviewers or adjudicators to enhance annotation quality and accuracy.

Ethical Considerations

Below we detail relevant ethical aspects related to the construction and dissemination of the GutBrainIE dataset:

- **Intellectual property and data sources**: the GutBrainIE dataset comprises only titles and abstracts of biomedical articles retrieved from PubMed, a publicly accessible electronic database. These documents are available for reuse under terms that permit research and educational purposes, and no full-text content or content under restricted licenses has been included.

- **Annotators privacy**: although we collected information about the annotators to manage task assignments and quality control during the annotation process, the released version of the dataset includes only anonymized identifiers, designed to preserve the utility of the data while ensuring that no individual annotator can be personally identified from the released dataset.

- **Annotators compensation**: all annotations were performed by volunteers, who were informed in advance that no monetary compensation would be provided. Participation was entirely voluntary and conducted in a non-commercial academic context.

- **Data transparency and characteristics**: detailed information about the annotation schema, workflow, and data characteristics is provided in the appendix and in the annotation guidelines file.

- **Potential data quality issues**: although a significant portion of data has been manually annotated by experts following strict guidelines, we acknowledge the possibility of residual noise or inconsistencies also in the highest-quality Gold and Platinum collections. Such issues are common in most publicly available corpora and are not expected to critically impact downstream applications.

- **Use of generative AI**: during the preparation of this work, the authors used GPT-4o and Grammarly for grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

We believe that the publication and use of GutBrainIE will contribute positively to the development of robust and effective IE systems across a variety of semantically rich and complex domains, including but not limited to Biomedical Natural Language Processing (BioNLP) and health-related applications.

Acknowledgements

This project has received funding from the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

References

Khaled Alomar, Halil Ibrahim Aysel, and Xiaohao Cai. 2023. Data augmentation in classification and segmentation: A survey and new strategies. *Journal of Imaging*, 9(2):46.

Omar Alonso and Gary Marchionini. 2019. *The Practice of Crowdsourcing*. Morgan & Claypool Publishers.

Sultan Alrowili and Vijay Shanker. 2021. BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva, and Guenter Neumann. 2020. A data-driven approach for noise reduction in distantly supervised biomedical relation extraction. *arXiv preprint arXiv:2005.12565*.

Saadullah Amin, Pasquale Minervini, David Chang, Pontus Stenetorp, and Günter Neumann. 2022. Meddistant19: towards an accurate benchmark for broad-coverage biomedical relation extraction. *arXiv preprint arXiv:2204.04779*.

Jeremy Appleton. 2018. The gut-brain axis: influence of microbiota on mood and mental health. *Integrative Medicine: A Clinician's Journal*, 17(4):28.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data.

Marilia Carabotti, Annunziata Scirocco, Maria Antonietta Maselli, and Carola Severi. 2015. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Annals of gastroenterology: quarterly publication of the Hellenic Society of Gastroenterology*, 28(2):203.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. 2021. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831, Online. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Nigel Collier, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi, and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78.

Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. 2017. A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18:1–14.

John F Cryan, Kenneth J O'Riordan, Kiran Sandhu, Veronica Peterson, and Timothy G Dinan. 2020. The gut microbiome in neurological disorders. *The Lancet Neurology*, 19(2):179–194.

Kartik Detroja, C.K. Bhensdadia, and Brijesh S. Bhatt. 2023. A survey on Relation Extraction. *Intelligent Systems with Applications*, 19:200244.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3558–3565.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kuzman Ganchev, Fernando Pereira, Mark Mandel, Steven Carroll, and Peter White. 2007. Semi-automated named entity annotation. In *Proceedings of the linguistic annotation workshop*, pages 53–56.

Shivani Ghaisas, Joshua Maher, and Anumantha Kanthasamy. 2016. Gut microbiome in health and disease: Linking the microbiome–gut–brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases. *Pharmacology & therapeutics*, 158:52–62.

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, S Ramaneswaran, and Dinesh Manocha. 2023. Aclm: A selective-denoising based generative data augmentation approach for low-resource complex ner. *arXiv preprint arXiv:2306.00928*.

Robert Greinacher and Franziska Horn. 2018. The DALPHI annotation framework & how its pre-annotations can improve annotator efficiency. *arXiv preprint arXiv:1808.05558*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yue Guo and Yi Yang. 2024. Improving weak-to-strong generalization with reliability-aware alignment. *arXiv preprint arXiv:2406.19032*.

Gibong Hong, Veronica Hindle, Nadine M Veasley, Hannah D Holscher, and Halil Kilicoglu. 2025. DiMB-RE: mining the scientific literature for diet-microbiome associations. *Journal of the American Medical Informatics Association*, 32(6):998–1006.

