Overview of GutBrainIE@CLEF 2026: Gut-Brain Interplay Information Extraction

Marco Martinelli[1,*], Vanessa Bonato[2], Giorgio Maria Di Nunzio[1], Guglielmo Faggioli[1], Nicola Ferro[1], Benedikt Kantz[3], Ornella Irrera[1], Stefano Marchesin[1], Simone Merlo[1], Laura Menotti[1], Riccardo Michelotto[1], Gaia Tussardi[1], Federica Vezzani[2], Peter Waldert[3] and Gianmaria Silvello[1]

[1]*Department of Information Engineering, University of Padua, Italy*
[2]*Department of Linguistic and Literary Studies, University of Padua, Italy*
[3]*Graz University of Technology, Austria*

Abstract

Recent studies link the gut microbiota to mental health conditions and to neurodegenerative diseases such as Parkinson's, Alzheimer's, and Sclerosis. However, the rapid pace at which this research field is evolving presents a significant challenge for clinicians and researchers who need to stay up to date with new findings and discoveries. In this context, Information Extraction (IE) systems are becoming essential for automatically extracting and structuring information from scientific texts, aiming to support and enhance the understanding of the gut-brain axis. GutBrainIE promotes the development of Natural Language Processing (NLP) systems capable of extracting structured specialized knowledge from biomedical texts related to the gut-brain axis, aiming to accelerate biomedical discoveries through automated IE. GutBrainIE is part of the BioASQ Lab at CLEF 2026 and is promoted by the HEREDITARY research project, funded by the European Commission. The task includes four subtasks, two dealing with entity extraction (Named Entity Recognition and Named Entity Recognition and Disambiguation) and the other two with Relation Extraction (Mention-level Relation Extraction and Concept-level Relation Extraction). The GutBrainIE dataset comprises over 5,000 PubMed documents manually and automatically annotated for entities, concept-level links, and relations, organized into three quality tiers.

This extended overview describes the subtasks, dataset, evaluation methodology, results, and participant approaches for the GutBrainIE-2026 task.

Keywords

Gut-Brain Axis, Gut Microbiota, Mental Health, Neurodegenerative Diseases, Natural Language Processing (NLP), Information Extraction (IE), Named Entity Recognition (NER), Named Entity Linking (NEL), Named Entity Recognition and Disambiguation (NERD), Relation Extraction (RE)

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
martinell2@dei.unipd.it (M. Martinelli); vanessa.bonato@unipd.it (V. Bonato); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio); guglielmo.faggioli@unipd.it (G. Faggioli); nicola.ferro@unipd.it (N. Ferro); benedikt.kantz@tugraz.at (B. Kantz); ornella.irrera@unipd.it (O. Irrera); stefano.marchesin@unipd.it (S. Marchesin); simone.merlo@phd.unipd.it (S. Merlo); laura.menotti@unipd.it (L. Menotti); riccardo.michelotto.1@studenti.unipd.it (R. Michelotto); gaia.tussardi@phd.unipd.it (G. Tussardi); federica.vezzani@unipd.it (F. Vezzani); pwaldert@tugraz.at (P. Waldert); gianmaria.silvello@unipd.it (G. Silvello)
https://www.dei.unipd.it/~martinell2/ (M. Martinelli); https://www.dei.unipd.it/~dinunzio/ (G. M. Di Nunzio); https://www.dei.unipd.it/~faggioli/ (G. Faggioli); https://www.dei.unipd.it/~ferro/ (N. Ferro); https://dakantz.at/ (B. Kantz); https://www.dei.unipd.it/~irreraorne/ (O. Irrera); https://www.dei.unipd.it/~marches1/ (S. Marchesin); https://www.dei.unipd.it/~merlosimon/ (S. Merlo); https://www.dei.unipd.it/~menottilau/ (L. Menotti); https://www.dei.unipd.it/~tussardig/ (G. Tussardi); https://github.com/MrP01 (P. Waldert); https://www.dei.unipd.it/~silvello/ (G. Silvello)
0009-0001-1596-8642 (M. Martinelli); 0009-0002-9918-282X (V. Bonato); 0000-0001-9709-6392 (G. M. Di Nunzio); 0000-0002-5070-2049 (G. Faggioli); 0000-0001-9219-6239 (N. Ferro); 0000-0003-3294-8421 (B. Kantz); 0000-0003-2284-5699 (O. Irrera); 0000-0003-0362-5893 (S. Marchesin); 0009-0003-8003-4795 (S. Merlo); 0000-0002-0676-682X (L. Menotti); 0009-0009-7173-2683 (G. Tussardi); 0000-0003-2240-6127 (F. Vezzani); 0009-0004-8459-7381 (P. Waldert); 0000-0003-4970-4554 (G. Silvello)

1. Introduction

Recent biomedical research increasingly links gut microbiota to neurological and psychiatric disorders, including Parkinson's, Alzheimer's, Multiple Sclerosis, mood disorders and, more in general, to mental health [1, 2, 3, 4]. As shown in Figure 1, publications on PubMed concerning the gut–brain axis and its implications for major neurological disorders more than doubled between 2020 and 2025, rising from 300 to over 700 articles per year. This poses a growing challenge for clinicians and researchers in identifying and interpreting relevant findings across unstructured scientific texts.

The GutBrainIE task, part of the BioASQ Lab and inserted in the context of the EU-supported project HEREDITARY,[1] builds on this challenge, focusing on the development and benchmarking of domain-specific Information Extraction (IE) systems capable of handling the linguistic complexity and conceptual specificity of gut–brain biomedical texts [5, 6]. As in its first edition, the GutBrainIE-2026 task proposes an IE challenge comprising 4 Natural Language Processing (NLP) tasks, requiring participants to extract structured information from unstructured PubMed abstracts related to the gut-brain axis [7, 8, 9]. The final goal of this task is to support the automatic and scalable processing of scientific literature and extraction of information, thereby contributing to biomedical knowledge discovery and, in the long term, informed clinical decision-making.

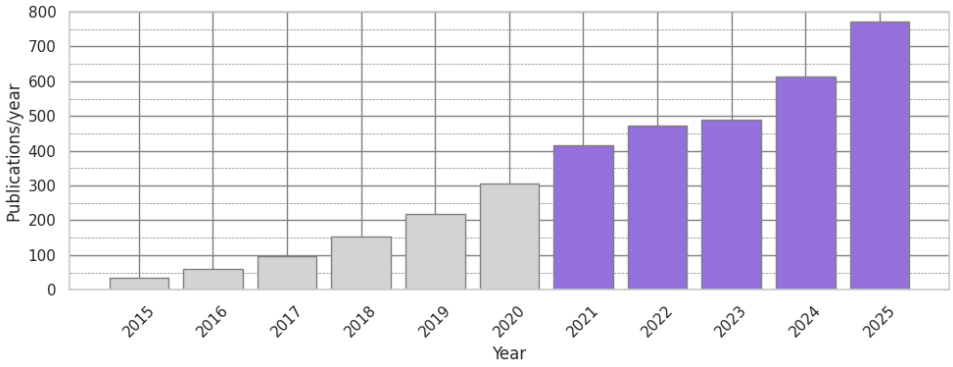

Figure 1: Annual rate of biomedical publications on PubMed related to the gut–brain axis, retrieved with the query *((gut) AND (microbiota OR microbiome)) AND (dementia OR Alzheimer OR multiple sclerosis OR Amyotrophic lateral sclerosis)* on June 10, 2026. Purple bars highlight years with more than 400 records.

GutBrainIE-2026 is the second iteration of the task, in which we kept the number of subtasks fixed at four, with two subtasks remaining the same as in the previous iteration and the other two being replaced. In particular, we added annotations for concept-level links, which enabled subtasks requiring the linkage of entity mentions to concepts from reference vocabularies. Of the four proposed subtasks, two concern entity mentions and the other two concern semantic relations:

- **Subtask 6.1.1 – Named Entity Recognition (NER)**: participants are asked to identify and classify specific text spans (entity mentions) into one of the 13 predefined categories (e.g., *bacteria*, *chemical*, *microbiota*).
- **Subtask 6.1.2 – Named Entity Recognition and Disambiguation (NERD)**: this subtask extends NER by also requiring linking the identified entity mentions to a concept identifier from one of the defined biomedical reference resources.
- **Subtask 6.2.1 – Mention-level Relation Extraction (M-RE)**: participants are provided with a set of predefined relation types, each defined by a combination of compatible head and tail entities (e.g., *Chemical* → *Microbiome* via *Impact* or *Produced by*), and are asked to identify which entity mention pairs are in relation within a document and to assign the correct relation predicate.
- **Subtask 6.2.2 – Concept-level Relation Extraction (C-RE)**: this subtask is similar to Mention-level RE (M-RE), but requires relations to be framed at the concept level. Namely, relations should

[1] https://hereditary-project.eu/

be expressed between linked concepts rather than between their textual mentions, making them abstracted from surface forms and lexical variations.

All subtasks target PubMed abstracts, with an annotated corpus of biomedical documents related to the gut-brain axis. The GutBrainIE-2026 collection includes the entire GutBrainIE-2025 collection, whose annotations were extended with concept-level links to support the newly introduced NERD and Concept-level RE (C-RE) subtasks. For this reason, the current release should be considered the reference collection for the task. In addition, we added and annotated new documents retrieved with the query illustrated in Figure 1. Therefore, the GutBrainIE-2026 dataset consists of over 5,000, more than three times the size of the 2025 collection. We kept the tiered annotation quality, organized as follows:

- **Gold Annotations**: high-quality, expert-curated;
- **Silver Annotations**: mid-quality, created by trained students under expert supervision;
- **Bronze Annotations**: automatically generated with no manual corrections.

The dataset is split into Training, Development, and Test sets, with the latter two containing only expert annotations from the gold-quality fold.

Submissions are evaluated using standard macro- and micro-averaged Precision, Recall, and F1 metrics. Results are compared against a baseline system shared with participants at the beginning of the challenge to provide a reference baseline. As in last year's iteration, these tasks are proposed in an end-to-end setting, meaning that no gold labels for the test set was provided. Therefore, participants wishing to submit to NERD needed first to perform NER. Similarly, for participating in the Relation Extraction (RE) subtasks, they needed to first perform entity detection.

This paper provides a comprehensive overview of the GutBrainIE-2026 task. Section 2 presents the subtasks and their structure; Section 3 introduces the dataset structure and annotation schema; Section 4 presents participating teams and evaluation procedures; Section 5 reports the results and leaderboards across subtasks; Section 6 describes the systems, models, and approaches employed by participating teams; finally, Section 7 concludes the paper and proposes future directions.

2. Task Overview

In its second iteration, GutBrainIE-2026 proposed four subtasks:

1. **Named Entity Recognition (NER)**,
2. **Named Entity Recognition and Disambiguation (NERD)**,
3. **Mention-level RE (M-RE)**,
4. **Concept-level RE (C-RE)**.

In developing their IE systems for these subtasks, participants were not constrained in any aspect regarding system architecture, type of training, or data to use to achieve the best performance. Participation increased compared to last year's iteration [8, 9], with 18 teams participating and submitting a total of 516 runs across the four subtasks.

In the remainder of this Section, we describe each subtask in detail.

2.1. Subtask 6.1.1: Named Entity Recognition (NER)

The NER subtask retained the same framing as in the 2025 iteration, focusing on extracting and classifying entity mentions into one of the 13 predefined categories. Specifically, given PubMed article abstracts related to the gut-brain axis, participants were asked to identify specific text spans corresponding to one of the 13 categories defined in Table 1.

Each entity mention is composed of:

- **Location**, indicating whether the entity mention appears in the title or in the abstract of the PubMed document,

- **Start and end indices**, denoting the start and end character offsets of the entity mention within the text,
- **Text span**, representing the actual string of text corresponding to the mention,
- **Label**, specifying the assigned entity label.

A predicted entity mention is considered correct only if the ground truth contains an entry exactly matching all these fields. Examples of ground truth entries for NER are illustrated in Figure 2.

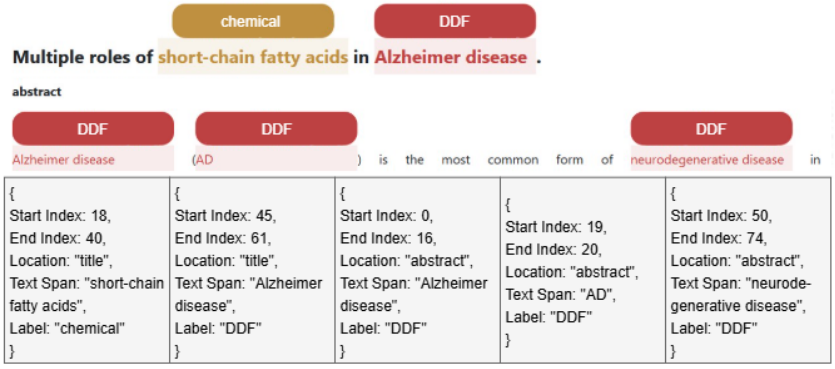

Figure 2: Example of annotated entity mentions for the NER subtask. The figure shows entities tagged in both the title and abstract of a PubMed article.

2.2. Subtask 6.1.2: Named Entity Recognition and Disambiguation (NERD)

The NERD task is one of the two subtasks added in the 2026 iteration of the GutBrainIE task. It extends NER by requiring the extraction and classification of entity mentions into one of the 13 categories and linking them to a concept Uniform Resource Identifier (URI) from one of the reference vocabularies. Since multiple valid concepts may be available across the different vocabularies considered, participants were provided with the full list of concept URIs used in the collection.

In this subtask, predicted entities must match ground truth annotations in all NER fields plus URI to be considered correct. Annotated entities with concept-level links are shown in Figure 3.

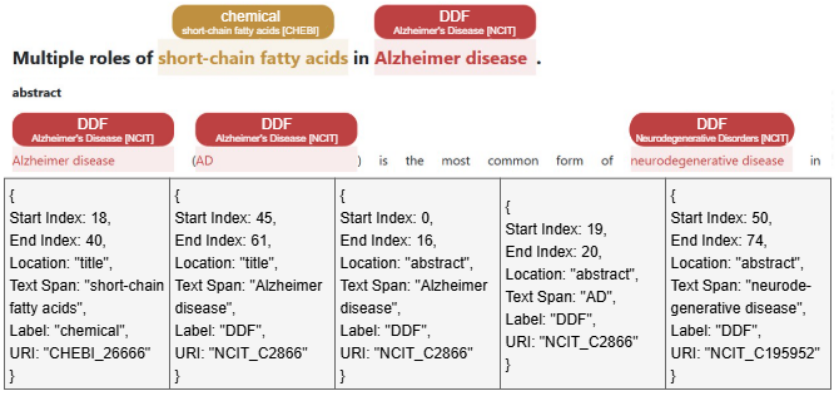

Figure 3: Example of annotated entity mentions for the NERD subtask. The figure shows the same annotated entity mentions illustrated in Figure 2 with concept-level links.

2.3. Subtask 6.2.1: Mention-level Relation Extraction (M-RE)

The M-RE subtask inherits the framing of the Ternary Mention-based Relation Extraction (TM-RE) subtask proposed in GutBrainIE-2025 [9]. In M-RE, participants are required to identify the entity mentions involved in a relation, predict their entity labels, and specify the relation predicate that links

them. The full set of defined relation predicates is reported in Table 2. Predicted relations for M-RE will be 5-tuples in the format: `(subject text span, subject label, predicate, object text span, object label)`. Similarly to the other subtasks, predicted relations are considered correct only if they fully match a ground truth 5-tuple. Figure 4 shows examples of annotated relations at the mention-level.

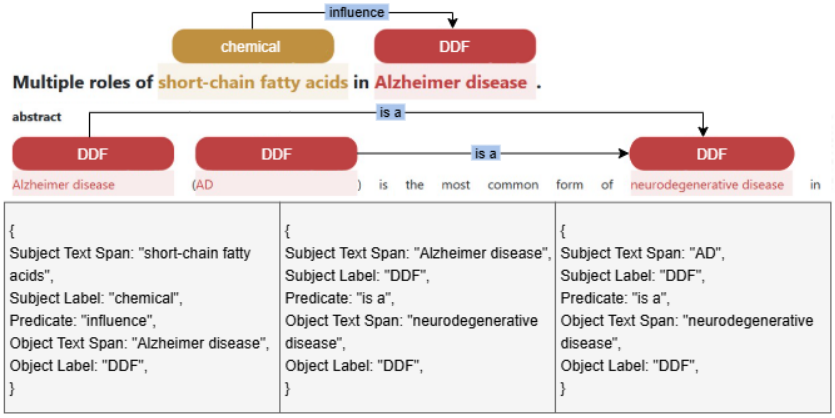

Figure 4: Example of annotated relations for the M-RE subtask. The figure shows pairs of entities identified as being in relation within a PubMed article, specifying the relation predicate and the entity mentions involved.

2.4. Subtask 6.2.2: Concept-level Relation Extraction (C-RE)

The C-RE subtask is the second subtask introduced in the 2026 iteration of GutBrainIE. It mirrors the M-RE subtask but requires entity mentions to be represented at the concept-level rather than at the mention-level. Therefore, C-RE relations are expressed as 5-tuples in the format: `(subject URI, subject label, predicate, object URI, object label)`. The correctness of predicted relations is assessed as in M-RE. Examples of annotated relations at the concept-level are shown in Figure 5.

