Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction

Marco Martinelli[1,*], Gianmaria Silvello[1], Vanessa Bonato[2], Giorgio Maria Di Nunzio[1], Nicola Ferro[1], Ornella Irrera[1], Stefano Marchesin[1], Laura Menotti[1] and Federica Vezzani[2]

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

Abstract

Recent studies link the gut microbiota to mental health conditions and to neurodegenerative diseases such as Parkinson's and Alzheimer's. However, the rapid speed at which this research field is evolving presents a significant challenge for clinicians and researchers who have to keep pace with an ever-expanding volume of biomedical literature. In this context, automatic tools for extracting and structuring information from scientific texts are becoming essential to support 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 Information Extraction (IE).

GutBrainIE is part of the BioASQ Lab at CLEF 2025 and is organized within the context of the research project HEREDITARY, funded by the European Commission. The task includes four subtasks of increasing complexity, one dealing with Named Entity Recognition (NER) and the other three with Relation Extraction (RE), and comprises a dataset manually annotated for entities and relations structured into four quality tiers.

This extended overview describes the subtasks, dataset, evaluation methodology, results, and participant approaches for the GutBrainIE-2025 task.

Keywords

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

1. Introduction

Scientific evidence increasingly supports a connection between gut microbiota and several mental and neurological disorders, including Parkinson's, Alzheimer's, Multiple Sclerosis, and dementia. This emerging line of research on the gut–brain axis suggests that the gut microbiota may play a critical role in regulating brain function, also impacting on mental health [1, 2, 3, 4].

The growing interest in the gut-brain axis and its promising potential is driving a surge in biomedical research publications in this field. Looking at Figure 1, we can observe that in 2020 approximately 200 papers were published on the relationship between gut microbiota and neurological diseases; by 2024, this number had more than doubled, reaching over 500 publications. This rapid increase in the number of publications represents a considerable challenge for clinicians and researchers attempting to identify and interpret relevant findings across unstructured scientific texts.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
martinell2@dei.unipd.it (M. Martinelli); gianmaria.silvello@unipd.it (G. Silvello); vanessa.bonato@unipd.it (V. Bonato); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio); nicola.ferro@unipd.it (N. Ferro); ornella.irrera@unipd.it (O. Irrera); stefano.marchesin@unipd.it (S. Marchesin); laura.menotti@unipd.it (L. Menotti); federica.vezzani@unipd.it (F. Vezzani)
https://www.dei.unipd.it/~martinell2/ (M. Martinelli); https://www.dei.unipd.it/~silvello/ (G. Silvello); https://www.dei.unipd.it/~dinunzio/ (G. M. Di Nunzio); https://www.dei.unipd.it/~ferro/ (N. Ferro); https://www.dei.unipd.it/~irreraorne/ (O. Irrera); https://www.dei.unipd.it/~marches1/ (S. Marchesin); https://www.dei.unipd.it/~menottilau/ (L. Menotti)
0009-0001-1596-8642 (M. Martinelli); 0000-0003-4970-4554 (G. Silvello); 0009-0002-9918-282X (V. Bonato); 0000-0001-9709-6392 (G. M. Di Nunzio); 0000-0001-9219-6239 (N. Ferro); 0000-0003-2284-5699 (O. Irrera); 0000-0003-0362-5893 (S. Marchesin); 0000-0002-0676-682X (L. Menotti); 0000-0003-2240-6127 (F. Vezzani)

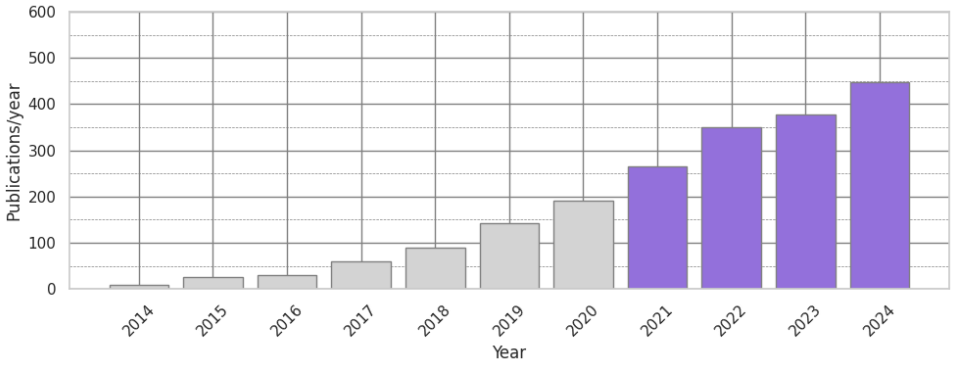

Figure 1: Annual rate of biomedical publications on PubMed related to the gut–brain axis, retrieved with the query *"gut microbiota" AND ("Parkinson" OR "mental health")* on July 1, 2025. Purple bars highlight years with more than 250 records.