Karim M Ibrahim, Elena V Epure, Geoffroy Peeters, and Gael Richard. 2020. Confidence-based weighted loss for multi-label classification with missing labels. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 291–295.

Ornella Irrera, Stefano Marchesin, and Gianmaria Silvello. 2024. MetaTron: advancing biomedical annotation empowering relation annotation and collaboration. *BMC bioinformatics*, 25(1):112.

Rezarta Islamaj, Chih-Hsuan Wei, David Cissel, Nicholas Miliaras, Olga Printseva, Oleg Rodionov, Keiko Sekiya, Janice Ward, and Zhiyong Lu. 2021. NLM-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of biomedical informatics*, 118:103779.

Peter D Karp. 2016. How much does curation cost? *Database*, 2016:baw110.

Jung-Jae Kim, Xu Han, Vivian K Lee, and Dietrich Rebholz Schuhmann. 2013. GRO Task: Populating the Gene Regulation Ontology with events and relations. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 50–57.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(Suppl 1):S2.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1):170.

Sang Gyu Kwak and Jong Hae Kim. 2017. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Ying Lin, Liyuan Liu, Heng Ji, Dong Yu, and Jiawei Han. 2019. Reliability-aware dynamic feature composition for name tagging. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 165–174.

Ting Liu, Xueli Pan, Xu Wang, K Anton Feenstra, Jaap Heringa, and Zhisheng Huang. 2021. Exploring the microbiota-gut-brain axis for mental disorders with knowledge graphs. *Journal of Artificial Intelligence for Medical Sciences*, 1(3):30–42.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

M Martinelli, G Silvello, V Bonato, GM Di Nunzio, N Ferro, O Irrera, S Marchesin, L Menotti, F Vezzani, et al. 2025. Overview of GutBrainIE@ CLEF 2025: Gut-Brain Interplay Information Extraction. In *CLEF 2025 Working Notes*, volume 4038, pages 65–98.

Marco Martinelli, Stefano Marchesin, and Gianmaria Silvello. 2026. Efficient and Reliable Estimation of Named Entity Linking Quality: A Case Study on GutBrainIE.

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, and Jan Hajič. 2023. Quality and efficiency of manual annotation: Pre-annotation bias. *arXiv preprint arXiv:2306.09307*.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, et al. 2025. Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 173–198. Springer.

Yesol Park, Gyujin Son, and Mina Rho. 2024. Biomedical flat and nested named entity recognition: Methods, challenges, and advances. *APPLIED SCIENCES-BASEL*, 14(20):1–23.

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. Overview of the cancer genetics (cg) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66.

Peng Su, Gang Li, Cathy Wu, and K Vijay-Shanker. 2019. Using distant supervision to augment manually annotated data for relation extraction. *PloS one*, 14(7):e0216913.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED–Addressing the False Negative Problem in Relation Extraction. *arXiv preprint arXiv:2205.12696*.

Bui Duc Tho, Minh-Tien Nguyen, Dung Tien Le, Lin-Lung Ying, Shumpei Inoue, and Tri-Thanh Nguyen. 2024. Improving biomedical Named Entity Recognition with additional external contexts. *Journal of Biomedical Informatics*, 156:104674.

Erik M Van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A Kors, and Laura I Furlong. 2012. The

EU-ADR corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, 45(5):879–884.

Ruili Wang, Feng Hou, Steven F Cahan, Li Chen, Xiaoyun Jia, and Wanting Ji. 2022. Fine-grained entity typing with a type taxonomy: a systematic review. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4794–4812.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining Language Models with Document Links. In *Association for Computational Linguistics (ACL)*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

A GutBrainIE Benchmark at a Glance

The entire GutBrainIE benchmark is summarized in Figures 1-2. In particular, Figure 2 shows the four-stage annotation workflow adopted to build the GutBrainIE collection. The *Named Entity Linking Phase* is highlighted with a purple-to-yellow gradient to indicate that it combines automatic and manual annotation. Automatic annotations for the remaining documents were produced using the system from which our baseline system is derived.

B Shared Task Results

The results obtained by teams participating in the international shared task are summarized in Tables 4 (NER) and 5 (M-RE) (Nentidis et al., 2025; Martinelli et al., 2025). For each participating team, we report the performance of their best submitted run, considering the micro-averaged F_1-score as the reference metric. In both tables, the entry BASELINE (highlighted in blue) corresponds to the results achieved by our baseline system.