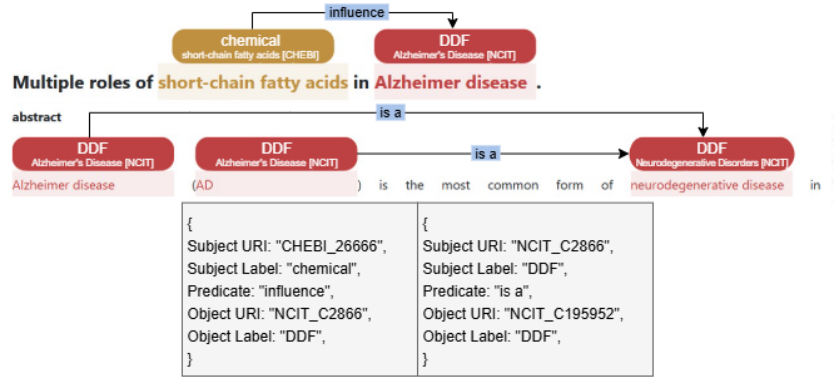

Figure 5: Example of annotated relations for the C-RE subtask. The figure shows the same relations illustrated in Figure 4 framed at the concept level.

3. Dataset

The GutBrainIE-2026 dataset consists of titles and abstracts of biomedical articles retrieved from PubMed, focused on the gut-brain axis and its implications for neurological and mental health conditions. The collection extends the one released for the 2025 challenge, which primarily covered Parkinson's disease and mental health, by adding newly annotated documents on Alzheimer's disease, sclerosis, and

Table 1
Overview of the 13 entity labels used in the GutBrainIE-2026 challenge, including their corresponding URIs and definitions.

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.

dementia. These additional documents were retrieved using the query reported in the caption of Figure 1.

The corpus includes annotations for entity mentions, concept-level links, and relations. These

Table 2
Overview of the relations used in the GutBrainIE-2026 challenge, expressed as head-predicate-tail triples.

Head Entity	Tail Entity	Predicate
Anatomical Location	Human Animal	Located in
Bacteria	Bacteria Chemical Drug	Interact
Bacteria	DDF	Influence
Bacteria	Gene	Change expression
Bacteria	Human Animal	Located in
Bacteria	Microbiome	Part of
Chemical	Anatomical Location Human Animal	Located in
Chemical	Chemical	Interact Part of
Chemical	Microbiome	Impact Produced by
Chemical Dietary Supplement Drug Food	Bacteria Microbiome	Impact
Chemical Dietary Supplement Food	DDF	Influence
Chemical Dietary Supplement Drug Food	Gene	Change expression
Chemical Dietary Supplement Drug Food	Human Animal	Administered
DDF	Anatomical Location	Strike
DDF	Bacteria Microbiome	Change abundance
DDF	Chemical	Interact
DDF	DDF	Affect Is a
DDF	Human Animal	Target
Drug	Chemical Drug	Interact
Drug	DDF	Change effect
Human Animal Microbiome	Biomedical Technique	Used by
Microbiome	Anatomical Location Human Animal	Located in
Microbiome	Gene	Change expression
Microbiome	DDF	Is linked to
Microbiome	Microbiome	Compared to

annotations were manually curated either by experts or by trained students.[2]

3.1. Dataset Creation

The dataset creation process described in this subsection concerns the extension developed for the 2026 edition of GutBrainIE. Details on the construction and curation of the 2025 portion of the collection are reported in [6] and [9].

To build the GutBrainIE-2026 extension, we first retrieved documents from PubMed using the query: `((gut) AND (microbiota OR microbiome)) AND (dementia OR Alzheimer OR multiple sclerosis OR Amyotrophic lateral sclerosis)`. The retrieval was performed on September 16, 2025, and returned 3,100 documents. We then removed documents published before 2015, because of the limited volume of relevant literature in earlier years, as well as documents already included in the GutBrainIE-2025 dataset. After filtering, the retrieval resulted in 2,458 documents.

Before manually annotating the new documents, we enriched the GutBrainIE-2025 collection with concept-level links. This step was applied only to expert-curated annotations, since student-curated annotations were not considered sufficiently accurate for reliable concept mapping. Concept-level links were generated through a semi-supervised pipeline that first normalizes entity text spans using heuristic rules and then matches the normalized strings against concept names in the reference biomedical vocabularies. When no suitable concept was available in the considered resources, a new concept was instantiated in our ontology. Further details on this pipeline are provided in [6]. The automatically generated links were first revised by students. We then involved expert annotators to estimate their overall quality through a sampling-based evaluation framework with statistical guarantees, obtaining an estimated accuracy of 0.915 ± 0.047 [10].

After this enrichment step, and before starting the manual annotation of the newly retrieved documents, we pre-annotated them for NERD to facilitate and speed up the annotation process. Entity mention pre-annotations were produced using the system submitted by team *Gut-Instincts* in the previous edition of the challenge, one of the top-performing systems for the NER subtask, fine-tuned on the 2025 annotations [11]. Concept-level link pre-annotations were then added through a two-stage string-matching pipeline. First, the text spans produced by the NER pre-annotation system were normalized using the same normalization heuristics adopted for the 2025 concept-linking process. Then, for each normalized mention, we searched for an exact match among the concept-linked annotations of the 2025 collection, requiring both the normalized text span and the entity label to match. If no match was found, the search was repeated by considering only the normalized text span and dropping the label constraint. When no match was found in either stage, no concept-level link was added as pre-annotation. As in the 2025 edition, we did not pre-annotate documents for RE. The relation extraction results obtained in the previous iteration of the challenge indicated that the risk of introducing noisy relation pre-annotations was higher than the expected benefit of providing useful annotation suggestions. Such noise could bias annotators and ultimately affect the quality of the final dataset [12].

The retrieved articles were then assigned to expert and student annotators. The annotation process was organized into two phases, with the annotations produced in the first phase used to improve the pre-annotations before the second phase. Overall, 13 experts and 55 students contributed to the annotation effort.

The annotation cycle lasted four months, from October 2025 to January 2026. During this period, expert annotators met once per month to monitor progress, discuss difficult cases encountered during annotation, and revise the annotation guidelines when needed. These guidelines, publicly available at https://hereditary.dei.unipd.it/challenges/gutbrainie/2026/files/GutBrainIE_2026_Annotation_Guidelines.pdf, were also shared with participants to help them design and tune their systems.

Once manual annotation was completed, we fine-tuned the *Gut-Instincts* system [11] for NER and ATLOP [13] for RE, using the complete set of entity and relation annotations produced across the 2025

[2]The students involved in the annotation process are enrolled in the Master's Degree in Modern Languages for International Communication and Cooperation at the University of Padua. They received specific training in medical terminology as part of the course "Translation-Oriented Terminography".

and 2026 annotation cycles. For more details on the fine-tuning of the RE module, refer to Section 4.4. These fine-tuned systems were then used to automatically annotate the remaining unannotated documents from the original retrieval. Concept-level links were added with the same linking pipeline used during pre-annotation. This pipeline was also used to generate concept-level links for the student annotations included in the 2025 collection.

3.2. Dataset Folds

The released collection merges annotations from the 2025 and 2026 annotation cycles. The training set is divided into three parts:

1. **Gold Collection**: high-quality annotations produced by experts.
2. **Silver Collection**: mid-quality annotations curated by trained students under expert supervision. Students were divided into two clusters:
   - *StudentA*, including those with more consistent annotation performance,
   - *StudentB*, including those with less consistent annotation performance.

   Silver annotations produced for the GutBrainIE-2025 edition are kept as a separate subset, since their concept-level links were generated automatically rather than manually curated.
3. **Bronze Collection**: automatically generated annotations, produced as described above. No manual revision was performed on this subset.

The development and test sets are held-out selections of documents from the gold collection, selected to ensure full representativeness and coverage of all entity and relation types. In particular, for both sets half of their documents were drawn from the 2025 annotations and the other half from 2026 annotations.

Detailed statistics on the number of documents, annotated entities, and annotated relations in each fold are reported in Table 3. Figures 6 and 7 show, respectively, the distribution of annotated entities and relations across dataset folds and labels.

Table 3
Dataset statistics for GutBrainIE-2026.

Collection	# Docs	# Entities	Ents/Doc	# Rels	Rels/Doc
Train Gold	639	20,530	32.13	8,556	13.39
Train Silver	811	26,134	32.22	10,907	13.45
Train Silver 2025	499	15,275	30.61	10,616	21.27
Train Bronze	2,972	89,987	30.28	29,692	9.99
Development Set	80	2,521	31.51	1,261	15.76
Test Set	80	2,850	35.62	1,285	16.06

3.3. Dataset Format

Annotations are provided in JSON format. Each entry correspond to a PubMed article, keyed by its PubMed ID (PMID), and contains the following fields:

- **Metadata:** Article-level information including:
  - *title*, *author*, *journal*, *year*, *abstract*;
  - *annotator_id*: one of `expert_1`–`expert_13`, `student_A`, `student_B`, or `distant` (automatically generated). Participants may decide to filter or weight examples differently based on the annotator.
- **Entities:** An array of objects, each with:
  - *start*, *end*: character offsets of the text span associated to the entity mention;
  - *location*: "title" or "abstract";

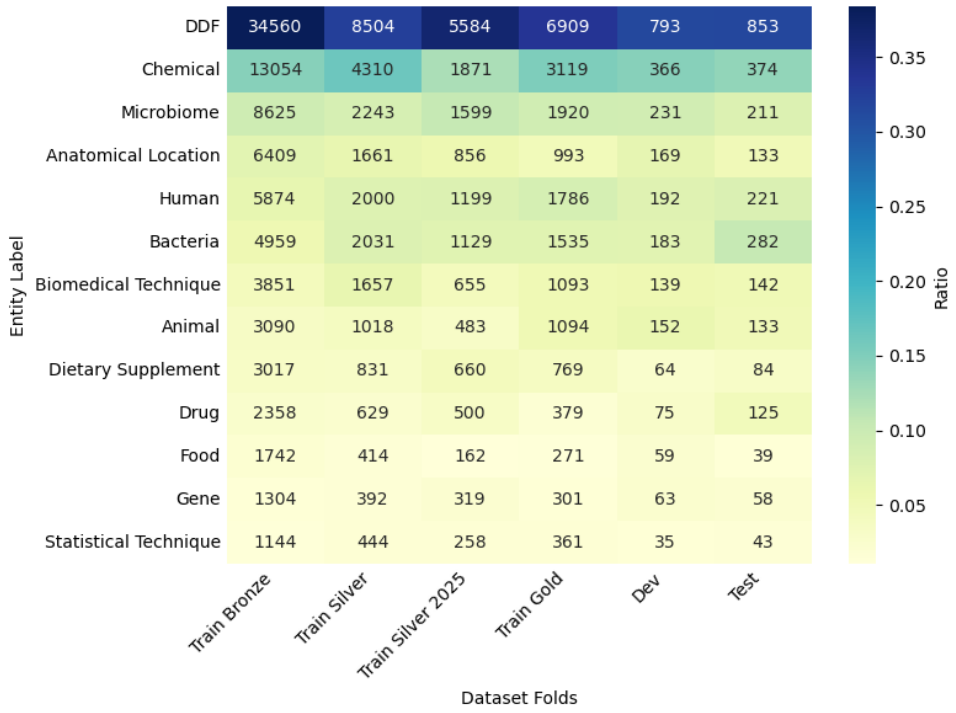

Figure 6: Distribution of annotated entities across dataset folds and entity labels.

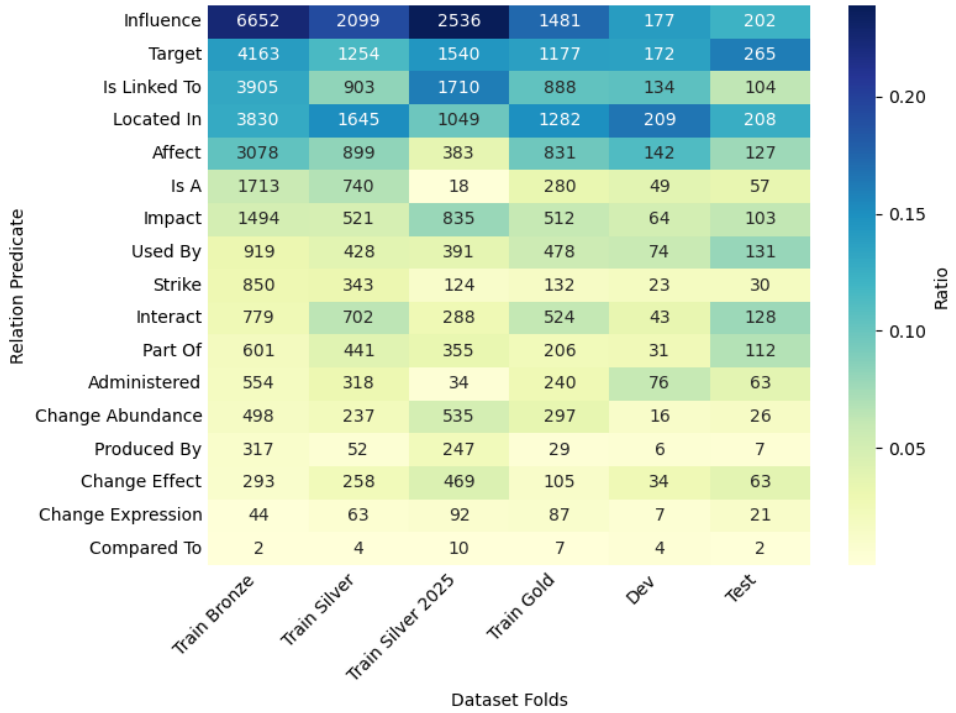

Figure 7: Distribution of annotated relations across dataset folds and relation predicates.

- *text_span*: the actual text span of the mention;
- *label*: the annotated entity label (such as *bacteria*, *microbiome*).
- *uri*: the concept-level link assigned to the entity mention, expressed as the URI of the corresponding concept in one of the reference biomedical resources or in the custom GutBrainIE ontology.

- **Relations:** An array of objects representing relations, each with:
  - *subject_start*, *subject_end*, *subject_location*, *subject_text_span*, *subject_label*: the subject entity mention;
  - *predicate*: the annotated relation predicate;
  - *object_start*, *object_end*, *object_location*, *object_text_span*, *object_label*: the object entity mention.
- **Mention-level Relations:** Extracted from the `Relations` array as mention-level tuples ⟨`subject_text_span`, `subject_label`, `predicate`, `object_text_span`, `object_label`⟩. Information about the location of entity mentions in the text is neglected.
- **Concept-level Relations:** Extracted from the `Relations` array as concept-level tuples ⟨`subject_uri`, `subject_label`, `predicate`, `object_uri`, `object_label`⟩. Information about the location of entity mentions in the text is neglected.

3.3.1. Alternative Dataset Formats

For users preferring tabular data, each field above is also provided in both CSV and TSV formats:

- `metadata.csv` — `metadata.tsv`
- `entities.csv` — `entities.tsv`
- `relations.csv` — `relations.tsv`
- `mention_level_relations.csv` — `mention_level_relations.tsv`
- `concept_level_relations.csv` — `concept_level_relations.tsv`

CSV files use the pipe symbol (`|`) as a delimiter, while TSV files use the tab character (`\t`).

4. Participation and Evaluation

This Section provides a concise overview of the teams that participated in GutBrainIE-2026. A comprehensive description of the submitted systems can be found in Section 6 and in the participants' individual papers listed in Table 5.

As in the previous iteration of the challenge, teams could participate in any of the four subtasks independently and submit up to 25 runs per subtask.

Although 80 teams from 19 different countries registered for the challenge, the final number of teams submitting at least one run was 18, resulting in 516 submitted runs. Among these, 10 teams also submitted a participant paper describing their methodologies, approaches, and systems. However, the discussion presented in Section 6 includes all 18 teams that submitted at least one run. Table 4 summarizes participation across the various subtasks.

4.1. Guidelines

Participating teams were required to satisfy the following guidelines:

- Runs should be submitted in the JSON format described below;
- Each team could submit a maximum of 25 runs per subtask.

Each submitted run had to follow exactly the format specified for the corresponding subtask, include all required fields, and adhere to the defined JSON syntax. No specific ordering of documents or predictions was required.

Table 4
Overview of the participating teams and submitted runs. Column "Runs" indicates the total number of runs presented by the team, while the other columns report the number of runs per subtask submitted.

Team	Runs	NER	NERD	M-RE	C-RE
AAU-Gut-Brain	6	4	2	—	—
ELiRF-UPV	11	10	1	—	—
GBA	40	10	10	10	10
GBIE	2	2	—	—	—
GetGut@AAU	12	3	3	3	3
Graphwise	100	25	25	25	25
GutHub	4	1	1	1	1
MKTE	2	—	1	1	—
MindGut link	8	2	2	2	2
NightSun	100	25	25	25	25
NLP-DE	4	2	2	—	—
NOVALINCS	5	5	—	—	—
SMTE	5	3	—	2	—
TEXA	4	1	1	1	1
ToGS	98	24	24	25	25
TUGW	18	8	8	—	2
TWIX	95	23	24	24	24
unibuc-bionlp-bp	2	2	—	—	—

4.1.1. Subtask 6.1.1 (NER) Run Format

Runs must be submitted as a JSON file (`.json`) with the following structure:

```
{
    "34870091": {
        "entities": [
            {
                "start_idx": 75,
                "end_idx": 82,
                "location": "title",
                "text_span": "patients",
                "label": "human"
            },
            {
                "start_idx": 250,
                "end_idx": 270,
                "location": "abstract",
                "text_span": "intestinal microbiome",
                "label": "microbiome"
            }
        ]
    }
}
```

where:

- The top-level key (e.g. "`34870091`") is the PubMed ID of the document.
- `entities` is a list of entity objects.
- Each entity object represents a predicted entity and contains:
  - `start_idx` and `end_idx`: character offsets of the span,
  - `location`: "title" or "abstract",
  - `text_span`: the actual text,
  - `label`: the entity type.