In response to this challenge, the GutBrainIE-2025 task, part of the BioASQ Lab [5, 6] and inserted in the context of the EU-funded project HEREDITARY,[1] introduces a Natural Language Processing (NLP) challenge focused on extracting structured information from PubMed abstracts related to the gut–brain axis. The task aims to foster the development of robust and effective Information Extraction (IE) systems that support experts in analyzing the scientific literature, thereby contributing to biomedical knowledge discovery and, in the long term, informed clinical decision-making.

In its first edition, GutBrainIE proposes four subtasks of increasing complexity:

- **Subtask 6.1 - 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 6.2.1 - Binary Tag-based Relation Extraction (BT-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 entities are in relation within a document, without specifying the exact predicate or entity mentions involved.
- **Subtask 6.2.2 - Ternary Tag-based Relation Extraction (TT-RE)**: this subtask extends BT-RE by requiring participants to predict the specific relation predicate connecting each head-tail entity pair.
- **Subtask 6.2.3 - Ternary Mention-based Relation Extraction (TM-RE)**: this is the most challenging subtask, demanding to identify the exact entity mention involved in a relation and assign the correct relation predicate.

All subtasks target PubMed abstracts, leveraging a corpus of biomedical documents related to the gut-brain axis. Each document contains a title and abstract, both annotated with entity mentions and relations. Specifically, the GutBrainIE-2025 dataset consists of over 1500 annotated documents, split into Training, Development, and Test sets. A noteworthy feature of the dataset is its tiered annotation quality, organized as follows:

- **Platinum Annotations**: highest-quality annotations, expert-curated and internally reviewed;
- **Gold Annotations**: high-quality annotations and expert-curated;
- **Silver Annotations**: mid-quality annotations, created by trained students under expert supervision;
- **Bronze Annotations**: automatically generated annotations with no manual correction.

In particular, the Development and Test sets contain only expert annotations (Platinum- and Gold-Standard Annotations).

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

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.

This paper provides a comprehensive overview of the GutBrainIE-2025 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 first edition, GutBrainIE-2025 featured four subtasks:

1. **Named Entity Recognition** (NER).
2. **Binary Tag-based Relation Extraction** (BT-RE).
3. **Ternary Tag-based Relation Extraction** (TT-RE).
4. **Ternary Mention-based Relation Extraction** (TM-RE).

Participants were free to develop their systems without constraints on architecture, training methodology, or external resources, aiming to achieve the best possible performance. Overall, 17 teams submitted a total of 395 runs. In the remainder of this section, we describe each task in detail.

2.1. Subtask 1: Named Entity Recognition (NER)

The NER subtask focuses on classifying entity mentions into one of the 13 predefined categories. Participants were provided with PubMed abstracts related to the gut-brain axis and asked to identify specific text spans corresponding to one of the 13 categories defined in Table 1.

Each entity mention consists of the following elements:

- **Location**, indicating whether the entity mention appears in the title or in the abstract.
- **Start and end indices**, denoting character offsets of the entity mention within the text.
- **Text span**, representing the actual string of text corresponding to the mention.
- **Label**, specifying the entity label assigned to the mention.

A predicted entity mention is considered correct only if all its fields exactly match an entry in the ground truth.

2.2. Subtask 2: Binary Tag-based Relation Extraction (BT-RE) Subtask

The BT-RE subtask is one of the three GutBrainIE-2025 subtasks dealing with RE. In this subtask, participants have to determine which pairs of entities are in relation within a document, considering the set of relations defined in Table 2.