Across submissions on both tasks, participants largely relied on supervised pipelines built around biomedical transformer encoders, with performance gains mainly coming from system-level modifications rather than fundamentally different paradigms. For NER, strong systems consistently relied on fine-tuned transformers used for token classification (often enhanced with structured decoding such as CRF-style label dependency modeling). Performance gains have been observed with ensembling, confidence thresholding, and lightweight post-processing to correct boundaries and merge overlaps. Several teams experimented with incorporating lower-quality (bronze/silver) annotations or pseudo-labeled data, typically after cleaning or reweighting, suggesting that additional weak supervision can be beneficial in some configurations but requires careful filtering to avoid introducing noise. In a minority of cases, generation- or schema-driven Large Language Model (LLM) extraction was explored (zero-shot or with limited supervision), but these approaches exhibited recurring practical issues such as span misalignment and hallucinated markup that required non-trivial post-processing and, in most cases, resulted in underperforming tranformer-based systems.

For M-RE, most approaches framed the task as classification over candidate entity pairs, using explicit entity markers (or query-style formulations) to condition predictions and mitigate the large negative space via negative sampling and class rebalancing. Performance improvements were most often associated with stronger candidate construction and sampling strategies, dataset cleaning (e.g., sampling or filtering extreme relation density cases), and ensembling or fusion schemes. A few submissions explored alternative formulations (e.g., document-level interaction modeling, end-to-end relation generation, or multi-stage reasoning with auxiliary LLM components), but the general trend remained that fully prompting-based, zero-shot LLM solutions were not competitive on these fine-grained biomedical RE settings and tended to be highly unreliable.

For a complete, method-by-method discussion, implementation details, and the full set of results for all submitted runs, we refer the reader to the shared task overview (Martinelli et al., 2025).

C Human Performance

To estimate human-level performance on the benchmark tasks proposed in the evaluation campaign, we evaluated layperson annotations on the shared honeypot documents. Each student's annotations were treated as an individual system submission and evaluated using the same script and metrics applied to participant test runs, with the final annotated version of each honeypot document used as ground truth. To establish a fair comparison, we re-trained our baseline system leaving out the honeypot documents to prevent data leakage, then ran inference on the honeypot set and evaluated its predictions.

For the NER task, all laypeople achieved micro-averaged precision, recall, and F_1 scores above 0.40, with average scores of 0.79 precision, 0.77 recall, and 0.77 F_1. Although lower, results were still robust for M-RE, where, on average, laypeople scored higher on precision (0.61) and slightly lower on recall (0.52) and consequently on F_1 (0.53). The baseline system achieved a micro-averaged precision, recall, and F_1 of 0.83 for NER, and 0.44 precision, 0.31 recall, and 0.37 F_1 for M-RE. These results indicate that, while NER can be effectively tackled by automatic systems, achieving results comparable to those of non-expert annotators, RE remains significantly more complex. Indeed, in this task layperson annotators consistently outperformed the baseline across all metrics, highlighting the semantic and contextual difficulty of our bench-

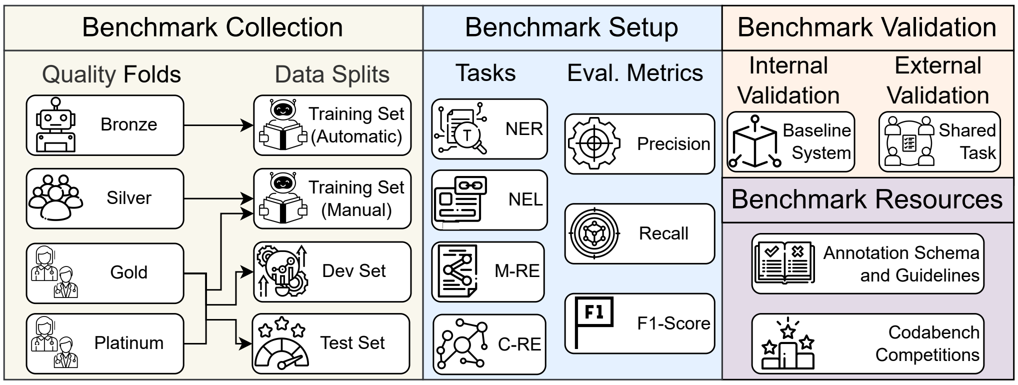


Figure 1: Summary of the main features and contributions of the GUTBRAINIE benchmark.