4.1.2. Subtask 6.1.2 (NERD) Run Format

Submissions must be provided as a JSON file (`.json`) with the following structure:

```
{
    "34870091": {
        "entities": [
            {
                "start_idx": 75,
                "end_idx": 82,
                "location": "title",
                "text_span": "patients",
                "label": "human",
                "uri": "http://id.nlm.nih.gov/mesh/D010361"
            },
            {
                "start_idx": 250,
                "end_idx": 270,
                "location": "abstract",
                "text_span": "intestinal microbiome",
                "label": "microbiome",
                "uri": "http://id.nlm.nih.gov/mesh/D000069196"
            }
        ]
    }
}
```

Each entity object follows the same structure defined for NER, with the additional `uri` field specifying the concept-level link assigned to the entity mention.

4.1.3. Subtask 6.2.1 (M-RE) Run Format

Submissions must be provided as a JSON file (`.json`) with the following structure:

```
{
    "34870091": {
        "mention_level_relations": [
            {
                "subject_text_span": "intestinal microbiome",
                "subject_label": "microbiome",
                "predicate": "located in",
                "object_text_span": "patients",
                "object_label": "human"
            }
        ]
    }
}
```

where:

- The top-level key (e.g. "`34870091`") is the PubMed ID of the document.
- `mention_level_relations` is a list of relation objects.
- Each relation object represents a predicted mention-level relation and contains:
  - `subject_text_span`: the exact character sequence of the subject mention,
  - `subject_label`: the entity type of the subject mention,
  - `predicate`: the relation type between the subject and object,
  - `object_text_span`: the exact character sequence of the object mention,
  - `object_label`: the entity type of the object mention.

4.1.4. Subtask 6.2.2 (C-RE) Run Format

Submissions must be provided as a JSON file (`.json`) with the following structure:

```
{
    "34870091": {
        "concept_level_relations": [
            {
                "subject_uri": "http://id.nlm.nih.gov/mesh/D010361",
                "subject_label": "microbiome",
                "predicate": "located in",
                "object_uri": "http://id.nlm.nih.gov/mesh/D000069196",
                "object_label": "human"
            }
        ]
    }
}
```

where:

- The top-level key (e.g. "`34870091`") is the PubMed ID of the document.
- `concept_level_relations` is a list of relation objects.
- Each relation object represents a predicted concept-level relation and contains:
  - `subject_uri`: the URI of the concept corresponding to the subject entity mention,
  - `subject_label`: the entity type of the subject mention,
  - `predicate`: the relation type between the subject and object,
  - `object_uri`: the URI of the concept corresponding to the object entity mention,
  - `object_label`: the entity type of the object mention.

4.1.5. Submission Upload

All runs must be submitted as a single ZIP archive named `<teamID>_GutBrainIE_2026.zip`. Within this archive, each run must be placed in its own folder named `<teamID>_<taskID>_<runID>_<systemDesc>` (without spaces or special characters), where:

- `<teamID>` is the name of the participating team;
- `<taskID>` is the identifier of the subtask the run is being submitted to (one of `T611` for NER, `T612` for NERD, `T621` for M-RE, or `T622` for C-RE);
- `<runID>` is a unique alphanumeric string (a–z, A–Z, 0–9) chosen by the team to distinguish among their runs;
- `<systemDesc>` is an optional short label describing the system.

Each run folder is required to contain exactly two files:

- `<teamID>_<taskID>_<runID>_<systemDesc>.json`
- `<teamID>_<taskID>_<runID>_<systemDesc>.meta`

The `.json` file holds the team's predictions for the specified subtask on the test set. The accompanying `.meta` file must include the following information:

- Team ID, Task ID, and Run ID;
- Type of training applied;
- Pre-processing methods;
- Training data used;
- Relevant details of the run;
- A link to a public repository enabling reproducibility.

4.2. Participants

A total of 80 teams registered for the GutBrainIE-2026 task, of which 18 submitted at least one run and thus participated in the evaluation.

In total, 516 runs were submitted: 150 for NER, 129 for NERD, 119 for M-RE, and 118 for C-RE. Table 4 shows which tasks each team participated in and how many runs they submitted, while Table 5 reports their affiliations, countries of origin, and associated resources.

Table 5
Teams participating in GutBrainIE-2026.

Team ID	Affiliation	Country	Repository	Paper
AAU-Gut-Brain	Aalborg University	Denmark	–	–
ELiRF-UPV	Universitat Politècnica de València	Spain	–	–
GBA	University of Tübingen	Germany	–	–
GBIE	University of Bucharest	Romania	Anonymous repository	–
GetGut@AAU	Aalborg University	Denmark	GitHub	[14]
Graphwise	Graphwise	Bulgaria	Available upon request	[15]
GutHub	Aalborg University	Denmark	GitHub	[16]
MKTE	University of Padua	Italy	–	–
MindGut link	Universidad Nacional de Educación a Distancia	Spain	–	–
NightSun	University of Zurich	Switzerland	GitHub	[17]
NLP-DE	Technical University of Munich	Germany	–	–
NOVALINCS	Universidade Nova de Lisboa	Portugal	–	[18]
SMTE	University of Padua	Italy	GitHub	[19]
TEXA	University of Padua	Italy	–	[20]
ToGS	Graz University of Technology	Austria	GitHub	[21]
TUGW	Graz University of Technology Graphwise	Austria Bulgaria	GitHub	[21, 15]
TWIX	University of Padua	Italy	GitHub	[22]
unibuc-bionlp-bp	University of Bucharest	Romania	–	–

4.3. Evaluation

All submitted runs are evaluated using standard IE metrics of precision (P), recall (R), and F1-score (F_1), assessed with both macro- and micro-averaging. The same metrics apply to all four subtasks.

Let TP_ℓ, FP_ℓ, and FN_ℓ denote, respectively, the number of true positives, false positives, and false negatives for label ℓ. We define the label set $\mathscr{L}$ as:

- for subtasks 6.1.1 and 6.1.2 (NER and NERD): the set of entity types;
- for subtasks 6.2.1 and 6.2.2 (M-RE and C-RE): the set of triples *(subject label, predicate, object label)*.

The macro-averaged metrics are computed as:

$$P_{\text{macro}} = \frac{1}{|\mathscr{L}|} \sum_{\ell \in \mathscr{L}} \frac{TP_\ell}{TP_\ell + FP_\ell}, \tag{1a}$$

$$R_{\text{macro}} = \frac{1}{|\mathscr{L}|} \sum_{\ell \in \mathscr{L}} \frac{TP_\ell}{TP_\ell + FN_\ell}, \tag{1b}$$

$$F_{1,\text{macro}} = \frac{2\, P_{\text{macro}}\, R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}}. \tag{1c}$$

The micro-averaged metrics aggregate counts before division:

$$P_{\text{micro}} = \frac{\sum_{\ell \in \mathscr{L}} TP_\ell}{\sum_{\ell \in \mathscr{L}} (TP_\ell + FP_\ell)}, \tag{2a}$$

$$R_{\text{micro}} = \frac{\sum_{\ell \in \mathscr{L}} TP_\ell}{\sum_{\ell \in \mathscr{L}} (TP_\ell + FN_\ell)}, \tag{2b}$$

$$F_{1,\text{micro}} = \frac{2\, P_{\text{micro}}\, R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}. \tag{2c}$$

For each subtask, the micro-averaged F1-score (Eq. 2c) is adopted as the reference metric for the final leaderboard.

4.4. Baseline

To support participants and provide a reference for performance evaluation, we developed a baseline system covering all four GutBrainIE subtasks. The baseline extends the one released for the 2025 edition of the challenge, adapted to the expanded 2026 dataset by fine-tuning its components on the updated annotations [9]. Since the 2026 edition introduces subtasks involving concept-level links, we extended the baseline with an additional Named Entity Linking (NEL) module.

The baseline is organized into three independent components: a NER module based on GLiNER [23], a NEL module that combines exact matching and embedding-based semantic similarity in a hierarchical linking strategy, and a RE module based on ATLOP [13].

The NER module employs GLiNER, a bidirectional transformer encoder trained for instruction-based named entity recognition [23]. We used the NuNERZero checkpoint [24] and fine-tuned the model on the Gold and Silver portions of the training data, applying a confidence threshold of 0.6. After inference, we merged predicted entities having adjacent or overlapping spans.

The NEL module follows a three-stage hierarchical linking strategy. For each mention predicted by the upstream NER module, the system first searches for an exact match among the concept-linked annotations in the gold and silver folds, requiring both the text span and the predicted entity label to match. If no match is found, the search is repeated by considering only the text span and ignoring the entity label. If this second step also fails, the module relies on semantic similarity based on PubMedBERT embeddings.[3] In this final stage, the predicted mention is compared both against annotated text spans already linked to concepts and against the concept names available in the reference resources. The candidate obtaining the highest similarity score is retained, and its associated concept is assigned to the mention.

The RE module uses ATLOP, a document-level relation extraction model that employs localized context pooling and adaptive thresholding [13]. ATLOP receives the document text and the entities predicted by the NER module and predicts relational triples within each document. The resulting relations are filtered to exclude any relation not listed in Table 2. For fine-tuning, we used the Gold and Silver collections as manually annotated sets and the Bronze collection as the distantly supervised annotated set.

Table 6 reports, for each participating team, the number of submitted runs that surpassed the baseline system out of the total number of runs submitted for each subtask (considering the micro-averaged F1 score as the reference metric).

The code implementing the baseline system is available at the following GitHub repository: https://github.com/MMartinelli-hub/GutBrainIE_2026_Baseline.

5. Results

This section presents the performance results for each subtask, based on the evaluation metrics described in Section 4.3.

For each subtask, we report the leaderboard tables showing the best-performing run per team, ranked by micro-averaged F1 score. In the reported leaderboards, submissions from team *TWIX* are shown separately and are not considered for the official ranking, since the team includes members of the task organizing committee. We allowed TWIX to participate in the evaluation but excluded it from the ranked leaderboard to ensure a fair comparison among external participants. Accordingly, TWIX runs are reported at the bottom of each leaderboard, separated from the official ranking. Complete scores for every submitted run can be found in the appendix.

[3] https://huggingface.co/NeuML/pubmedbert-base-embeddings

Table 6
Number of submitted runs surpassing the baseline system. For each team and subtask, the table reports the number of submitted runs that achieved a higher micro-averaged F1 score than the baseline system out of the total number of runs submitted.

Team	Total Runs	NER	NERD	M-RE	C-RE
AAU-Gut-Brain	0/6 (0%)	0/4 (0%)	0/2 (0%)	–	–
ELiRF-UPV	0/11 (0%)	0/10 (0%)	0/1 (0%)	–	–
GBA	3/40 (7.5%)	0/10 (0%)	3/10 (30%)	0/10 (0%)	0/10 (0%)
GBIE	0/2 (0%)	0/2 (0%)	–	–	–
GetGut@AAU	12/12 (100%)	3/3 (100%)	3/3 (100%)	3/3 (100%)	3/3 (100%)
Graphwise	68/100 (68%)	24/25 (96%)	25/25 (100%)	6/25 (24%)	13/25 (52%)
GutHub	1/4 (25%)	0/1 (0%)	0/1 (0%)	1/1 (100%)	0/1 (0%)
MKTE	0/2 (0%)	–	0/1 (0%)	0/1 (0%)	–
MindGut link	2/8 (25%)	0/2 (0%)	2/2 (100%)	0/2 (0%)	0/2 (0%)
NightSun	100/100 (100%)	25/25 (100%)	25/25 (100%)	25/25 (100%)	25/25 (100%)
NLP-DE	1/4 (25%)	1/2 (50%)	0/2 (0%)	–	–
NOVALINCS	0/5 (0%)	0/5 (0%)	–	–	–
SMTE	3/5 (60%)	1/3 (33.3%)	–	2/2 (100%)	–
TEXA	3/4 (75%)	1/1 (100%)	1/1 (100%)	0/1 (0%)	1/1 (100%)
ToGS	0/98 (0%)	0/24 (0%)	0/24 (0%)	0/25 (0%)	0/25 (0%)
TUGW	2/18 (11.1%)	0/8 (0%)	2/8 (25%)	–	0/2 (0%)
TWIX	89/95 (93.7%)	17/23 (73.9%)	24/24 (100%)	24/24 (100%)	24/24 (100%)
unibuc-bionlp-bp	1/2 (50%)	1/2 (50%)	–	–	–

5.1. Subtask 6.1.1 (NER) Results

Most participating teams in the NER subtask adopted supervised approaches based on biomedical transformer encoders or GLiNER-style span models, with the most employed backbones including BiomedBERT, BioLinkBERT, BioBERT, SciBERT, BiomedELECTRA, and biomedical variants of GLiNER [25, 26, 27, 28, 29, 23]. The best-performing systems relied on supervised fine-tuning combined with well-designed preprocessing and decoding strategies. In particular, CRF layers, span-based decoding, overlap resolution, and label-aware post-processing were commonly used to improve performance.

While most systems used the manually curated gold and silver folds, some teams also included bronze annotations in their training through relabeling, filtering, or re-weighting strategies.

Ensemble approaches were also widely adopted, combining different biomedical encoders, random seeds, training subsets, or decoding thresholds. These approaches generally improved robustness, although some submissions showed a clear precision–recall trade-off, with stricter voting strategies producing very high precision at the cost of lower recall.

A smaller number of teams explored generative or more naive entity matching approaches. These systems were generally less competitive, tending to suffer from either low precision or low recall.

The leaderboard for the NER subtask is shown in Table 7.

5.2. Subtask 6.1.2 (NERD) Results

Most participating teams approached NERD as a pipeline composed of a NER module followed by a NEL component. The best-performing systems reused strong supervised NER models, mainly based on biomedical transformer encoders or GLiNER-style span models, and combined them with linking strategies based on training dictionaries, biomedical embeddings, concept definitions, and type constraints.

Dictionary-based matching was widely adopted as a first linking step, usually relying on mappings from mention text and entity label to URI extracted from the training data. When exact matching failed,

Table 7
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.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NightSun	run21	0.8226	0.7309	**0.7693**	0.8756	0.8110	**0.8420**
Graphwise	18	0.8114	0.7362	0.7665	0.8613	0.8169	0.8385
TEXA	1	0.7353	0.7529	0.7413	0.7980	0.8302	0.8138
unibuc-bionlp-bp	1	0.7174	**0.7679**	0.7390	0.7816	**0.8425**	0.8109
SMTE	R1	0.7367	0.7311	0.7302	0.8109	0.7932	0.8019
GetGut@AAU	1	0.7318	0.7212	0.7227	0.8011	0.8017	0.8014
NLP-DE	1	0.7150	0.7459	0.7276	0.7847	0.8173	0.8007
Organizers	BASELINE	0.7114	0.7480	0.7267	0.7782	0.8221	0.7996
MindGut link	BiomedBERT	0.7765	0.6733	0.7089	0.8292	0.7702	0.7986
NOVALINCS	run4	0.7086	0.6733	0.6848	0.8056	0.7850	0.7952
GutHub	1	0.7579	0.6015	0.6637	0.8704	0.7220	0.7893
GBIE	DuoHead3	**0.8392**	0.5719	0.6673	**0.8956**	0.6805	0.7734
AAU-Gut-Brain	SciBERT	0.6829	0.7072	0.6900	0.7451	0.7813	0.7628
GBA	25	0.6187	0.7638	0.6785	0.6535	0.8150	0.7254
TUGW	naiveentitiesGraphwiseT61113NEREXP8SUB1union	0.4765	0.7543	0.5753	0.5272	0.8414	0.6482
ELiRF-UPV	test6	0.7087	0.4467	0.5379	0.7637	0.4841	0.5926
ToGS	hermesragnaivebeamnonehermes318Bbase	0.3353	0.4506	0.3713	0.3496	0.5823	0.4369
TWIX	merged6	0.9242	0.7333	0.8114	0.9548	0.8058	0.8740

several teams introduced semantic retrieval based on SapBERT or related biomedical embedding models, often using FAISS indexes to retrieve candidate concepts [30, 31]. Some systems further refined candidate selection through type-compatibility priors, concept-definition encoders, cross-encoder rerankers, or voting across multiple linking models. Runs including bronze data improved performance by applying filtering or relabeling strategies.

Overall, the results indicate that NERD is substantially harder than NER, with URI assignment representing a strong bottleneck. The most effective systems combined domain-specific supervised NER with entity-linking modules based on dictionaries, semantic similarity, concept definitions, and type-aware candidate filtering.

Table 8 shows the leaderboard for the NERD subtask.

5.3. Subtask 6.2.1 (M-RE) Results

Most participating teams approached M-RE as a supervised pairwise classification task over candidate entity pairs produced by an upstream NER module. The strongest systems relied on biomedical transformer encoders, particularly on BiomedBERT and PubMedBERT, and explicitly marked subject and object mentions in the input text using entity markers [25, 32]. Candidate pairs were often filtered according to the set of legal entity-type combinations defined by the annotation schema.