Within BT-RE, participants are not required to predict a relation predicate. Therefore, a predicted relation for this subtask will be a pair *(subjectEntityLabel; objectEntityLabel)*, where entity labels are taken from the ones reported in Table 1.

2.3. Subtask 3: Ternary Tag-based Relation Extraction (TT-RE) Subtask

The TT-RE subtask complements BT-RE by requiring participants to predict, along with the pair of entities in relation, the predicate of the relation holding among them. As in BT-RE, the set of relations to be considered is reported in Table 2.

Predicted relations for TT-RE will be triples *(subjectEntityLabel; relationPredicate; objectEntityLabel)*.

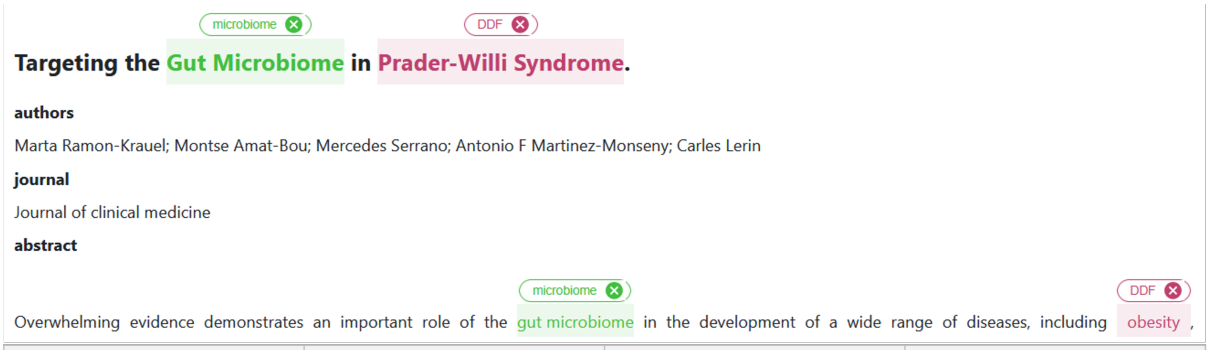

Figure 2: Example of annotated entity mentions for the NER subtask. The figure shows entities tagged in both the title and abstract of a PubMed article.

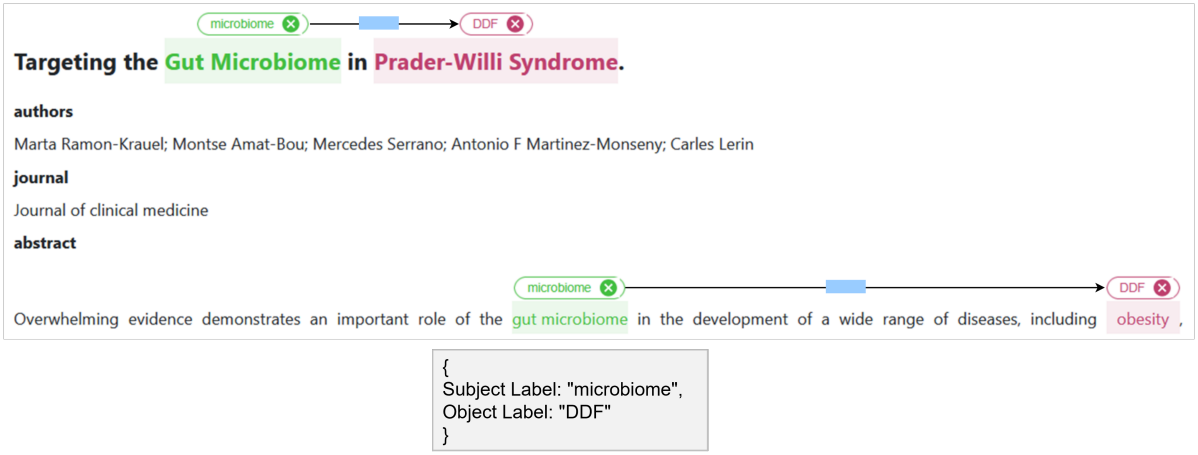

Figure 3: Example of an annotated binary relation for the BT-RE subtask. The figure shows pairs of entities identified as being in relation within a PubMed article, without specifying the relation predicate and the entity mentions involved. Duplicated tag-based binary relations are merged into a single entry.