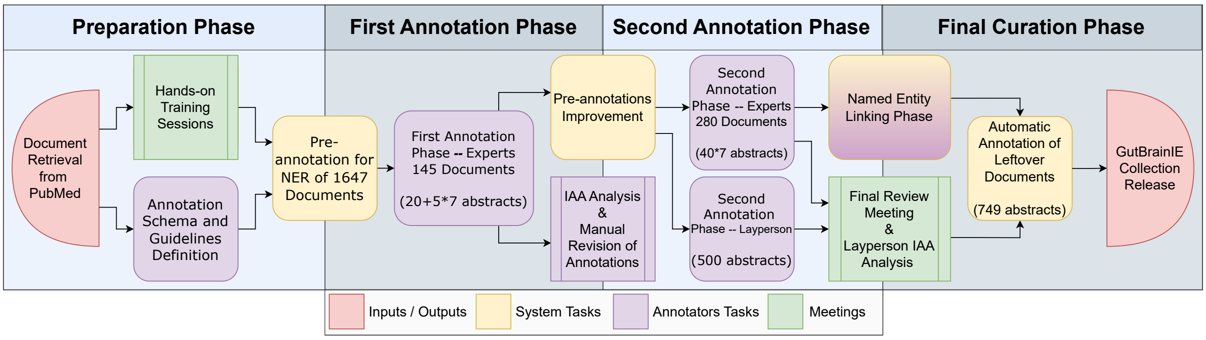

Preparation Phase
First Annotation Phase
Second Annotation Phase
Final Curation Phase
Document Retrieval from PubMed
Hands-on Training Sessions
Annotation Schema and Guidelines Definition
Pre-annotation for NER of 1647 Documents
First Annotation Phase -- Experts 145 Documents
(20+5*7 abstracts)
Pre-annotations Improvement
IAA Analysis & Manual Revision of Annotations
Second Annotation Phase -- Experts 280 Documents
(40*7 abstracts)
Second Annotation Phase -- Layperson
(500 abstracts)
Named Entity Linking Phase
Final Review Meeting & Layperson IAA Analysis
Automatic Annotation of Leftover Documents
(749 abstracts)
GutBrainIE Collection Release
Inputs / Outputs
System Tasks
Annotators Tasks
Meetings

Figure 2: The four-stage workflow followed in the curation of the GUTBRAINIE collection.

Table 4: Performance metrics of each team's top run for NER. For each evaluation metric, the best result is in bold, the second-best is underlined. Runs are ranked based on micro-averaged F_1-score

			Macro-avg			Micro-avg		
Team ID	**Country**	**Affiliation**	P	R	F_1	P	R	F_1
GutUZH	Switzerland	University of Zurich	0.7950	0.7736	**0.7613**	**0.8384**	0.8432	**0.8408**
Gut-Instincts	Denmark	Aalborg University	0.7619	0.7813	0.7591	0.8286	0.8480	0.8382
NLPatVCU	United States	Virginia Commonwealth University	0.8139	0.7161	0.7169	0.8255	0.8488	0.8370
ICUE	United Kingdom	The University of Edinburgh	**0.8216**	0.7451	0.7546	0.8369	0.8294	0.8331
LYX-DMIIP-FDU	China	Fudan University	0.7605	**0.7910**	0.7347	0.8020	**0.8513**	0.8259
ata2425ds	Italy	University of Padua	0.7199	0.7546	0.7217	0.7914	0.8432	0.8164
greenday	United States	Stony Brook University	0.7368	0.7682	0.7471	0.7956	0.8278	0.8114
Graphswise-1	Bulgaria	Graphwise	0.7691	0.7398	0.7185	0.8066	0.7955	0.8010
BASELINE	–	–	0.6883	0.7690	0.7047	0.7639	0.8238	0.7927
ataupd2425-gainer	Italy	University of Padua	0.5808	0.5322	0.5281	0.8333	0.7397	0.7837
DS@GT-bioasq-task6	United States	NA	0.6342	0.7849	0.6872	0.7337	0.8197	0.7743
DS@GT-BioNER	Canada	NA	0.6731	0.6497	0.6469	0.7783	0.7437	0.7606
ataupd2425-pam	Italy	University of Padua	0.6400	0.7435	0.6763	0.6809	0.7745	0.7247
Schemalink	Italy	University of Milan	0.4813	0.5038	0.4650	0.5547	0.5659	0.5602
BIU-ONLP	Israel	Bar Ilan University	0.4393	0.3585	0.3711	0.4916	0.4721	0.4816
lasigeBioTM	Portugal	Universidade de Lisboa	0.2206	0.1034	0.0863	0.3471	0.1964	0.2509

mark.

D Conceptual Schema

Entities and relations to be annotated, and thus to be predicted by IE systems, within the GUTBRAINIE benchmark are defined by the conceptual schema shown in Figure 3. It was collaboratively designed by expert annotators and subsequently validated by external biomedical specialists. During the initial development phase, we explored a wider range of entity types and relation predicates. However, after preliminary pilot annotations, we filtered out those that were excessively underrepresented.