Table 8
Performance metrics of each team's top run for NERD. For each evaluation metric, the best result is in bold, the second-best is underlined.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NightSun	run16	0.5376	**0.5107**	**0.5215**	0.6290	**0.6053**	**0.6169**
Graphwise	9	0.5235	0.4883	0.5029	0.6217	0.5897	0.6053
TUGW	naiveentitiesGraphwiseT612224ENSEXP4SUB1ELfreqgazeintersect	**0.6342**	0.3626	0.4470	**0.7723**	0.4437	0.5636
MindGut link	BiomedBERT	0.4814	0.4305	0.4484	0.5738	0.5330	0.5527
GetGut@AAU	1	0.4603	0.4534	0.4550	0.5515	0.5519	0.5517
GBA	5	0.3618	0.4353	0.3923	0.4385	0.5263	0.4784
TEXA	1	0.4039	0.4168	0.4090	0.4581	0.4766	0.4672
Organizers	BASELINE	0.3820	0.4045	0.3916	0.4281	0.4522	0.4398
MKTE	run1	0.3109	0.3117	0.2991	0.3974	0.3617	0.3787
GutHub	1	0.3815	0.3018	0.3338	0.3856	0.3199	0.3497
AAU-Gut-Brain	SciBERT	0.2697	0.2764	0.2723	0.3442	0.3477	0.3459
NLP-DE	1	0.2699	0.2848	0.2764	0.3260	0.3395	0.3326
ELiRF-UPV	test11	0.3324	0.2280	0.2675	0.4198	0.2706	0.3291
ToGS	naivefilteredentities	0.2530	0.2549	0.2429	0.3057	0.3091	0.3074
TWIX	24merged6	0.6695	0.5431	0.5965	0.7527	0.6353	0.6890

To address the strong imbalance between positive and negative relation instances, several teams employed negative sampling, hard negative mining, class-aware losses, or confidence threshold tuning. Some systems also restricted decoding to valid relation predicates, applied distance-aware filtering, or tuned per-predicate thresholds on the development set. These strategies proved effective in reducing false positives while preserving satisfactory recall.

In several cases, relation extraction relied on local context windows around the candidate entity pair rather than the full document, suggesting that focused contextual information can be more effective for mention-level classification. Ensemble strategies over different seeds, epochs, or model variants were also explored, although the gains were generally smaller than in the NER subtask.

Overall, M-RE results confirm that relation extraction remains substantially harder than entity recognition and entity linking. The most effective systems combined reliable entity predictions, supervised biomedical encoders, explicit entity-pair marking, type constraints, negative sampling, and threshold-based post-processing. The leaderboard for the M-RE subtask is shown in Table 9.

5.4. Subtask 6.2.2 (C-RE) Results

Most participating teams approached C-RE by reusing their M-RE pipelines and projecting the resulting mention-level predictions to concept-level tuples through a NEL module.

The best-performing systems combined strong M-RE classifiers with robust NERD components. Supervised biomedical encoders, especially BiomedBERT- and PubMedBERT-based pairwise classifiers with typed entity markers and negative sampling were reused by top teams and extended with URI assignment strategies based on dictionaries and SapBERT-style semantic similarity [25, 32, 30]. Several teams adopted direct projection strategies, converting predicted mention-level relations into concept-level relations using lookup dictionaries derived from training data.

Compared with M-RE, performance dropped substantially across almost all teams. This confirms what emerged in Section 5.2: concept-level normalization is a critical bottleneck for IE performance. Even systems with competitive M-RE results experienced large decreases in performance when exact

Table 9
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.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NightSun	run18	0.3773	**0.4025**	**0.3360**	0.4459	**0.4593**	**0.4525**
SMTE	R2	0.3459	0.3369	0.3084	0.4476	0.3997	0.4223
GetGut@AAU	1	**0.3961**	0.3167	0.3251	0.4546	0.3658	0.4054
GutHub	1	0.3111	0.2760	0.2751	0.4347	0.3754	0.4029
Graphwise	BGPT5515GEXP8	0.2583	0.3818	0.2868	0.3517	0.4497	0.3947
Organizers	BASELINE	0.3660	0.2862	0.3003	0.4444	0.3453	0.3886
TEXA	1	0.3804	0.2761	0.2947	**0.4548**	0.3382	0.3880
GBA	7	0.2318	0.1478	0.1467	0.2198	0.1621	0.1866
MindGut link	BiomedBERT	0.0725	0.1079	0.0775	0.1318	0.1749	0.1503
ToGS	hermeslorasbeamnonehermes318Bs	0.1111	0.0758	0.0649	0.1114	0.1166	0.1139
MKTE	run1	0.0813	0.1568	0.0895	0.0800	0.1365	0.1009
TWIX	mre10	0.6588	0.7308	0.6859	0.8996	0.4651	0.6132

Table 10
Performance metrics of each team's top run for C-RE. For each evaluation metric, the best result is in bold, the second-best is underlined.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NightSun	run18	**0.1995**	0.2397	**0.1889**	**0.2484**	0.2798	**0.2632**
Graphwise	BGPT5515GEXP4	0.1582	**0.2651**	0.1814	0.2092	**0.2963**	0.2452
GetGut@AAU	1	0.1481	0.1505	0.1385	0.2123	0.1926	0.2020
TEXA	1	0.1063	0.0991	0.0981	0.1556	0.1383	0.1464
Organizers	BASELINE	0.1009	0.1021	0.0966	0.1403	0.1292	0.1345
GutHub	1	0.1010	0.1100	0.0996	0.1305	0.1284	0.1295
GBA	25	0.1135	0.0742	0.0786	0.1717	0.0988	0.1254
MindGut link	BiomedBERT	0.0756	0.0974	0.0764	0.1042	0.1374	0.1186
ToGS	hermeslorasbeamnonehermes318Bs	0.0454	0.0347	0.0251	0.0598	0.0700	0.0645
TUGW	hermeslorasbeamnonehermes323BsGraphwiseT61113NEREXP8SUB1union	0.0283	0.0386	0.0266	0.0602	0.0683	0.0640
TWIX	11mre12	0.3699	0.4257	0.3902	0.4903	0.2691	0.3475

URI matching was introduced.

6. Discussion

This Section provides an overview of the approaches adopted by participating teams in the GutBrainIE-2026 task. We organize the discussion into four subsections, one for each subtask.

6.1. Subtask 6.1.1 (NER) Discussion

Team *NightSun* [17] fine-tuned BERT-based token classifiers augmented with a Conditional Random Field (CRF) layer over BIO-tagged sequences [33]. Preprocessing involved splitting articles at sentence boundaries into chunks of at most 512 tokens, processing titles and abstracts separately, aligning WordPiece tokenization with BIO labels, and merging predictions back through character offsets [34]. The team experimented with multiple biomedical transformer backbones, including BiomedBERT, BioLinkBERT, BioBERT, SciBERT, and BioClinicalModernBERT [25, 26, 27, 28, 35]. Several runs relied on three-seed ensembles and CRF parameter averaging to improve prediction stability. Training data included gold, silver, 2025 silver, and bronze annotations. In some configurations, bronze-quality annotations were first relabeled using a stronger model and then reused for training, aiming to reduce the noise introduced by weakly supervised data. The submitted runs also explored different voting strategies, with stricter unanimous voting favoring precision and more balanced configurations improving the final micro-averaged F1 score.

Team *Graphwise* [15] fine-tuned a biomedical GLiNER checkpoint `Ihor/gliner-biomed-large-v1.0`, using a confidence threshold of 0.7 [23, 36]. Training data included gold, silver, 2025 silver, and bronze annotations. To improve the quality of weakly supervised data, bronze annotations were re-annotated using an internal ensemble before being incorporated into training. The team also explored several BERT+CRF models based on BiomedBERT [33, 25]. These variants included annotation-quality-aware loss weighting, augmented training data, and model ensembling. Their best-performing configuration was based on the fine-tuned GLiNER model, which outperformed the submitted majority-vote ensemble runs. This suggests that, in their setup, a strong span-based biomedical model combined with threshold tuning and refined weak supervision was more effective than larger ensembles of token-classification models.

Team *TEXA* [20] fine-tuned a GLiNER-based model following a pipeline closely aligned with the organizers' baseline [23]. The system mirrored the baseline configuration in both pre-processing and post-processing steps, with training performed on the manually annotated folds of the dataset, as in the baseline system. Their submitted run achieved $F1_{\mu} = 0.8138$, slightly improving over the baseline system. Team *unibuc-bionlp-bp* similarly fine-tuned a GLiNER-based model, adopting a lower confidence threshold during inference. This configuration favored recall, leading to the highest recall among the best-per-team submissions ($R_{\mu} = 0.8425$).

Team *SMTE* [19] trained an ensemble of BioBERT and PubMedBERT token-classification models on gold, silver, and bronze annotations [27, 32]. Their pipeline included several post-processing steps to improve prediction consistency, including BIO-tag repair, two-pass thresholding, label-aware remapping, false-positive filtering, deduplication, and overlap pruning.

Team *GetGut@AAU* [14] fine-tuned a GLiNER-based model, adopting a sliding-window strategy with overlapping chunks to handle long inputs [23]. Titles and abstracts were processed at the section level, with offset tracking used to map predictions back to the original documents.

Team *NLP-DE* used a zero-shot GLiNER setup for entity extraction, using the NuNERZero pre-trained checkpoint, the same used by the organizers in their baseline system [23, 24].

Team *MindGut link* submitted two runs, exploring both token-classification and span-based approaches. The first run fine-tuned BiomedBERT as a token-classification model using gold and silver annotations [25]. Preprocessing involved converting annotations to BIO format, applying WordPiece tokenization, and using a sliding-window strategy to handle long input sequences. The second run fine-tuned GLiNER-BioMed for span-based NER, using the `gliner_large_bio-v0.1` checkpoint [37].

This configuration was also trained on the gold and silver annotated collections, with the confidence threshold fixed to 0.5.

Team *NOVALINCS* [18] employed GLiNER with the NuNERZero checkpoint, similarly to the organizers' baseline [23, 24]. Across runs, they experimented by fine-tuning with different combinations of training data and data augmentation strategies, including: replacing 15% of non-entity tokens with synonyms, replacing entity mentions with similar words based on word2vec embeddings, and combining both strategies.

Team *GutHub* [16] employed GLiNER and pre-trained biomedical transformers fine-tuned for token classification, combining their outputs through an original ensembling strategy. The base models included PubMedBERT, BioLinkBERT, BioELECTRA, BioClinical ModernBERT, OpenMed-NER, and GLiNER [32, 26, 29, 35, 38, 23]. Predictions were obtained through a stacking ensemble based on an XGBoost meta-classifier [39]. Instead of majority voting, model outputs were encoded as confidence-scaled one-hot features, allowing the ensemble to learn the strengths and weaknesses of each base model. Post-processing included dynamic boundary correction to remove whitespace artifacts introduced by subword tokenization and recover character offsets aligned with the original text.

Team *GBIE* proposed a dual-head NER architecture based on BioClinical-ModernBERT [35]. The model decomposes NER into boundary detection and span classification. The boundary head predicts entity start and end positions and learns to pair them, while the span classification head assigns candidate spans to one of the 13 entity labels or to a non-entity class. Training was organized in three stages: first, the LoRA encoder and boundary head were trained on the bronze set; then, the span classification head was trained on bronze annotations and hard negatives; finally, the full model was jointly fine-tuned on gold and silver annotations. Preprocessing concatenated titles and abstracts without cleaning or removing HTML tags, while post-processing realigned offsets, removed punctuation artifacts introduced by tokenization, and applied confidence-based filtering. One submitted run further used class-specific thresholds to balance precision and recall across entity categories.

Team *AAU-Gut-Brain* submitted runs based on training spaCy and scispaCy models [40, 41]. Preprocessing converted the challenge annotations into the format required by their training pipeline. Training used all available annotations, independently of their quality tier.

Team *GBA* submitted runs based on fine-tuning GLiNER2 models [42]. Most runs used LoRA fine-tuning, while one configuration applied full GLiNER2 fine-tuning. Training data included weighted combinations of gold, silver, silver 2025, bronze, and, in some runs, development annotations. Preprocessing was limited to reshaping the data into the expected input format. Some runs incorporated a GraphLinker component trained from scratch, with variants differing in the training data used and embedding configuration.

Team *ELiRF-UPV* submitted runs based on fine-tuning transformer models for token classification. The team experimented with DeBERTa-v3-large, BioLinkBERT-large, BiomedBERT base and large variants, and BiomedELECTRA [43, 26, 25, 29]. Several configurations further added a BiLSTM layer on top of the transformer encoder. Preprocessing relied on the format conversion provided by the organizers, with annotations converted to BIO labels [34]. All runs were trained on the gold, silver, and 2025 silver collections.

Team *ToGS* [21] adopted an IE pipeline based on grammar-constrained generative IE. Preprocessing split titles and abstracts into sentences, recalculating offsets before and after prediction. Named entities were first detected using spaCy and fine-tuned token-level models, including DistilBERT and PubMedBERT-MNLI-MedNLI, with the latter model used to guide downstream grammar construction [40, 44, 45]. The generated grammars were built from the ontology grounding the annotation schema provided by the organizers and the detected entity candidates, constraining the language model output to valid entity structures. The team experimented with RAG, LoRA fine-tuning of Hermes 3 models, and constrained beam-search decoding. Final outputs were parsed with regular expressions and realigned to the original document offsets.

Team *TUGW* submitted collaborative runs, combining the runs of teams *ToGS* and *Graphwise* using either union or intersection strategies [21, 15].

Team *TWIX* [22] proposed a two-stage approach using fine-tuned biomedical transformer models.

Their approach decomposed NER into two stages: Term Extraction, framed as BIO token classification with a single generic `term` label, and Term Classification, framed as sequence classification over candidate spans marked with `[E1]` and `[/E1]`. Candidate spans classified as `not an entity` were discarded. The team experimented with seven biomedical transformers, including BioLinkBERT, BiomedBERT, and BiomedELECTRA variants [26, 25, 29].

6.2. Subtask 6.1.2 (NERD) Discussion

All participants approached NERD as a pipeline in which entity mentions are first predicted by an upstream NER module and then linked to concept URIs. Therefore, in the following we focus only on the NEL component of the submitted systems, since the NER modules correspond to those already discussed in Section 6.1.

Team *NightSun* [17] employed a three-stage linking pipeline combining dictionary matching, dense retrieval, and reranking. They built a unified knowledge base covering all 13 entity types, integrating aliases from biomedical and domain-specific resources, together with URI mappings extracted from the training and development annotations. Mentions were first resolved through exact matching over normalized surface-form/type pairs. Unmatched mentions were encoded with SapBERT and linked by retrieving candidate URIs from a FAISS index [30, 31]. Retrieved candidates were finally reranked by combining embedding similarity with a type-conditioned frequency prior, mainly to resolve type-ambiguous mentions.

Team *Graphwise* [15] used a two-stage entity linking approach combining gazetteers and embedding-based retrieval. The team first applied two frequency-based gazetteers built from training and development annotations. The first matched mentions by both text span and entity category, while the second relied only on the text span, returning in both cases the most frequent URI observed in the training data. Out-of-vocabulary mentions were resolved through dense retrieval over the task knowledge base. The team experimented with PubMedBERT and a fine-tuned SapBERT model trained using URI labels and alternatives generated with a Large Language Model (LLM) [32, 30]. At inference time, mentions were embedded and matched against category-specific vector indices, with ensemble variants comparing candidates returned by different embedding models.

Team *MindGut* proposed a three-stage dictionary-based entity linking pipeline similar to the one implemented by the organizers' baseline. Mentions were first linked using exact dictionary matching on gold and silver annotations, then through a label-based fallback strategy, and finally using SapBERT for unresolved cases [30].

Team *GetGut@AAU* [14] employed a lightweight dictionary-based linking approach relying on contextual embeddings. The team built a reference dictionary mapping entities to URIs by pairing each entity mention with the sentence in which it appears. These context-enriched strings were encoded with SapBERT using SentenceTransformers to obtain fixed-size embeddings [30]. At inference time, each predicted entity and its surrounding sentence were encoded separately and compared against the dictionary through cosine similarity. The final URI assignment was based on a hybrid score combining entity-level and context-level similarity, selecting the top-ranked URI as prediction.

Team *GBA* used a custom graph-based linker trained from scratch. This model embeds entity mentions and URI concepts into a shared space, enriching URI representations through a graph neural network over the concept graph, and selects the URI with the highest similarity to the mention embedding. Across runs, the team experimented with different entity embedding backbones.

Team *TEXA* [20] proposed a system that closely mirrors the three-stage linking pipeline implemented in the organizers' baseline.

Team *MKTE* implemented a dictionary-based entity linking approach based on a knowledge base built from gold training annotations by mapping each span-label pair to its corresponding URI. At inference time, mentions were first resolved through case-insensitive exact matching and then through partial substring matching restricted to the same entity label. Unresolved mentions were assigned a default ontology-level URI according to their entity type.

Team *GutHub* [16] approached the NERD subtask with a two-stage linking architecture based on

dense retrieval and reranking. The team first built a retrieval index from URI-associated text spans and synonyms collected from the training annotations and linked resources. These entries were encoded with SapBERT and stored in a FAISS index for efficient candidate retrieval [30, 31]. At inference time, each entity mention was encoded with SapBERT to retrieve the top candidate URIs. A cross-encoder reranker then scored each mention-candidate pair, selecting the highest-scoring URI as the final prediction.

Team *AAU-Gut-Brain* employed a similarity-based entity linking approach. They fine-tuned a SciBERT sentence encoder using definition pairs associated with the same KB concept and Multiple Negatives Ranking Loss [28]. The resulting model embedded mentions and KB definitions into a shared vector space, assigning URIs through nearest-neighbor retrieval.

Teams *NLP-DE* and *ELiRF-UPV* complemented entity mentions extracted by their NER module with concept-level links using the baseline system provided by the organizers.