E Types of Named Entities

GUTBRAINIE includes annotations for 13 distinct entity types, listed in Table 6. Each entity type is associated with a unique Uniform Resource Iden-

Table 5: Performance metrics of each team's top run for M-RE. For each evaluation metric, the best result is in bold, the second-best is underlined. Runs are ranked based on micro-averaged F_1-score

			Macro-avg			Micro-avg		
Team ID	**Country**	**Affiliation**	P	R	F_1	P	R	F_1
Gut-Instincts	Denmark	Aalborg University	0.3310	0.4303	**0.3497**	0.4215	0.5147	**0.4635**
Graphswise-1	Bulgaria	Graphwise	0.3323	0.2369	0.2603	0.4686	0.3097	0.3729
ICUE	United Kingdom	The University of Edinburgh	0.2509	0.4239	0.2825	0.2858	0.5054	0.3651
LYX-DMIIP-FDU	China	Fudan University	0.2106	0.2418	0.1990	0.3682	0.3257	0.3457
ONTUG	Austria	University of Graz + Ontotext	0.2589	0.2293	0.2266	0.3529	0.3231	0.3373
BASELINE	–	–	**0.3514**	0.1829	0.2123	**0.4986**	0.2453	0.3288
Schemalink	Italy	University of Milan	0.2265	0.4088	0.2546	0.1948	0.4665	0.2749
ataupd2425-pam	Italy	University of Padua	0.1940	0.2764	0.1982	0.2278	0.3432	0.2738
ataupd2425-gainer	Italy	University of Padua	0.2203	0.1384	0.1538	0.4272	0.1810	0.2542
NLPatVCU	United States	Virginia Commonwealth University	0.1522	**0.5041**	0.2163	0.1423	**0.6005**	0.2300
BIU-ONLP	Israel	Bar Ilan University	0.1171	0.0854	0.0879	0.2339	0.1461	0.1799
ToGS	Austria	University of Graz	0.0249	0.0180	0.0203	0.1702	0.0536	0.0815
lasigeBioTM	Portugal	Universidade de Lisboa	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

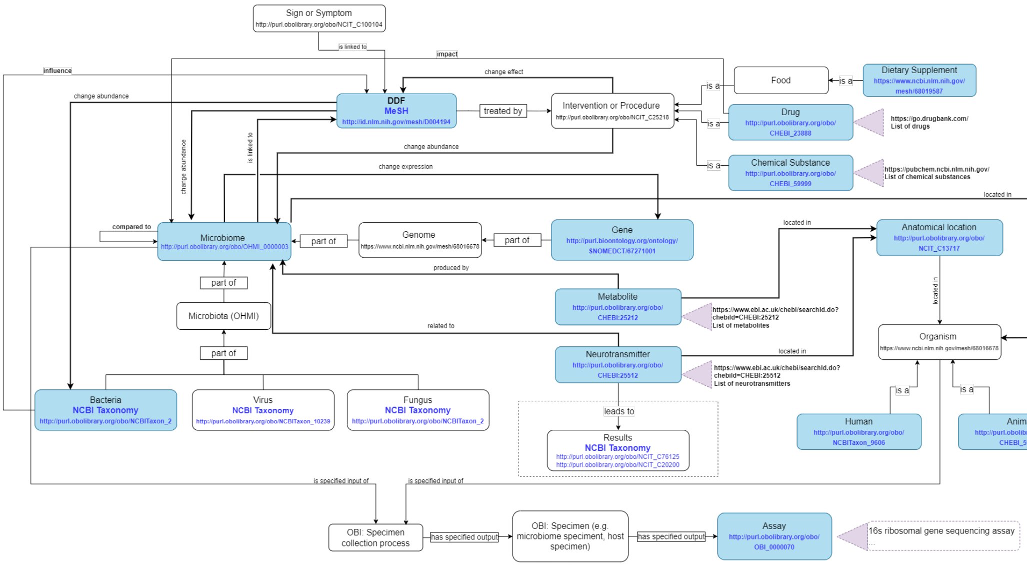

Sign or Symptom
http://purl.obolibrary.org/obo/NCIT_C100104
is linked to
impact
influence
change effect
change abundance
DDF
MeSH
http://id.nlm.nih.gov/mesh/D004194
treated by
Intervention or Procedure
http://purl.obolibrary.org/obo/NCIT_C25218
is a
Food
is a
Dietary Supplement
https://www.ncbi.nlm.nih.gov/mesh/68019587
is a
Drug
http://purl.obolibrary.org/obo/CHEBI_23888
https://go.drugbank.com/
List of drugs
is a
Chemical Substance
http://purl.obolibrary.org/obo/CHEBI_59999
https://pubchem.ncbi.nlm.nih.gov/
List of chemical substances
change abundance
is linked to
change abundance
change expression
located in
compared to
Microbiome
http://purl.obolibrary.org/obo/OHMI_0000003
part of
Genome
https://www.ncbi.nlm.nih.gov/mesh/68016678
part of
Gene
http://purl.bioontology.org/ontology/SNOMEDCT/67271001
located in
Anatomical location
http://purl.obolibrary.org/obo/NCIT_C13717
produced by
part of
Metabolite
http://purl.obolibrary.org/obo/CHEBI:25212
https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:25212
List of metabolites
located in
Microbiota (OHMI)
related to
Organism
https://www.ncbi.nlm.nih.gov/mesh/68016678
part of
located in
Neurotransmitter
http://purl.obolibrary.org/obo/CHEBI:25512
https://www.ebi.ac.uk/chebi/searchId.do?chebiId=CHEBI:25512
List of neurotransmitters
is a
is a
Bacteria
NCBI Taxonomy
http://purl.obolibrary.org/obo/NCBITaxon_2
Virus
NCBI Taxonomy
http://purl.obolibrary.org/obo/NCBITaxon_10239
Fungus
NCBI Taxonomy
http://purl.obolibrary.org/obo/NCBITaxon_2
leads to
Results
NCBI Taxonomy
http://purl.obolibrary.org/obo/NCIT_C76125
http://purl.obolibrary.org/obo/NCIT_C20200
Human
http://purl.obolibrary.org/obo/NCBITaxon_9606
Animal
http://purl.obolibrary.org/obo/CHEBI_59999
is specified input of
is specified input of
OBI: Specimen collection process
has specified output
OBI: Specimen (e.g. microbiome speciment, host specimen)
has specified output
Assay
http://purl.obolibrary.org/obo/OBI_0000070
16s ribosomal gene sequencing assay
...

Figure 3: Conceptual schema defining entity types and relation predicates captured within GUTBRAINIE. Blue rectangles represent annotated entity types, while white rectangles indicate concepts considered during schema design but excluded from annotation due to low frequency. Arrows indicate valid relation directions and predicates between entities.

tifier (URI) that links it to a standardized concept in a reference vocabulary and is accompanied by an explanation that defines its scope and semantic meaning. Moreover, Figure 4a shows the distribution of entity types across the full GUTBRAINIE dataset and within the manually and automatically annotated subsets.

F Types of Relations

GUTBRAINIE features annotations for 17 distinct relation predicates, each of which possibly connects multiple combinations of head and tail entity types, resulting in over 50 possible (*head, predicate, tail*) triples. Table 7 lists all relation predicates used in GUTBRAINIE, represented as (*head, predicate, tail*) combinations according to the conceptual schema depicted in Figure 3. In addition, Figure 4b illustrates the distribution of relation predicates across the full GUTBRAINIE dataset and within the manually and automatically annotated subsets.

G Types of Concept-Level Links

To support semantic normalization and concept-level reasoning, entities annotated in the Platinum and Gold collections have been linked to concepts in standardized biomedical vocabularies. We tried our best to minimize the number of different vocabularies employed, resulting in a total of six biomedical vocabularies and a custom-defined ontology. Each entity type is linked to a number of vocabular-

Table 6: Overview of the 13 entity labels used in the GutBrainIE corpus, including their corresponding URIs and explanations.

Entity Label	URI	Explanation
Anatomical Location	NCIT_C13717	Named locations of or within the body.
Animal	NCIT_C14182	A non-human living organism that has membranous cell walls, requires oxygen and organic foods, and is capable of voluntary movement, as distinguished from a plant or mineral.
Biomedical Technique	NCIT_C15188	Research concerned with the application of biological and physiological principles to clinical medicine.
Bacteria	NCBITaxon_2	One of the three domains of life (the others being Eukarya and ARCHAEA), also called Eubacteria. They are unicellular prokaryotic microorganisms which generally possess rigid cell walls, multiply by cell division, and exhibit three principal forms: round or coccal, rodlike or bacillary, and spiral or spirochetal.
Chemical	CHEBI_59999	A chemical substance is a portion of matter of constant composition, composed of molecular entities of the same type or of different types. This category also includes metabolites, which in biochemistry are the intermediate or end product of metabolism, and neurotransmitters, which are endogenous compounds used to transmit information across the synapses.
Dietary Supplement	MESH_68019587	Products in capsule, tablet or liquid form that provide dietary ingredients, and that are intended to be taken by mouth to increase the intake of nutrients. Dietary supplements can include macronutrients, such as proteins, carbohydrates, and fats; and/or micronutrients, such as vitamins; minerals; and phytochemicals.
Disease, Disorder, or Finding (DDF)	NCIT_C7057	A condition that is relevant to human neoplasms and non-neoplastic disorders. This includes observations, test results, history and other concepts relevant to the characterization of human pathologic conditions.
Drug	CHEBI_23888	Any substance which when absorbed into a living organism may modify one or more of its functions. The term is generally accepted for a substance taken for a therapeutic purpose, but is also commonly used for abused substances.
Food	NCIT_C1949	A substance consumed by humans and animals for nutritional purpose.
Gene	SNOMEDCT_67261001	A functional unit of heredity which occupies a specific position on a particular chromosome and serves as the template for a product that contributes to a phenotype or a biological function.
Human	NCBITaxon_9606	Members of the species Homo sapiens.
Microbiome	OHMI_0000003	This term refers to the entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions.
Statistical Technique	NCIT_C19044	A method of calculating, analyzing, or representing statistical data.