Team *ToGS* [21] addressed entity linking with a simple matching-based URI assignment strategy. After parsing the extracted entities, the system assigned URIs through direct matching against the collection of known entity text–URI mappings, similarly to what is done by the baseline system in the first two stages. To improve coverage for entities not found through exact matching, the team additionally used TF-IDF similarity to retrieve close textual matches and assign the corresponding URI.

Team *TWIX* [22] addressed the NERD subtask with a two-stage pipeline composed of candidate retrieval and reranking. Given an entity mention and its context, the retriever first selected the top candidate concepts from the reference vocabularies using dense vector similarity through a dual-encoder architecture. Retrieved candidates were then processed by a cross-encoder reranker, which jointly encoded each mention-candidate pair and selected the highest-scoring concept as final URI. Employed models included BioLinkBERT-base and BiomedBERT-base for both stages [26, 25].

Team *TUGW* submitted collaborative runs, combining the runs of teams *ToGS* and *Graphwise* using either union or intersection strategies [21, 15].

6.3. Subtask 6.2.1 (M-RE) Discussion

All submitted M-RE systems predict relations over entity mentions produced by an upstream NER component. Therefore, in the following, we focus solely on the RE modules adopted by participants, while the corresponding NER components are those already discussed in Section 6.1.

Team *NightSun* [17] participated with a BiomedBERT-base classifier fine-tuned for relation extraction [25]. Candidate pairs were encoded by inserting typed markers around subject and object mentions, and relation labels were predicted from the concatenated hidden states of the marker tokens. Training used positive relations from the 2026 gold data and 2025 silver annotations, while negative examples were sampled from type-valid entity pairs within the same document. To address class imbalance, the team combined cross-entropy and Dice loss, testing different loss weights and random seeds. For long documents, they adopted windowed tokenization centered on the entity markers, ensuring that both candidate entities remained visible within the 512-token input limit whenever possible.

Team *SMTE* [19] framed mention-level relation extraction as pairwise classification over predicted entity mentions. Candidate pairs compatible with the official relation schema were classified into one of the 17 relation predicates or a no-relation class. Their submitted runs fine-tuned PubMedBERT-large using entity markers and mention-mean pooling, where the token representations inside each marked entity span were averaged and concatenated before classification [32]. The two runs differed only in the input context: one used the full title–abstract text, while the other used a local window around the candidate entity pair. Training relied on hard negative sampling to address the large number of no-relation pairs. At inference time, candidate pairs were generated from the team's NER ensemble predictions, decoded under schema constraints, filtered with distance-aware and predicate-specific thresholds, and deduplicated by retaining the highest-scoring prediction.

Team *GetGut@AAU* [14] fine-tuned BiomedBERT-large for relation extraction, framing the task as sequence classification over marked entity pairs [25]. Subject and object mentions were enclosed with special markers and classified into relation predicates or a no-relation class. Training followed a

curriculum over annotation-quality tiers, progressing from bronze to silver and gold data, with loss weights reflecting corpus reliability. Negative examples were generated through controlled sampling of unrelated entity pairs. At inference time, NLTK sentence splitting was used to build sentence-level candidate pairs, inject entity markers, classify relations, and filter out no-relation predictions [46].

A similar framing has been employed by team GutHub [16]. For each document, candidate pairs were generated from predicted entities and marked with `[E1]/[E2]` tokens before classification. To reduce false positives, inference used class-specific confidence thresholds and defaulted uncertain predictions to no relation. The team further applied deterministic post-processing rules to remove invalid relations, including self-relations and relations between overlapping or nested entity mentions. Class imbalance was handled by filtering candidate pairs through the official annotation schema, applying distance-based negative sampling, and weighting positive relation classes more strongly than the no-relation class during training.

Team *Graphwise* [15] addressed relation extraction mainly by extending the ATLOP-based architecture adopted in the baseline system [13]. They experimented with several biomedical and general-domain encoders, including BERT, RoBERTa, PubMedBERT, BiomedELECTRA, BiomedBERT, MedCPT, and BioLinkBERT, training on the gold, silver, silver-2025, and bronze sets [47, 48, 32, 29, 25, 49, 26]. The team also explored parallel encoder ensembles by combining models sharing the same tokenizer and aggregating their logits during training. In addition, the team tested LLM-based approaches, including structured multi-shot prompting with OpenAI models and supervised fine-tuning with LoRA adapters on Qwen3.5-4B [50, 51]. These methods used the document text and upstream entity predictions as input, produced JSON-formatted relation outputs, and applied schema-based post-processing to remove invalid or duplicated predictions.

Also team *TEXA* [20] used an ATLOP-based system for relation extraction [13]. They used the same pre-processing and training strategies employed in the baseline system, using manually annotated folds (gold and silver) as strong supervision signals and automatically annotated data (bronze) as weak supervision for training. At inference time, relations were generated from predicted entities and filtered for validity with respect to the official annotation schema provided by the organizers.

Team *GBA* used their custom GraphLinker module trained from scratch, similarly to what they did for NERD.

Team *MindGut link* framed relation extraction as a classification task over marked entity pairs. Entity mentions were represented using inline markers and contextual hidden states, and the model was fine-tuned using only gold annotations. At inference time, predictions were selected using a fixed confidence threshold and then post-processed through schema-based filtering and deduplication.

Team *ToGS* [21] used the same constrained generative pipeline for relation extraction, in a manner similar to that for NER and NERD. Documents were split into sentences, entity offsets were adjusted, and relation-specific grammars were generated from the ontology and the entities detected upstream.

Team *MKTE* addressed M-RE with a feature-based LinearSVC classifier. Candidate pairs were first filtered using the official annotation schema and distance constraints. For each remaining pair, the team extracted numerical features capturing structural properties, distance, negation, and schema constraints, together with TF-IDF features computed over masked entity-pair contexts. The classifier was trained on gold and silver data with balanced class weights and sampled negative pairs. At inference time, entity pairs were generated from the team's NER predictions, classified, and predictions labeled as no relation were discarded.

Team *TWIX* [22] used an ATLOP-based relation extraction module with RoBERTa-large as the backbone [48]. They employed a two-stage training strategy: first, pre-training on silver or bronze+silver annotations, and then fine-tuning on gold annotations.

6.4. Subtask 6.2.2 (C-RE) Discussion

For C-RE, participants did not introduce substantially different architectures specifically tailored to the concept-level setting. Submitted systems generally reused the same RE modules adopted for M-RE, replacing upstream NER outputs with NERD predictions or projecting mention-level relations to

concept-level tuples through entity linking. Therefore, we do not provide a separate detailed discussion for this subtask, as it would largely repeat the methodologies already described for M-RE.

7. Conclusions and Future Work

GutBrainIE-2026 marked the second edition of a shared task dedicated to Information Extraction on the gut–brain axis, a research area of growing relevance in both neuroscience and microbiology.

This edition saw 80 teams registering and 18 teams submitting a total of 516 runs. Participants tackled subtasks concerning entity recognition, with Named Entity Recognition and Named Entity Recognition and Disambiguation, and relation extraction, with M-RE and C-RE. The results obtained by submitted runs highlight the effectiveness of supervised fine-tuning methods based on biomedical transformers for both entity recognition and relation extraction, while also showing that concept normalization remains the main bottleneck for end-to-end IE systems.

The released dataset, including over 5,000 annotated PubMed abstracts and stratified into annotation quality tiers, represents a valuable resource for training and evaluating biomedical NLP systems. Expert-curated annotations proved to be the most effective training data for achieving top-performing results across all four subtasks. At the same time, the mid-quality annotations manually curated by trained non-experts proved beneficial for improving system performance, whereas automatic annotations were useful only when combined with relabeling, filtering, or other noise-reduction strategies.

As future work, we plan to further improve the overall quality of the dataset by manually reviewing and annotating the current bronze fold, which is currently composed of fully automatic and non-revised annotations. We will also retrieve and annotate additional PubMed documents to increase the scope of our collection. Finally, we aim to extend the task by proposing each subtask in two complementary settings: the current end-to-end setting and a more standard controlled setting. In the latter, ground-truth entity mentions will be provided for NERD and M-RE, while ground-truth entity mentions together with concept-level links will be provided for C-RE. This will make it possible to evaluate systems both in a realistic end-to-end IE scenario and in a setting focused on the specific subtask, isolating the impact of upstream components such as entity recognition and concept normalization.

Acknowledgments

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.

Declaration on Generative AI

During the preparation of this work, the author used GPT-5.5 and Grammarly for rewriting, grammar checking, and spelling correction. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] M. Carabotti, A. Scirocco, M. A. Maselli, C. Severi, 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 (2015) 203.

[2] S. Ghaisas, J. Maher, A. Kanthasamy, 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 (2016) 52–62.

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

[4] J. F. Cryan, K. J. O'Riordan, K. Sandhu, V. Peterson, T. G. Dinan, The gut microbiome in neurological disorders, The Lancet Neurology 19 (2020) 179–194.

[5] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodriguez-Ortega, E. Rodriguez-Lopez, N. Loukachevitch, I. Rozhkov, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, D. Dimitriadis, A. Bekiaridou, A. Samaras, V. Patsiou, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, BioASQ at CLEF2026: The Fourteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2026, pp. 315–324. doi:10.1007/978-3-032-21321-1_42.

[6] M. Martinelli, S. Marchesin, V. Bonato, G. Di Nunzio, N. Ferro, O. Irrera, L. Menotti, F. Vezzani, G. Silvello, A Domain-Specific Curated Benchmark for Entity and Document-Level Relation Extraction, in: Findings of the Association for Computational Linguistics: EACL 2026, 2026, pp. 5693–5711.

[7] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. R. Ortega, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, et al., Bioasq at clef2025: The thirteenth edition of the large-scale biomedical semantic indexing and question answering challenge, in: European Conference on Information Retrieval, Springer, 2025, pp. 407–415.

[8] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, volume TBA of *Lecture Notes in Computer Science*, Springer, 2025, p. TBA.

[9] M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[10] M. Martinelli, S. Marchesin, G. Silvello, Efficient and reliable estimation of named entity linking quality: A case study on gutbrainie, 2026. URL: https://arxiv.org/abs/2601.06624. arXiv:2601.06624.

[11] L. R. Andersen, M. H. Dolmer, M. I. Gardshodn, J. M. Rodriguez, D. Dell'Aglio, Trusting gut instincts: Transformer-based extraction of structured data from gut-brain axis publications, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 99–119. URL: https://ceur-ws.org/Vol-4038/paper_6.pdf.

[12] V. Sanh, T. Wolf, Y. Belinkov, A. M. Rush, Learning from others mistakes: Avoiding dataset biases without modeling them, arXiv preprint arXiv:2012.01300 (2020).

[13] W. Zhou, K. Huang, T. Ma, J. Huang, Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[14] S. Jørgensen, E. Garavito Molina, J. M. Rodriguez, D. Dell'Aglio, GetGut: A Multi-Stage Pipeline for Gut-Brain Axis Information Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[15] I. Nikolova, M. Kuzmanov, A. Datseris, M. Barouh, R. Costa, C. Barbero, M. Canelas, M. Ramos, You Don't Need a Frontier LLM: Domain-Specific Encoders can compete in Gut–Brain Information Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[16] A. C. Rydder, S. B. Christensen, D. Austys, J. M. Rodriguez, D. Dell'Aglio, GutHub: Tackling the GutBrainIE Task Through Entity recognition, Disambiguation, and Relation Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[17] Y. Liu, Z. Niu, C. Li, NightSun at CLEF2026 GutBrainIE: A First-Place System for NER, Entity Linking, and Relation Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[18] M. Lourenço, A. Lamurias, Exploring Data Augmentation Strategies for Biomedical NER: A Study on the GutBrainIE 2026 Shared Task, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[19] S. Maule, G. M. Di Nunzio, SMTE – GutBrainIE@CLEF 2026: Biomedical Named Entity Recognition and Mention-Level Relation Extraction for the Gut-Brain Axis, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[20] D. Smirnov, B. Khashechian, G. M. Di Nunzio, TEXA at GutBrainIE 2026: A Multi-Stage Approach for Entity Recognition, Linking, and Relation Extraction in Gut-Brain Biomedical Abstracts, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[21] B. Kantz, P. Waldert, S. Lengauer, T. Schreck, CHASTE: Constrained Holistic Annotator System with onTological Entities, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[22] M. Martinelli, L. Menotti, TWIX: a Two-Stage Approach for End-To-End Named Entity Recognition and Relation Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.
[23] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5364–5376. URL: https://aclanthology.org/2024.naacl-long.300/. doi:10.18653/v1/2024.naacl-long.300.
[24] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé, E. Bernard, NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data, 2024. arXiv:2402.15343.
[25] S. Chakraborty, E. Bisong, S. Bhatt, T. Wagner, R. Elliott, F. Mosconi, BioMedBERT: A pre-trained biomedical language model for QA and IR, in: Proceedings of the 28th international conference on computational linguistics, 2020, pp. 669–679.
[26] M. Yasunaga, J. Leskovec, P. Liang, LinkBERT: Pretraining Language Models with Document Links, in: Association for Computational Linguistics (ACL), 2022.
[27] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: pre-trained biomedical language representation model for biomedical text mining, arXiv preprint arXiv:1901.08746 (2019).
[28] I. Beltagy, K. Lo, A. Cohan, SciBERT: Pretrained Language Model for Scientific Text, in: EMNLP, 2019. arXiv:arXiv:1903.10676.
[29] S. Alrowili, V. Shanker, BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 221–227. URL: https://www.aclweb.org/anthology/2021.bionlp-1.24.
[30] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4228–4238. URL: https://aclanthology.org/2021.naacl-main.334. doi:10.18653/v1/2021.naacl-main.334.
[31] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library, arXiv preprint arXiv:2401.08281 (2024).
[32] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.
[33] C. Sutton, A. McCallum, et al., An introduction to conditional random fields, Foundations and Trends® in Machine Learning 4 (2012) 267–373.
[34] L. Ramshaw, M. Marcus, Text Chunking using Transformation-Based Learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107/.

[35] T. Sounack, J. Davis, B. Durieux, A. Chaffin, T. J. Pollard, E. Lehman, A. E. W. Johnson, M. McDermott, T. Naumann, C. Lindvall, Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp, 2025. URL: https://arxiv.org/abs/2506.10896. arXiv:2506.10896.

[36] A. Yazdani, I. Stepanov, D. Teodoro, Gliner-biomed: A suite of efficient models for open biomedical named entity recognition, 2025. URL: https://arxiv.org/abs/2504.00676. arXiv:2504.00676.

[37] A. Yazdani, I. Stepanov, D. Teodoro, GLiNER-biomed: A Suite of Efficient Models for Open Biomedical Named Entity Recognition, 2025. URL: https://arxiv.org/abs/2504.00676. arXiv:2504.00676.

[38] M. Panahi, Openmed ner: Open-source, domain-adapted state-of-the-art transformers for biomedical ner across 12 public datasets, 2025. URL: https://arxiv.org/abs/2508.01630. arXiv:2508.01630.

[39] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 785–794. URL: https://doi.org/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.

[40] Y. Vasiliev, Natural language processing with Python and spaCy: A practical introduction, No Starch Press, 2020.

[41] M. Neumann, D. King, I. Beltagy, W. Ammar, ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 319–327. URL: https://www.aclweb.org/anthology/W19-5034. doi:10.18653/v1/W19-5034. arXiv:arXiv:1902.07669.

[42] U. Zaratiana, G. Pasternak, O. Boyd, G. Hurn-Maloney, A. Lewis, GLiNER2: Schema-driven multi-task learning for structured information extraction, in: I. Habernal, P. Schulam, J. Tiedemann (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Suzhou, China, 2025, pp. 130–140. URL: https://aclanthology.org/2025.emnlp-demos.10/.

[43] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. arXiv:2111.09543.

[44] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).

[45] P. Deka, A. Jurek-Loughrey, D. P, Multiple evidence combination for fact-checking of health-related information, in: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 237–247. URL: https://aclanthology.org/2023.bionlp-1.20.

[46] S. Bird, E. Loper, NLTK: The natural language toolkit, in: Proceedings of the ACL Interactive Poster and Demonstration Sessions, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 214–217. URL: https://aclanthology.org/P04-3031/.

[47] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[48] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[49] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval, Bioinformatics 39 (2023) btad651.

[50] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models, ICLR 1 (2022) 3.

[51] Qwen Team, Qwen3.5: Towards native multimodal agents, 2026. URL: https://qwen.ai/blog?id=qwen3.5.