ies ranging from 3 to 6. The vocabularies employed for each entity type are reported in Table 8, which also includes, for each vocabulary and entity type, the number of different unique URIs employed.

H LLM-Based Definition Generation for NEL

As stated in Section 2, the final stage of our NEL pipeline was reached when no match could be

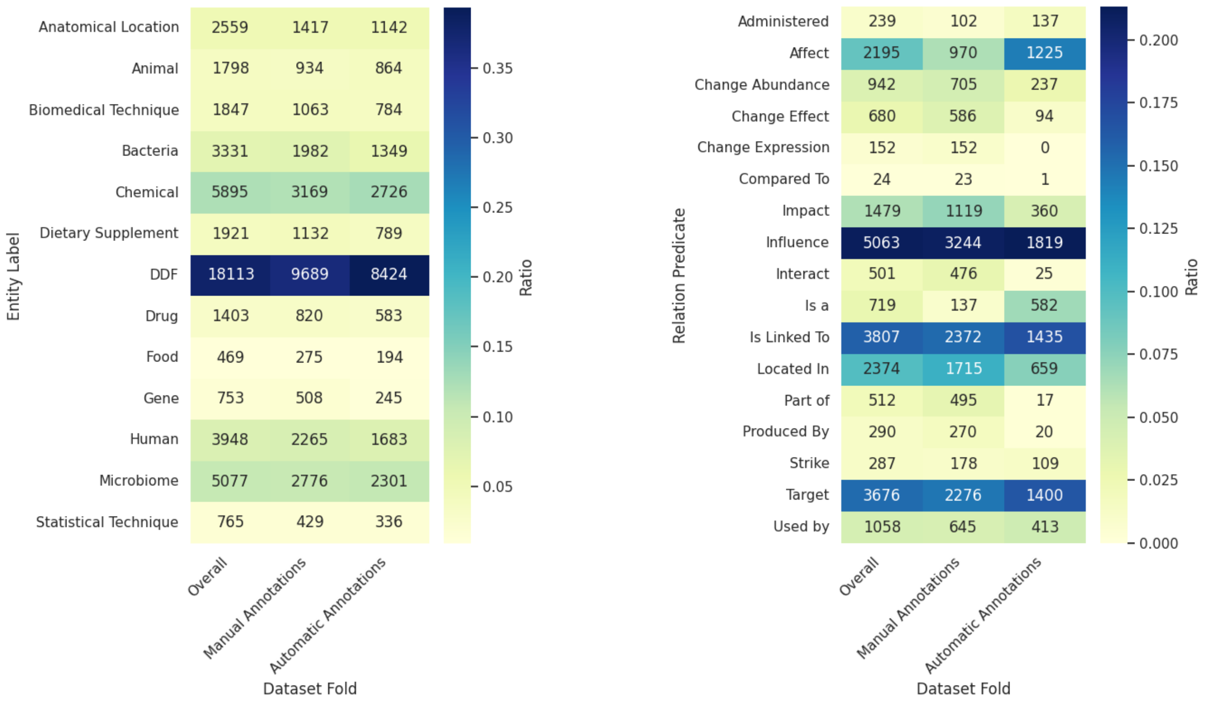
Table 7: Overview of the relations used in the GutBrainIE corpus, expressed as head-predicate-tail triples.