A. Subtask 6.1.1 (NER) Overall Results

Table 11: Performance metrics of each team's submitted runs for NER. For each evaluation metric, the best result is in bold, the second-best is underlined.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Organizers	BASELINE	0.7114	0.7480	0.7267	0.7782	0.8221	0.7996
AAU-Gut-Brain	SciBERT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AAU-Gut-Brain	SciBERT	0.6829	0.7072	0.6900	0.7451	0.7813	0.7628
AAU-Gut-Brain	roBERTa-biomed	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AAU-Gut-Brain	roBERTa-biomed	0.6821	0.7006	0.6887	0.7545	0.7620	0.7583
ELiRF-UPV	test1	0.6618	0.4438	0.5258	0.7412	0.4778	0.5810
ELiRF-UPV	test10	0.6830	0.4414	0.5325	0.7529	0.4789	0.5854
ELiRF-UPV	test2	0.6903	0.4396	0.5304	0.7557	0.4759	0.5840
ELiRF-UPV	test3	0.7037	0.4361	0.5345	0.7803	0.4752	0.5906
ELiRF-UPV	test4	0.6864	0.4392	0.5293	0.7646	0.4815	0.5909
ELiRF-UPV	test5	0.7113	0.4369	0.5339	0.7639	0.4737	0.5848
ELiRF-UPV	test6	0.7087	0.4467	0.5379	0.7637	0.4841	0.5926
ELiRF-UPV	test7	0.7087	0.4467	0.5379	0.7637	0.4841	0.5926
ELiRF-UPV	test8	0.6627	0.4393	0.5228	0.7473	0.4800	0.5845
ELiRF-UPV	test9	0.6865	0.4315	0.5243	0.7659	0.4778	0.5885
GBA	0	0.6233	0.7389	0.6713	0.6636	0.7954	0.7235
GBA	1	0.6241	0.7155	0.6570	0.6619	0.7843	0.7179
GBA	25	0.6187	0.7638	0.6785	0.6535	0.8150	0.7254
GBA	2	0.6428	0.6314	0.6243	0.6852	0.7365	0.7099
GBA	3	0.6233	0.7389	0.6713	0.6636	0.7954	0.7235
GBA	4	0.6233	0.7389	0.6713	0.6636	0.7954	0.7235
GBA	5	0.6198	0.7273	0.6638	0.6621	0.7947	0.7224
GBA	6	0.6241	0.7155	0.6570	0.6619	0.7843	0.7179
GBA	7	0.6043	0.7172	0.6500	0.6443	0.7895	0.7095
GBA	8	0.6043	0.7172	0.6500	0.6443	0.7895	0.7095
GBIE	DuoHead2	0.8261	0.5719	0.6704	0.9114	0.6138	0.7336
GBIE	DuoHead3	0.8392	0.5719	0.6673	0.8956	0.6805	0.7734
GetGut@AAU	1	0.7318	0.7212	0.7227	0.8011	0.8017	0.8014
GetGut@AAU	2	0.7318	0.7212	0.7227	0.8011	0.8017	0.8014
GetGut@AAU	3	0.7318	0.7212	0.7227	0.8011	0.8017	0.8014
Graphwise	1	0.7536	0.7340	0.7391	0.8195	0.8176	0.8186
Graphwise	2	0.7205	0.7469	0.7314	0.7958	0.8217	0.8085
Graphwise	3	0.7407	0.7251	0.7314	0.8238	0.8095	0.8166
Graphwise	4	0.7551	0.7456	0.7461	0.8172	0.8251	0.8211
Graphwise	5	0.7363	0.7463	0.7393	0.8051	0.8176	0.8113
Graphwise	6	0.7377	0.7015	0.7150	0.8203	0.7969	0.8084
Graphwise	7	0.7292	0.7063	0.7101	0.8029	0.8125	0.8077
Graphwise	8	0.6955	0.7284	0.7096	0.8126	0.8247	0.8186
Graphwise	9	0.7611	0.7472	0.7504	0.8265	0.8265	0.8265
Graphwise	10	0.7684	0.7409	0.7461	0.8138	0.8277	0.8207
Graphwise	11	0.7700	0.7270	0.7427	0.8266	0.8165	0.8216
Graphwise	12	0.7349	0.7469	0.7361	0.7972	0.8232	0.8100
Graphwise	13	0.7533	0.7399	0.7407	0.8157	0.8284	0.8220
Graphwise	14	0.7597	0.7474	0.7488	0.8164	0.8206	0.8185
Graphwise	15	0.7567	0.7522	0.7473	0.8148	0.8302	0.8225
Graphwise	16	0.7529	0.7599	0.7544	0.8220	0.8317	0.8268
Graphwise	17	0.7064	0.7289	0.7151	0.7809	0.8136	0.7969
Graphwise	18	0.8114	0.7362	0.7665	0.8613	0.8169	0.8385
Graphwise	19	0.7707	0.7586	0.7619	0.8333	0.8336	0.8334
Graphwise	20	0.7529	0.7605	0.7517	0.8196	0.8384	0.8289
Graphwise	21	0.7917	0.7129	0.7433	0.8607	0.8106	0.8349
Graphwise	22	0.7665	0.7369	0.7480	0.8340	0.8232	0.8286
Graphwise	23	0.7646	0.7306	0.7432	0.8358	0.8150	0.8253
Graphwise	24	0.7855	0.7336	0.7547	0.8449	0.8176	0.8310
Graphwise	25	0.7834	0.7399	0.7571	0.8420	0.8180	0.8299
GutHub	1	0.7579	0.6015	0.6637	0.8704	0.7220	0.7893
MindGut link	BiomedBERT	0.7765	0.6733	0.7089	0.8292	0.7702	0.7986
MindGut link	GlinerFlanT5	0.7765	0.6733	0.7089	0.8292	0.7702	0.7986
NightSun	run01	0.7782	0.7511	0.7600	0.8396	0.8302	0.8349
NightSun	run02	0.7754	**0.7759**	0.7725	0.8311	0.8410	0.8360
NightSun	run03	0.7431	0.7364	0.7361	0.8276	0.8202	0.8239
NightSun	run04	0.7833	0.7699	0.7722	0.8390	0.8380	0.8385
NightSun	run05	0.7578	0.7613	0.7559	0.8243	0.8310	0.8276
NightSun	run06	0.7813	0.7549	0.7635	0.8418	0.8321	0.8369
NightSun	run07	0.7849	0.7584	0.7669	0.8438	0.8332	0.8385
NightSun	run08	0.7824	0.7636	0.7685	0.8399	0.8340	0.8369
NightSun	run09	0.8064	0.7299	0.7611	0.8656	0.8143	0.8392
NightSun	run10	0.8276	0.6912	0.7459	0.8928	0.7872	0.8367
NightSun	run11	0.7217	0.7110	0.7137	0.8115	0.8076	0.8096
NightSun	run12	0.7367	0.6937	0.7081	0.8216	0.7869	0.8039
NightSun	run13	0.7330	0.7597	0.7427	0.8138	0.8373	0.8254
NightSun	run14	0.7722	0.7220	0.7427	0.8289	0.8150	0.8219

(Table 11 continued)

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NightSun	run15	0.7512	0.6772	0.7025	0.8241	0.7898	0.8066
NightSun	run16	0.7644	0.7732	0.7655	0.8267	**0.8432**	0.8349
NightSun	run17	0.8247	0.6947	0.7468	0.8877	0.7910	0.8365
NightSun	run18	0.8156	0.7022	0.7475	0.8805	0.7976	0.8370
NightSun	run19	0.7871	0.7701	0.7744	0.8406	0.8388	0.8397
NightSun	run20	0.7824	0.7600	0.7667	0.8404	0.8354	0.8379
NightSun	run21	0.8226	0.7309	0.7693	0.8756	0.8110	0.8420
NightSun	run22	0.7996	0.7464	0.7677	0.8571	0.8247	0.8406
NightSun	run23	0.7744	0.7638	0.7650	0.8337	0.8380	0.8359
NightSun	run24	0.7825	0.7635	0.7686	0.8395	0.8336	0.8365
NightSun	run25	0.7801	0.7628	0.7669	0.8387	0.8347	0.8367
NLP-DE	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
NLP-DE	1	0.7150	0.7459	0.7276	0.7847	0.8173	0.8007
NOVALINCS	run1	0.7194	0.6205	0.6526	0.7935	0.7220	0.7561
NOVALINCS	run2	0.7175	0.6839	0.6964	0.8033	0.7858	0.7945
NOVALINCS	run3	0.7085	0.6997	0.7025	0.7911	0.7746	0.7828
NOVALINCS	run4	0.7086	0.6733	0.6848	0.8056	0.7850	0.7952
NOVALINCS	run5	0.7305	0.6444	0.6738	0.8163	0.7706	0.7928
SMTE	R1	0.7367	0.7311	0.7302	0.8109	0.7932	0.8019
SMTE	R2	0.7667	0.6985	0.7270	0.8415	0.7554	0.7961
SMTE	R3	0.7290	0.6983	0.7101	0.8175	0.7739	0.7951
TEXA	1	0.7353	0.7529	0.7413	0.7980	0.8302	0.8138
ToGS	hermesbeamnonehermes318Bbase	0.1038	0.0297	0.0233	0.0347	0.0182	0.0238
ToGS	hermesbeamnonehermes323Bbase	0.0459	0.0200	0.0125	0.0310	0.0126	0.0179
ToGS	hermesloraentitiesnaivebeamnonehermes323Bentities	0.3236	0.3916	0.3334	0.3428	0.5545	0.4237
ToGS	hermesnaivebeamnonehermes318Bbase	0.2765	0.4159	0.3073	0.2894	0.4804	0.3612
ToGS	hermesnaivebeamnonehermes323Bbase	0.2806	0.4171	0.3149	0.3005	0.4922	0.3732
ToGS	...beamnonehermes318Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...beamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...loraentitiesnaivebeamnonehermes318Bentities	0.3051	0.4530	0.3544	0.3282	0.5322	0.4061
ToGS	...loraentitiesnaivebeamnonehermes323Bentities	0.3326	0.4495	0.3685	0.3286	0.5278	0.4051
ToGS	...naivebeamnonehermes318Bbase	0.3065	0.4472	0.3453	0.3420	0.5293	0.4155
ToGS	...naivebeamnonehermes323Bbase	0.3007	0.4429	0.3412	0.3408	0.5311	0.4152
ToGS	...ragbeamnonehermes318Bbase	0.0035	0.0018	0.0024	0.0005	0.0004	0.0004
ToGS	...ragbeamnonehermes323Bbase	0.0043	0.0018	0.0025	0.0005	0.0004	0.0004
ToGS	...ragloraentitiesnaivebeamnonehermes318Bentities	0.3279	0.4487	0.3655	0.3304	0.5319	0.4076
ToGS	...ragloraentitiesnaivebeamnonehermes323Bentities	0.3263	0.4466	0.3641	0.3264	0.5267	0.4030
ToGS	hermesragbeamnonehermes318Bbase	0.3747	0.2238	0.2488	0.3129	0.3710	0.3395
ToGS	hermesragbeamnonehermes323Bbase	0.3228	0.1744	0.1925	0.2885	0.3225	0.3045
ToGS	hermesragloraentitiesbeamnonehermes318Bentities	0.1977	0.0710	0.0472	0.3220	0.2706	0.2941
ToGS	hermesragloraentitiesbeamnonehermes323Bentities	0.1135	0.0695	0.0429	0.3099	0.2709	0.2891
ToGS	hermesragloraentitiesnaivebeamnonehermes323Bentities	0.3188	0.3897	0.3314	0.3448	0.5523	0.4246
ToGS	hermesragnaivebeamnonehermes318Bbase	0.3353	0.4506	0.3713	0.3496	0.5823	0.4369
ToGS	hermesragnaivebeamnonehermes323Bbase	0.3340	0.4413	0.3646	0.3450	0.5738	0.4309
ToGS	naiveentities	0.3072	0.4567	0.3560	0.3531	0.5397	0.4269
ToGS	naivefilteredentities	0.2842	0.2821	0.2712	0.3325	0.3362	0.3343
TUGW	...T61113NEREXP8SUB1union	0.4765	0.7543	0.5753	0.5272	0.8414	0.6482
TUGW	...T61124ENSEXP4SUB1union	0.4764	0.7542	0.5748	0.5239	0.8369	0.6444
TUGW	...T612224ENSEXP4SUB1ELfreqgazeintersection	0.7462	0.4223	0.5234	0.8858	0.5089	0.6464
TUGW	...T612224ENSEXP4SUB1ELfreqgazeunion	0.4757	0.7544	0.5744	0.5220	0.8340	0.6421
TUGW	...T612313NEREXP8SUB1ELfreqgazeintersection	0.7385	0.4277	0.5248	0.8693	0.5130	0.6452
TUGW	...T612313NEREXP8SUB1ELfreqgazeunion	0.4756	0.7531	0.5743	0.5251	0.8380	0.6456
TUGW	...T621BGPT5520Ggpt5.520shotgoldunion	0.3085	0.4602	0.3577	0.3560	0.5441	0.4304
TUGW	...T622BGPT5520Ggpt5.520shotgoldunion	0.3085	0.4602	0.3577	0.3560	0.5441	0.4304
TWIX	merged6	0.9242	0.7333	**0.8114**	0.9548	0.8058	**0.8740**
TWIX	merged5	**0.9380**	0.7262	0.8104	0.9543	0.8058	0.8738
TWIX	merged1	0.9175	0.7205	0.7982	0.9546	0.8021	0.8717
TWIX	merged3	0.9037	0.7188	0.7939	0.9508	0.8021	0.8701
TWIX	merged4	0.9227	0.7056	0.7871	0.9478	0.8002	0.8678
TWIX	merged2	0.9120	0.7151	0.7895	0.9456	0.7984	0.8658
TWIX	merged8	0.9325	0.6938	0.7841	**0.9618**	0.7835	0.8636
TWIX	merged7	0.8908	0.6888	0.7632	0.9372	0.7910	0.8579
TWIX	ensemble1	0.7860	0.6827	0.7248	0.8657	0.7809	0.8211
TWIX	ensemble2	0.7802	0.6845	0.7239	0.8603	0.7832	0.8199
TWIX	ensemble3	0.7806	0.6867	0.7246	0.8599	0.7850	0.8208
TWIX	ensemble4	0.7628	0.6989	0.7181	0.8416	0.8036	0.8221
TWIX	ensemble5	0.7906	0.6827	0.7227	0.8639	0.7884	0.8244
TWIX	ensemble6	0.8152	0.6685	0.7238	0.8807	0.7713	0.8224
TWIX	ensemble7	0.7833	0.6840	0.7217	0.8600	0.7898	0.8234
TWIX	ensemble8	0.7843	0.6846	0.7209	0.8587	0.7906	0.8232
TWIX	ner1	0.7026	0.7258	0.7111	0.7891	0.8128	0.8008
TWIX	ner3	0.7344	0.7158	0.7203	0.8046	0.7891	0.7968
TWIX	ner4	0.7081	0.6972	0.6980	0.7807	0.7917	0.7862
TWIX	ner5	0.6924	0.6974	0.6863	0.7710	0.7976	0.7841
TWIX	ner6	0.6781	0.6992	0.6842	0.7665	0.7984	0.7821
TWIX	ner2	0.6633	0.6975	0.6748	0.7682	0.7958	0.7817
TWIX	ner7	0.6710	0.6741	0.6700	0.7723	0.7895	0.7808
unibuc-bionlp-bp	1	0.7174	0.7679	0.7390	0.7816	0.8425	0.8109
unibuc-bionlp-bp	2	0.6049	0.7223	0.6538	0.7096	0.7980	0.7512

B. Subtask 6.1.2 (NERD) Overall Results

Table 12: Performance metrics of each team's submitted runs for NERD. For each evaluation metric, the best result is in bold, the second-best is underlined.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Organizers	BASELINE	0.3820	0.4045	0.3916	0.4281	0.4522	0.4398
AAU-Gut-Brain	SciBERT	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AAU-Gut-Brain	SciBERT	0.2697	0.2764	0.2723	0.3442	0.3477	0.3459
ELiRF-UPV	test11	0.3324	0.2280	0.2675	0.4198	0.2706	0.3291
GBA	0	0.3347	0.3979	0.3607	0.3912	0.4689	0.4265
GBA	1	0.3534	0.4040	0.3713	0.4116	0.4878	0.4465
GBA	25	0.3814	0.4438	0.4065	0.4334	0.5182	0.4720
GBA	2	0.3446	0.3476	0.3389	0.4090	0.4396	0.4237
GBA	3	0.3292	0.3902	0.3543	0.3834	0.4596	0.4181
GBA	4	0.3416	0.4058	0.3677	0.3939	0.4722	0.4295
GBA	5	0.3618	0.4353	0.3923	0.4385	0.5263	0.4784
GBA	6	0.3403	0.3900	0.3578	0.3919	0.4644	0.4251
GBA	7	0.3420	0.4032	0.3664	0.3987	0.4885	0.4390
GBA	8	0.3171	0.3737	0.3395	0.3714	0.4552	0.4091
GetGut@AAU	1	0.4603	0.4534	0.4550	0.5515	0.5519	0.5517
GetGut@AAU	2	0.4603	0.4534	0.4550	0.5515	0.5519	0.5517
GetGut@AAU	3	0.4603	0.4534	0.4550	0.5515	0.5519	0.5517
Graphwise	10	0.4964	0.4871	0.4900	0.5952	0.5875	0.5913
Graphwise	11	0.4821	0.4817	0.4800	0.5847	0.5834	0.5840
Graphwise	12	0.4683	0.4871	0.4765	0.5678	0.5864	0.5770
Graphwise	13	0.4833	0.4772	0.4794	0.5877	0.5775	0.5825
Graphwise	14	0.4871	0.4901	0.4866	0.5811	0.5867	0.5839
Graphwise	15	0.4729	0.4859	0.4781	0.5719	0.5808	0.5763
Graphwise	16	0.4845	0.4677	0.4736	0.5872	0.5704	0.5787
Graphwise	17	0.4865	0.4787	0.4786	0.5755	0.5823	0.5789
Graphwise	18	0.4620	0.4839	0.4716	0.5785	0.5871	0.5828
Graphwise	19	0.4845	0.4828	0.4819	0.5864	0.5864	0.5864
Graphwise	1	0.4546	0.4842	0.4676	0.5484	0.5793	0.5634
Graphwise	20	0.4889	0.4889	0.4850	0.5824	0.5923	0.5873
Graphwise	21	0.4965	0.4815	0.4864	0.5902	0.5830	0.5866
Graphwise	22	0.4707	0.4861	0.4760	0.5682	0.5867	0.5773
Graphwise	23	0.4869	0.4892	0.4858	0.5833	0.5864	0.5848
Graphwise	24	0.4846	0.4924	0.4856	0.5795	0.5904	0.5849
Graphwise	25	0.4845	0.4948	0.4886	0.5846	0.5915	0.5881
Graphwise	2	0.4994	0.4758	0.4854	0.5944	0.5752	0.5847
Graphwise	3	0.4895	0.4878	0.4851	0.5741	0.5830	0.5785
Graphwise	4	0.5186	0.4818	0.4969	0.6127	0.5812	0.5965
Graphwise	5	0.4918	0.4819	0.4850	0.5873	0.5797	0.5835
Graphwise	6	0.4597	0.4898	0.4729	0.5554	0.5867	0.5707
Graphwise	7	0.5039	0.4811	0.4903	0.6025	0.5830	0.5926
Graphwise	8	0.4931	0.4931	0.4898	0.5818	0.5908	0.5862
Graphwise	9	0.5235	0.4883	0.5029	0.6217	0.5897	0.6053
GutHub	1	0.3815	0.3018	0.3338	0.3856	0.3199	0.3497
MindGut link	BiomedBERT	0.4814	0.4305	0.4484	0.5738	0.5330	0.5527
MindGut link	GlinerFlanT5	0.4814	0.4305	0.4484	0.5738	0.5330	0.5527
MKTE	run1	0.3109	0.3117	0.2991	0.3974	0.3617	0.3787
NightSun	run01	0.5343	0.5091	0.5193	0.6240	0.6004	0.6120
NightSun	run02	0.5201	0.5113	0.5135	0.6112	0.6042	0.6076
NightSun	run03	0.5203	0.5149	0.5153	0.6084	0.6042	0.6063
NightSun	run04	0.5412	0.5138	0.5248	0.6279	0.6042	0.6158
NightSun	run05	0.5267	0.5157	0.5188	0.6145	0.6075	0.6110
NightSun	run06	0.5271	0.5195	0.5209	0.6122	0.6079	0.6100
NightSun	run07	0.5396	0.5121	0.5232	0.6263	0.6027	0.6143
NightSun	run08	0.5252	0.5140	0.5173	0.6130	0.6060	0.6095
NightSun	run09	0.5256	0.5179	0.5193	0.6107	0.6064	0.6085
NightSun	run10	0.5401	0.5128	0.5238	0.6267	0.6030	0.6147
NightSun	run11	0.5256	0.5147	0.5178	0.6134	0.6064	0.6099
NightSun	run12	0.5261	0.5186	0.5199	0.6110	0.6067	0.6089
NightSun	run13	0.5342	0.5087	0.5190	0.6240	0.6004	0.6120
NightSun	run14	0.5200	0.5106	0.5131	0.6108	0.6038	0.6073
NightSun	run15	0.5206	0.5145	0.5152	0.6084	0.6042	0.6063
NightSun	run16	0.5376	0.5107	0.5215	0.6290	0.6053	0.6169
NightSun	run17	0.5222	0.5115	0.5145	0.6138	0.6067	0.6103
NightSun	run18	0.5230	0.5158	0.5170	0.6122	0.6079	0.6100
NightSun	run19	0.5376	0.5107	0.5215	0.6290	0.6053	0.6169
NightSun	run20	0.5222	0.5115	0.5145	0.6138	0.6067	0.6103
NightSun	run21	0.5230	0.5158	0.5170	0.6122	0.6079	0.6100
NightSun	run22	0.5367	0.5101	0.5208	0.6256	0.6019	0.6135
NightSun	run23	0.5224	0.5120	0.5149	0.6123	0.6053	0.6088
NightSun	run25	0.5410	0.5137	0.5247	0.6275	0.6038	0.6154
NightSun	run26	0.5265	0.5156	0.5187	0.6142	0.6071	0.6106
NLP-DE	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

(Table 12 continued)

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NLP-DE	1	0.2699	0.2848	0.2764	0.3260	0.3395	0.3326
TEXA	1	0.4039	0.4168	0.4090	0.4581	0.4766	0.4672
ToGS	hermesbeamnonehermes318Bbase	0.0706	0.0175	0.0130	0.0234	0.0122	0.0161
ToGS	hermesbeamnonehermes323Bbase	0.0383	0.0116	0.0076	0.0191	0.0078	0.0111
ToGS	hermesloraentitiesnaivebeamnonehermes323Bentities	0.2234	0.2593	0.2238	0.2241	0.3625	0.2770
ToGS	hermesnaivebeamnonehermes318Bbase	0.1906	0.2770	0.2050	0.1998	0.3317	0.2494
ToGS	hermesnaivebeamnonehermes323Bbase	0.1992	0.2877	0.2181	0.2118	0.3469	0.2630
ToGS	...beamnonehermes318Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...beamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...loraentitiesnaivebeamnonehermes318Bentities	0.2158	0.3077	0.2455	0.2302	0.3732	0.2847
ToGS	...loraentitiesnaivebeamnonehermes323Bentities	0.2312	0.3023	0.2509	0.2299	0.3692	0.2833
ToGS	...naivebeamnonehermes318Bbase	0.2179	0.3118	0.2410	0.2435	0.3769	0.2959
ToGS	...naivebeamnonehermes323Bbase	0.2104	0.2995	0.2335	0.2364	0.3684	0.2880
ToGS	...ragbeamnonehermes318Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...ragbeamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...ragloraentitiesnaivebeamnonehermes318Bentities	0.2355	0.3159	0.2589	0.2379	0.3829	0.2934
ToGS	...ragloraentitiesnaivebeamnonehermes323Bentities	0.2259	0.3011	0.2480	0.2262	0.3651	0.2794
ToGS	hermesragbeamnonehermes318Bbase	0.1918	0.1211	0.1291	0.1929	0.2287	0.2093
ToGS	hermesragbeamnonehermes323Bbase	0.1575	0.0947	0.0983	0.1804	0.2016	0.1904
ToGS	hermesragloraentitiesbeamnonehermes318Bentities	0.0936	0.0468	0.0298	0.2184	0.1835	0.1994
ToGS	hermesragloraentitiesbeamnonehermes323Bentities	0.0288	0.0444	0.0246	0.2064	0.1805	0.1926
ToGS	hermesragloraentitiesnaivebeamnonehermes323Bentities	0.2142	0.2509	0.2168	0.2206	0.3532	0.2715
ToGS	hermesragnaivebeamnonehermes318Bbase	0.1959	0.2615	0.2147	0.2123	0.3536	0.2653
ToGS	hermesragnaivebeamnonehermes323Bbase	0.2052	0.2662	0.2206	0.2189	0.3640	0.2733
ToGS	naiveentities	0.2179	0.3155	0.2488	0.2502	0.3825	0.3026
ToGS	naivefilteredentities	0.2530	0.2549	0.2429	0.3057	0.3091	0.3074
TUGW	...T61113NEREXP8SUB1union	0.1110	0.1762	0.1339	0.1284	0.2050	0.1579
TUGW	...T61124ENSEXP4SUB1union	0.1136	0.1793	0.1366	0.1297	0.2072	0.1595
TUGW	...T612224ENSEXP4SUB1ELfreqgazeintersection	0.6342	0.3626	0.4470	**0.7723**	0.4437	0.5636
TUGW	...T612224ENSEXP4SUB1ELfreqgazeunion	0.2630	0.4208	0.3180	0.3146	0.5026	0.3870
TUGW	...T612313NEREXP8SUB1ELfreqgazeintersection	0.6283	0.3676	0.4485	0.7582	0.4474	0.5627
TUGW	...T612313NEREXP8SUB1ELfreqgazeunion	0.2666	0.4266	0.3227	0.3161	0.5044	0.3886
TUGW	...T621BGPT5520Ggpt5.520shotgoldunion	0.2187	0.3159	0.2493	0.2502	0.3825	0.3026
TUGW	...T622BGPT5520Ggpt5.520shotgoldunion	0.2187	0.3159	0.2493	0.2502	0.3825	0.3026
TWIX	24merged6	0.6695	**0.5431**	0.5965	0.7527	**0.6353**	**0.6890**
TWIX	16merged4	**0.6857**	0.5360	0.5938	0.7511	0.6342	0.6877
TWIX	20merged5	0.6741	0.5374	0.5934	0.7511	0.6342	0.6877
TWIX	4merged1	0.6643	0.5362	0.5891	0.7530	0.6327	0.6876
TWIX	12merged3	0.6703	0.5404	0.5941	0.7509	0.6334	0.6872
TWIX	3merged1	0.6720	0.5380	0.5923	0.7517	0.6316	0.6864
TWIX	23merged6	0.6730	0.5430	**0.5973**	0.7492	0.6323	0.6858
TWIX	8merged2	0.6745	0.5381	0.5916	0.7489	0.6323	0.6857
TWIX	1merged1	0.6730	0.5384	0.5929	0.7503	0.6305	0.6852
TWIX	15merged4	0.6851	0.5348	0.5927	0.7480	0.6316	0.6849
TWIX	19merged5	0.6768	0.5378	0.5944	0.7480	0.6316	0.6849
TWIX	7merged2	0.6738	0.5383	0.5914	0.7471	0.6308	0.6841
TWIX	11merged3	0.6655	0.5366	0.5899	0.7474	0.6305	0.6840
TWIX	21merged6	0.6727	0.5424	0.5968	0.7466	0.6301	0.6834
TWIX	17merged5	0.6793	0.5382	0.5954	0.7458	0.6297	0.6829
TWIX	13merged4	0.6849	0.5341	0.5922	0.7454	0.6294	0.6825
TWIX	5merged2	0.6746	0.5382	0.5917	0.7454	0.6294	0.6825
TWIX	2merged1	0.6673	0.5357	0.5895	0.7472	0.6279	0.6824
TWIX	9merged3	0.6676	0.5379	0.5915	0.7456	0.6290	0.6823
TWIX	22merged6	0.6687	0.5406	0.5944	0.7453	0.6290	0.6822
TWIX	18merged5	0.6719	0.5346	0.5908	0.7432	0.6275	0.6805
TWIX	10merged3	0.6662	0.5367	0.5902	0.7434	0.6271	0.6803
TWIX	14merged4	0.6803	0.5313	0.5888	0.7428	0.6271	0.6801
TWIX	6merged2	0.6712	0.5351	0.5884	0.7423	0.6268	0.6797

C. Subtask 6.2.1 (M-RE) Overall Results

Table 13: Performance metrics of each team's submitted runs for M-RE. For each evaluation metric, the best result is in bold, the second-best is underlined.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Organizers	BASELINE	0.3660	0.2862	0.3003	0.4444	0.3453	0.3886
GBA	0	0.2288	0.1207	0.1278	0.2282	0.1256	0.1620
GBA	1	0.2413	0.1123	0.1267	0.2420	0.1204	0.1608
GBA	25	0.2528	0.1041	0.1269	0.2945	0.1166	0.1670
GBA	2	0.1739	0.0198	0.0327	0.1972	0.0269	0.0474
GBA	3	0.2288	0.1207	0.1278	0.2282	0.1256	0.1620
GBA	4	0.2288	0.1207	0.1278	0.2282	0.1256	0.1620
GBA	5	0.2518	0.0934	0.1171	0.2401	0.1166	0.1570
GBA	6	0.2413	0.1123	0.1267	0.2420	0.1204	0.1608
GBA	7	0.2318	0.1478	0.1467	0.2198	0.1621	0.1866
GBA	8	0.2318	0.1478	0.1467	0.2198	0.1621	0.1866
GetGut@AAU	1	0.3961	0.3167	0.3251	0.4546	0.3658	0.4054
GetGut@AAU	2	0.3961	0.3167	0.3251	0.4546	0.3658	0.4054
GetGut@AAU	3	0.3961	0.3167	0.3251	0.4546	0.3658	0.4054
Graphwise	BABB	0.2956	0.1879	0.2073	0.4890	0.1281	0.2030
Graphwise	BABE	0.2438	0.1698	0.1754	0.4661	0.1320	0.2057
Graphwise	BABLB	0.2455	0.2038	0.2019	0.4574	0.1409	0.2155
Graphwise	BABLL	0.2424	0.2065	0.1977	0.4795	0.1345	0.2101
Graphwise	BABOF	0.2721	0.1836	0.1911	0.4750	0.1339	0.2089
Graphwise	BABO	0.2687	0.2068	0.2088	0.5184	0.1352	0.2144
Graphwise	BAM	0.2384	0.1926	0.1906	0.4493	0.1307	0.2025
Graphwise	BAPBEBO	0.2476	0.2231	0.2020	0.4554	0.1473	0.2227
Graphwise	BAR	0.2566	0.1968	0.1961	0.4756	0.1249	0.1979
Graphwise	BGPT5515DEV	0.2526	0.3687	0.2752	0.3446	0.4254	0.3807
Graphwise	BGPT5515GEXP4	0.2473	0.3715	0.2768	0.3350	0.4228	0.3738
Graphwise	BGPT5515GEXP8	0.2583	0.3818	0.2868	0.3517	0.4497	0.3947
Graphwise	BGPT5515G	0.2519	0.3721	0.2822	0.3532	0.4471	0.3947
Graphwise	BGPT5520DEV	0.2614	0.3841	0.2773	0.3454	0.4471	0.3897
Graphwise	BGPT5520GEXP4	0.2408	0.3511	0.2627	0.3386	0.4126	0.3719
Graphwise	BGPT5520GEXP8	0.2780	0.3827	0.2978	0.3579	0.4395	0.3945
Graphwise	BGPT5520GEXP8	0.2244	0.3597	0.2530	0.3224	0.4158	0.3632
Graphwise	BGPT5520G	0.2667	0.3901	0.2880	0.3507	0.4433	0.3916
Graphwise	BGPT5525DEV	0.2495	0.3699	0.2703	0.3478	0.4452	0.3906
Graphwise	BGPT5525GEXP4	0.2398	0.3515	0.2626	0.3282	0.4126	0.3656
Graphwise	BGPT5525GEXP8	0.2300	0.3654	0.2570	0.3265	0.4087	0.3630
Graphwise	BGPT5525G	0.2427	0.3734	0.2712	0.3255	0.4190	0.3664
Graphwise	BQWENEXP4	0.3608	0.1714	0.2141	0.5059	0.2191	0.3058
Graphwise	BQWEN	0.3760	0.1950	0.2308	0.5219	0.2447	0.3332
Graphwise	PABEBIOBL	0.2601	0.2499	0.2219	0.4328	0.1608	0.2345
GutHub	1	0.3111	0.2760	0.2751	0.4347	0.3754	0.4029
MindGut link	BiomedBERT	0.0725	0.1079	0.0775	0.1318	0.1749	0.1503
MindGut link	GlinerFlanT5	0.0872	0.0845	0.0514	0.1295	0.1416	0.1353
MKTE	run1	0.0813	0.1568	0.0895	0.0800	0.1365	0.1009
NightSun	run01	0.3878	0.3968	0.3316	0.4457	0.4465	0.4461
NightSun	run02	0.3879	0.3968	0.3317	0.4454	0.4465	0.4459
NightSun	run03	0.3890	0.3993	0.3314	0.4403	0.4510	0.4456
NightSun	run04	0.3834	0.4071	0.3328	0.4233	0.4612	0.4414
NightSun	run05	0.3829	0.3857	0.3323	0.4560	0.4452	0.4506
NightSun	run06	0.3824	0.3857	0.3320	0.4557	0.4452	0.4504
NightSun	run07	0.3823	0.4086	0.3327	0.4226	0.4632	0.4419
NightSun	run08	0.3639	0.3857	0.3247	0.4321	0.4446	0.4383
NightSun	run09	0.3729	0.4018	0.3301	0.4291	0.4555	0.4419
NightSun	run10	0.3779	0.4020	0.3361	0.4466	0.4580	0.4522
NightSun	run11	0.3736	0.3951	0.3306	0.4255	0.4536	0.4391
NightSun	run12	0.3865	0.4026	0.3304	0.4297	0.4600	0.4443
NightSun	run13	0.3998	0.3908	0.3337	0.4418	0.4446	0.4432
NightSun	run14	0.3619	0.4065	0.3321	0.4275	0.4593	0.4429
NightSun	run15	0.3748	0.3951	0.3312	0.4260	0.4536	0.4393
NightSun	run16	0.3940	0.3924	0.3346	0.4542	0.4452	0.4497
NightSun	run17	0.3722	0.4023	0.3298	0.4284	0.4561	0.4418
NightSun	run18	0.3773	0.4025	0.3360	0.4459	0.4593	0.4525
NightSun	run19	0.4019	0.3909	0.3346	0.4432	0.4452	0.4442
NightSun	run20	0.3757	0.4065	0.3331	0.4270	0.4625	0.4440
NightSun	run21	0.3805	0.4064	0.3343	0.4402	0.4644	0.4520
NightSun	run22	0.3674	0.4065	0.3350	0.4393	0.4587	0.4488
NightSun	run23	0.3688	0.4086	0.3370	0.4414	0.4612	0.4511
NightSun	run24	0.3768	0.4072	0.3336	0.4268	0.4632	0.4442
NightSun	run25	0.3816	0.4031	0.3339	0.4300	0.4542	0.4417
SMTE	R1	0.3301	0.3239	0.2973	0.4349	0.4042	0.4190
SMTE	R2	0.3459	0.3369	0.3084	0.4476	0.3997	0.4223
TEXA	1	0.3804	0.2761	0.2947	0.4548	0.3382	0.3880
ToGS	hermesbeamnonehermes318Bbase	0.0007	0.0089	0.0010	0.0023	0.0038	0.0029