Head Entity	**Tail Entity**	**Predicate**
Anatomical Location	Human Animal	(1) Located in
Bacteria	Bacteria Chemical Drug	(2) Interact
Bacteria	DDF	(3) Influence
Bacteria	Gene	(4) Change expression
Bacteria	Human Animal	(1) Located in
Bacteria	Microbiome	(5) Part of
Chemical	Anatomical Location Human Animal	(1) Located in
Chemical	Chemical	(2) Interact (5) Part of
Chemical	Microbiome	(6) Impact (7) Produced by
Chemical Dietary Supplement Drug Food	Bacteria Microbiome	(6) Impact
Chemical Dietary Supplement Food	DDF	(3) Influence
Chemical Dietary Supplement Drug Food	Gene	(4) Change expression
Chemical Dietary Supplement Drug Food	Human Animal	(8) Administered
DDF	Anatomical Location	(9) Strike
DDF	Bacteria Microbiome	(10) Change abundance
DDF	Chemical	(2) Interact
DDF	DDF	(11) Affect (12) Is a
DDF	Human Animal	(13) Target
Drug	Chemical Drug	(2) Interact
Drug	DDF	(14) Change effect
Human Animal Microbiome	Biomedical Technique	(15) Used by
Microbiome	Anatomical Location Human Animal	(1) Located in
Microbiome	Gene	(4) Change expression
Microbiome	DDF	(16) Is linked to
Microbiome	Microbiome	(17) Compared to

found between an entity mention and the defined reference vocabularies. In these cases, we created a new individual in our custom ontology and prompted a LLM to generate an appropriate definition. Figure 5 shows an example of the prompt and response for the entity mention "*psychobiotics*", labeled as a "*dietary supplement*"

Table 8: Biomedical vocabularies used for concept-level linking of entity mentions. For each entity label, the table lists the reference vocabularies, ordered accordingly to the priority considered in the NEL pipeline (see Section 2), and reports for each of these the number of unique URIs assigned. *GBIE* indicates our custom-defined ontology.

Entity Label	**Linked Vocabularies**
Anatomical Location	UMLS (17), NCIT (70), GBIE (3)
Animal	UMLS (8), NCIT (7), NCBITaxon (5), GBIE (3)
Bacteria	UMLS (23), NCIT (6), NCBITaxon (136), MESH (34), OMIT (1), GBIE (6)
Biomedical Technique	UMLS (65), NCIT (16), OMIT (2), NCBITaxon (4), GBIE (53)
Chemical	UMLS (75), NCIT (94), CHEBI (209), OMIT (2), GBIE (14)
Dietary Supplement	NCIT (34), UMLS (11), CHEBI (13), NCBITaxon (4), OMIT (2), MESH (2), GBIE (3)
DDF	UMLS (179), NCIT (259), OMIT (36), NCBITaxon (1), GBIE (27)
Drug	UMLS (22), NCIT (9), CHEBI (34), OMIT (1), NCBITaxon (1), GBIE (3)
Food	UMLS (23), NCIT (17), GBIE (3)
Gene	UMLS (52), NCIT (42), OMIT (4), CHEBI (1), GBIE (13)
Human	UMLS (38), MESH (8), GBIE (24)
Microbiome	UMLS (2), NCIT (1), NCBITaxon (3), GBIE (15)
Statistical Technique	UMLS (35), NCIT (21), GBIE (23)



(a) Distribution of annotated entity mentions across dataset folds and entity labels.

(b) Distribution of annotated relations across dataset folds and relation predicates.

Figure 4: Distribution of entity types and relation predicates across manually and automatically annotated folds.

Prompt:

```
##Role##
You are a biomedical expert tasked with enriching a biomedical ontology. Given the sentence
below, identify the entity mention marked between **...**, which has been labeled as a dietary
supplement. Write a concise and ontologically appropriate definition for this entity, suitable
for use as an rdfs:comment or skos:definition in a formal biomedical knowledge base.

##Text##
Psychobiotics exhibit positive effects on the intestinal barrier, immune response, cortisol
levels, and the hypothalamic-pituitary-adrenal (HPA) axis. They are a new class of probiotics
that have been shown to improve mental health, including the treatment [...]

##Instructions##
- Output only the definition; do not include any additional commentary, introductions, or
explanations.
- Define the entity as a class-level concept (not as an individual instance).
- Ensure the definition is concise (2–3 sentences), scientifically sound, and suitable for
ontological use.
- Emphasize biomedical function, mechanisms of action, and therapeutic relevance.
- Use a formal, precise tone appropriate for ontology documentation.
```

Generated Answer:

```
Psychobiotics are live microorganisms that, when administered in adequate amounts, produce a
health benefit in the host by modulating the gut-brain axis and influencing brain function
or behavior.
```

Figure 5: Example prompt and response for the LLM-based definition generation stage of our NEL pipeline (see Section 2).