(Table 13 continued)

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ToGS	hermesbeamnonehermes323Bbase	0.0000	0.0014	0.0001	0.0009	0.0006	0.0007
ToGS	hermeslorasbeamnonehermes318Bs	0.1111	0.0758	0.0649	0.1114	0.1166	0.1139
ToGS	hermeslorasbeamnonehermes323Bs	0.0770	0.0613	0.0490	0.1154	0.1121	0.1137
ToGS	hermeslorasnaivebeamnonehermes318Bs	0.1111	0.0758	0.0649	0.1114	0.1166	0.1139
ToGS	hermeslorasnaivebeamnonehermes323Bs	0.0770	0.0613	0.0490	0.1154	0.1121	0.1137
ToGS	hermesnaivebeamnonehermes318Bbase	0.0007	0.0089	0.0010	0.0023	0.0038	0.0029
ToGS	hermesnaivebeamnonehermes323Bbase	0.0000	0.0014	0.0001	0.0009	0.0006	0.0007
ToGS	...beamnonehermes318Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...beamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...lorasnaivebeamnonehermes323Bs	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...naivebeamnonehermes318Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...naivebeamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...ragbeamnonehermes318Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...ragbeamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...raglorasnaivebeamnonehermes318Bs	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...raglorasnaivebeamnonehermes323Bs	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	hermesragbeamnonehermes318Bbase	0.0934	0.0709	0.0635	0.0668	0.0871	0.0756
ToGS	hermesragbeamnonehermes323Bbase	0.0576	0.0548	0.0458	0.0593	0.0730	0.0655
ToGS	hermesraglorasbeamnonehermes318Bs	0.0373	0.0475	0.0222	0.0994	0.0993	0.0993
ToGS	hermesraglorasbeamnonehermes323Bs	0.0411	0.0457	0.0226	0.1065	0.1006	0.1035
ToGS	hermesraglorasnaivebeamnonehermes318Bs	0.0373	0.0475	0.0222	0.0994	0.0993	0.0993
ToGS	hermesraglorasnaivebeamnonehermes323Bs	0.0411	0.0457	0.0226	0.1065	0.1006	0.1035
ToGS	hermesragnaivebeamnonehermes318Bbase	0.0934	0.0709	0.0635	0.0668	0.0871	0.0756
ToGS	hermesragnaivebeamnonehermes323Bbase	0.0576	0.0548	0.0458	0.0593	0.0730	0.0655
TWIX	mre10	0.6588	**0.7308**	0.6859	**0.8996**	**0.4651**	**0.6132**
TWIX	mre11	0.6588	**0.7308**	0.6859	**0.8996**	**0.4651**	**0.6132**
TWIX	mre8	**0.6598**	**0.7308**	**0.6866**	**0.8996**	**0.4651**	**0.6132**
TWIX	mre7	0.6570	**0.7308**	0.6847	0.8974	**0.4651**	0.6127
TWIX	mre9	0.6513	**0.7308**	0.6812	0.8821	**0.4651**	0.6091
TWIX	mre12	0.6509	**0.7308**	0.6810	0.8811	**0.4651**	0.6088
TWIX	mre13	0.6472	**0.7308**	0.6808	0.8726	**0.4651**	0.6068
TWIX	mre14	0.6477	**0.7308**	0.6810	0.8715	**0.4651**	0.6065
TWIX	mre17	0.6476	**0.7308**	0.6810	0.8715	**0.4651**	0.6065
TWIX	mre16	0.6469	**0.7308**	0.6806	0.8695	**0.4651**	0.6060
TWIX	mre20	0.6412	**0.7308**	0.6754	0.8612	**0.4651**	0.6040
TWIX	mre22	0.6403	**0.7308**	0.6748	0.8612	**0.4651**	0.6040
TWIX	mre23	0.6403	**0.7308**	0.6748	0.8612	**0.4651**	0.6040
TWIX	mre19	0.6388	**0.7308**	0.6737	0.8582	**0.4651**	0.6032
TWIX	mre15	0.6409	**0.7308**	0.6761	0.8511	**0.4651**	0.6015
TWIX	mre21	0.6368	**0.7308**	0.6716	0.8511	**0.4651**	0.6015
TWIX	mre24	0.6346	**0.7308**	0.6710	0.8481	**0.4651**	0.6007
TWIX	mre18	0.6370	**0.7308**	0.6740	0.8471	**0.4651**	0.6005
TWIX	mre5	0.6139	**0.7308**	0.6563	0.8374	**0.4651**	0.5980
TWIX	mre1	0.6124	**0.7308**	0.6552	0.8354	**0.4651**	0.5975
TWIX	mre2	0.6068	**0.7308**	0.6522	0.8354	**0.4651**	0.5975
TWIX	mre4	0.6121	**0.7308**	0.6550	0.8335	**0.4651**	0.5970
TWIX	mre3	0.6086	**0.7308**	0.6527	0.8259	**0.4651**	0.5951
TWIX	mre6	0.6053	**0.7308**	0.6505	0.8222	**0.4651**	0.5941

D. Subtask 6.2.2 (C-RE) Overall Results

Table 14: Performance metrics of each team's submitted runs for C-RE. For each evaluation metric, the best result is in bold, the second-best is underlined.

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Organizers	BASELINE	0.1009	0.1021	0.0966	0.1403	0.1292	0.1345
GBA	0	0.1227	0.0695	0.0758	0.1153	0.0774	0.0926
GBA	1	0.0939	0.0493	0.0552	0.1128	0.0691	0.0857
GBA	25	0.1135	0.0742	0.0786	0.1717	0.0988	0.1254
GBA	2	0.0492	0.0173	0.0225	0.0986	0.0173	0.0294
GBA	3	0.1094	0.0617	0.0712	0.1178	0.0790	0.0946
GBA	4	0.1037	0.0540	0.0637	0.1115	0.0749	0.0896
GBA	5	0.1418	0.0839	0.0901	0.1580	0.0955	0.1190
GBA	6	0.1080	0.0585	0.0643	0.1053	0.0650	0.0804
GBA	7	0.1136	0.0819	0.0847	0.1083	0.0979	0.1029
GBA	8	0.1031	0.0716	0.0756	0.0898	0.0815	0.0854
GetGut@AAU	1	0.1481	0.1505	0.1385	0.2123	0.1926	0.2020
GetGut@AAU	2	0.1481	0.1505	0.1385	0.2123	0.1926	0.2020
GetGut@AAU	3	0.1481	0.1505	0.1385	0.2123	0.1926	0.2020
Graphwise	BABB	0.0675	0.0510	0.0551	0.1289	0.0412	0.0624
Graphwise	BABE	0.0779	0.0459	0.0513	0.1364	0.0469	0.0698
Graphwise	BABLB	0.0559	0.0558	0.0516	0.1242	0.0461	0.0672
Graphwise	BABLL	0.0618	0.0692	0.0568	0.1202	0.0412	0.0613
Graphwise	BABOF	0.0857	0.0576	0.0626	0.1348	0.0469	0.0696
Graphwise	BABO	0.0800	0.0793	0.0700	0.1554	0.0494	0.0750
Graphwise	BAM	0.0516	0.0472	0.0467	0.1101	0.0395	0.0581
Graphwise	BAPBEBO	0.0620	0.0819	0.0617	0.1245	0.0494	0.0707
Graphwise	BAR	0.0667	0.0716	0.0616	0.1450	0.0469	0.0709
Graphwise	BGPT5515DEV	0.0931	0.1742	0.1084	0.1181	0.1728	0.1403
Graphwise	BGPT5515GEXP4	0.1582	0.2651	0.1814	0.2092	**0.2963**	0.2452
Graphwise	BGPT5515GEXP8	0.1373	0.2446	0.1668	0.1976	0.2741	0.2297
Graphwise	BGPT5515G	0.0927	0.1738	0.1119	0.1330	0.1942	0.1579
Graphwise	BGPT5520DEV	0.1016	0.1692	0.1149	0.1247	0.1794	0.1471
Graphwise	BGPT5520GEXP4	0.1299	0.2296	0.1489	0.1771	0.2617	0.2112
Graphwise	BGPT5520GEXP8	0.1343	0.2434	0.1602	0.1864	0.2724	0.2213
Graphwise	BGPT5520GEXP8	0.1311	0.2496	0.1573	0.1826	0.2840	0.2223
Graphwise	BGPT5520G	0.0876	0.1750	0.1072	0.1156	0.1753	0.1393
Graphwise	BGPT5525DEV	0.0839	0.1569	0.1000	0.1193	0.1712	0.1406
Graphwise	BGPT5525GEXP4	0.1295	0.2308	0.1519	0.1804	0.2617	0.2136
Graphwise	BGPT5525GEXP8	0.1322	0.2508	0.1624	0.1863	0.2790	0.2234
Graphwise	BGPT5525G	0.0849	0.1544	0.1025	0.1194	0.1778	0.1429
Graphwise	BQWENEXP4	0.1301	0.0942	0.1019	0.1748	0.0971	0.1249
Graphwise	BQWEN	0.1301	0.0942	0.1019	0.1748	0.0971	0.1249
Graphwise	PABEBIOBL	0.0611	0.0892	0.0637	0.1287	0.0576	0.0796
GutHub	1	0.1010	0.1100	0.0996	0.1305	0.1284	0.1295
MindGut link	BiomedBERT	0.0756	0.0974	0.0764	0.1042	0.1374	0.1186
MindGut link	GlinerFlanT5	0.0849	0.0725	0.0432	0.1059	0.1111	0.1084
NightSun	run01	0.2005	0.2325	0.1838	0.2440	0.2675	0.2552
NightSun	run02	0.2006	0.2325	0.1839	0.2440	0.2675	0.2552
NightSun	run03	0.1979	0.2368	0.1846	0.2441	0.2741	0.2582
NightSun	run04	0.2096	0.2447	0.1897	0.2382	0.2831	0.2587
NightSun	run05	0.2066	0.2318	0.1884	0.2486	0.2650	0.2566
NightSun	run06	0.2061	0.2318	0.1880	0.2485	0.2650	0.2565
NightSun	run07	0.2086	0.2447	0.1887	0.2368	0.2831	0.2579
NightSun	run08	0.1959	0.2261	0.1793	0.2386	0.2667	0.2518
NightSun	run09	0.1967	0.2384	0.1848	0.2381	0.2749	0.2552
NightSun	run10	0.2000	0.2391	0.1888	0.2487	0.2782	0.2626
NightSun	run11	0.1989	0.2385	0.1871	0.2360	0.2733	0.2532
NightSun	run12	0.2069	0.2388	0.1836	0.2384	0.2790	0.2571
NightSun	run13	0.2097	0.2296	0.1848	0.2401	0.2658	0.2523
NightSun	run14	0.1825	0.2394	0.1832	0.2365	0.2765	0.2549
NightSun	run15	0.1998	0.2385	0.1875	0.2365	0.2733	0.2535
NightSun	run16	0.2064	0.2313	0.1865	0.2502	0.2675	0.2586
NightSun	run17	0.1962	0.2384	0.1845	0.2376	0.2749	0.2549
NightSun	run18	0.1995	0.2397	0.1889	0.2484	0.2798	0.2632
NightSun	run19	0.2112	0.2298	0.1855	0.2413	0.2667	0.2533
NightSun	run20	0.1988	0.2423	0.1874	0.2400	0.2823	0.2595
NightSun	run21	0.1907	0.2394	0.1860	0.2450	0.2823	0.2623
NightSun	run22	0.1892	0.2362	0.1835	0.2376	0.2724	0.2538
NightSun	run23	0.1893	0.2362	0.1836	0.2374	0.2724	0.2537
NightSun	run24	0.1968	0.2423	0.1871	0.2397	0.2823	0.2593
NightSun	run25	0.2036	0.2421	0.1879	0.2385	0.2774	0.2565
TEXA	1	0.1063	0.0991	0.0981	0.1556	0.1383	0.1464
ToGS	hermesbeamnonehermes318Bbase	0.0013	0.0227	0.0020	0.0031	0.0049	0.0038
ToGS	hermesbeamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	hermeslorasbeamnonehermes318Bs	0.0454	0.0347	0.0251	0.0598	0.0700	0.0645
ToGS	hermeslorasbeamnonehermes323Bs	0.0283	0.0386	0.0266	0.0602	0.0683	0.0640

(Table 14 continued)

		Macro-averaging			Micro-averaging		
Team ID	**Run ID**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ToGS	hermeslorasnaivebeamnonehermes318Bs	0.0454	0.0347	0.0251	0.0598	0.0700	0.0645
ToGS	hermeslorasnaivebeamnonehermes323Bs	0.0283	0.0386	0.0266	0.0602	0.0683	0.0640
ToGS	hermesnaivebeamnonehermes318Bbase	0.0013	0.0227	0.0020	0.0031	0.0049	0.0038
ToGS	hermesnaivebeamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...beamnonehermes318Bbase	0.0002	0.0053	0.0004	0.0015	0.0016	0.0016
ToGS	...beamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...lorasnaivebeamnonehermes323Bs	0.0121	0.0066	0.0076	0.0090	0.0091	0.0090
ToGS	...naivebeamnonehermes318Bbase	0.0002	0.0077	0.0004	0.0015	0.0016	0.0016
ToGS	...naivebeamnonehermes323Bbase	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	...ragbeamnonehermes318Bbase	0.0091	0.0078	0.0065	0.0096	0.0107	0.0101
ToGS	...ragbeamnonehermes323Bbase	0.0207	0.0105	0.0111	0.0105	0.0107	0.0106
ToGS	...raglorasnaivebeamnonehermes318Bs	0.0186	0.0032	0.0045	0.0050	0.0041	0.0045
ToGS	...raglorasnaivebeamnonehermes323Bs	0.0005	0.0029	0.0008	0.0049	0.0041	0.0045
ToGS	hermesragbeamnonehermes318Bbase	0.0514	0.0377	0.0340	0.0363	0.0543	0.0435
ToGS	hermesragbeamnonehermes323Bbase	0.0327	0.0370	0.0274	0.0379	0.0535	0.0444
ToGS	hermesraglorasbeamnonehermes318Bs	0.0206	0.0328	0.0145	0.0542	0.0593	0.0566
ToGS	hermesraglorasbeamnonehermes323Bs	0.0178	0.0258	0.0109	0.0489	0.0527	0.0507
ToGS	hermesraglorasnaivebeamnonehermes318Bs	0.0206	0.0328	0.0145	0.0542	0.0593	0.0566
ToGS	hermesraglorasnaivebeamnonehermes323Bs	0.0178	0.0258	0.0109	0.0489	0.0527	0.0507
ToGS	hermesragnaivebeamnonehermes318Bbase	0.0514	0.0377	0.0340	0.0363	0.0543	0.0435
ToGS	hermesragnaivebeamnonehermes323Bbase	0.0327	0.0370	0.0274	0.0379	0.0535	0.0444
TUGW	...T61113NEREXP8SUB1union	0.0283	0.0386	0.0266	0.0602	0.0683	0.0640
TUGW	...T61124ENSEXP4SUB1union	0.0283	0.0386	0.0266	0.0602	0.0683	0.0640
TWIX	11mre12	**0.3699**	**0.4257**	**0.3902**	**0.4903**	0.2691	**0.3475**
TWIX	15mre7	**0.3699**	**0.4257**	**0.3902**	**0.4903**	0.2691	**0.3475**
TWIX	19mre8	**0.3699**	**0.4257**	**0.3902**	**0.4903**	0.2691	**0.3475**
TWIX	23mre9	**0.3699**	**0.4257**	**0.3902**	**0.4903**	0.2691	**0.3475**
TWIX	12mre12	0.3287	0.3878	0.3496	0.4895	0.2691	0.3473
TWIX	16mre7	0.3287	0.3878	0.3496	0.4895	0.2691	0.3473
TWIX	20mre8	0.3287	0.3878	0.3496	0.4895	0.2691	0.3473
TWIX	24mre9	0.3287	0.3878	0.3496	0.4895	0.2691	0.3473
TWIX	3mre10	**0.3699**	**0.4257**	**0.3902**	0.4888	0.2691	0.3471
TWIX	7mre11	**0.3699**	**0.4257**	**0.3902**	0.4888	0.2691	0.3471
TWIX	4mre10	0.3287	0.3878	0.3496	0.4881	0.2691	0.3469
TWIX	8mre11	0.3287	0.3878	0.3496	0.4881	0.2691	0.3469
TWIX	10mre12	0.3286	0.3879	0.3498	0.4814	0.2658	0.3425
TWIX	14mre7	0.3286	0.3879	0.3498	0.4814	0.2658	0.3425
TWIX	18mre8	0.3286	0.3879	0.3498	0.4814	0.2658	0.3425
TWIX	22mre9	0.3286	0.3879	0.3498	0.4814	0.2658	0.3425
TWIX	2mre10	0.3286	0.3879	0.3498	0.4799	0.2658	0.3422
TWIX	6mre11	0.3286	0.3879	0.3498	0.4799	0.2658	0.3422
TWIX	13mre7	0.3440	0.3996	0.3632	0.4717	0.2609	0.3360
TWIX	17mre8	0.3440	0.3996	0.3632	0.4717	0.2609	0.3360
TWIX	21mre9	0.3440	0.3996	0.3632	0.4717	0.2609	0.3360
TWIX	9mre12	0.3440	0.3996	0.3632	0.4717	0.2609	0.3360
TWIX	1mre10	0.3440	0.3996	0.3632	0.4703	0.2609	0.3356
TWIX	5mre11	0.3440	0.3996	0.3632	0.4703	0.2609	0.3356