2.4. Subtask 4: Ternary Mention-based Relation Extraction (TM-RE) Subtask

The TM-RE subtask is, among the three RE subtasks, the one most aligned with the standard NLP task of Relation Extraction [7]. Here, participants are required to identify the entity mentions involved in a relation, predict their entity labels, and specify the relation predicate that links them.

Predicted relations for TM-RE will be tuples *(subjectEntityTextSpan; subjectEntityLabel; relationPredicate; objectEntityTextSpan; objectEntityLabel)*.

3. Dataset

The released dataset for GutBrainIE-2025 consists of titles and abstracts of biomedical articles retrieved from PubMed, focusing on the gut-brain axis and its implications in neurological and mental health. Articles were manually annotated, either by experts or trained students,[2] for entity mentions (i.e., text

[2]The students we are referring to are enrolled in the Master Degree in Modern Languages for International Communication and Cooperation of the University of Padua. They received a specific training on medical terminology during the course of Translation-Oriented Terminography.

Figure 4: Example of an annotated ternary relation for the TT-RE subtask. The figure shows pairs of entities identified as being in relation within a PubMed article, specifying the relation predicate but not the entity mentions involved. Duplicated tag-based ternary relations are merged into a single entry.

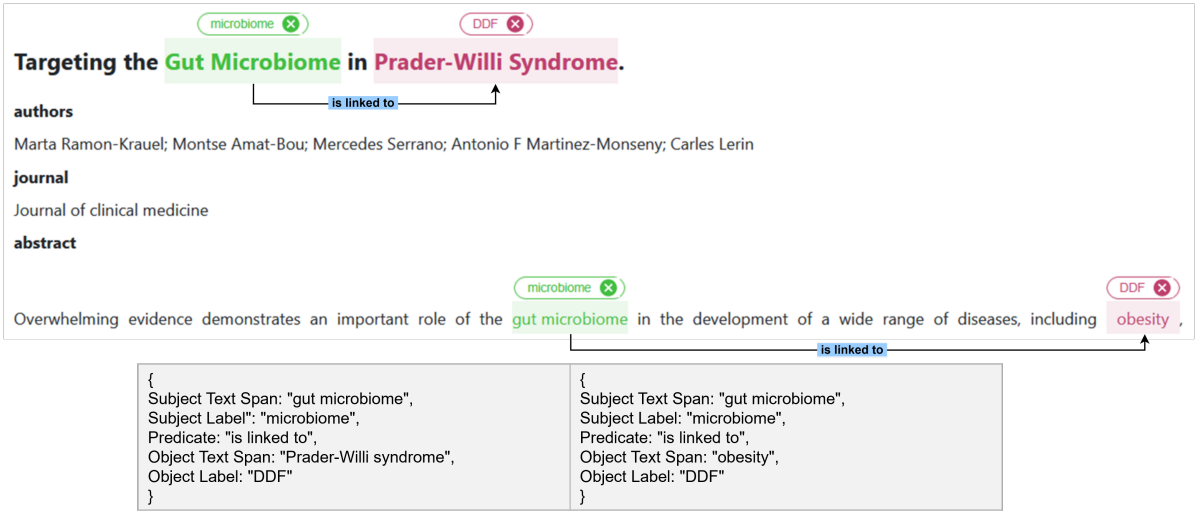

Figure 5: Example of an annotated ternary relation for the TM-RE subtask. The figure shows pairs of entities identified as being in relation within a PubMed article, specifying the relation predicate and the entity mentions involved.

spans mapped to one of the categories defined in Table 1) and relations (i.e., associations between entities defined in Table 2).

3.1. Dataset Creation

To build the GutBrainIE-2025 dataset, we first retrieved documents from PubMed using two separate queries: `"gut microbiota" AND "Parkinson"` and `"gut microbiota" AND "Mental Health"`. The first retrieval was performed on 09/05/2024 and yielded 828 documents. A second retrieval using the same queries was conducted on 31/10/2024, resulting in 834 additional documents not included in the first batch. We then filtered out documents from the years 2014–2019 (for the "Mental Health" query) and 2013–2020 (for the "Parkinson" query) due to the limited volume of relevant literature in those periods, discarding 16 documents in total. The final collection includes 1,647 documents.

Before starting manual annotation, documents were pre-annotated for NER using GLiNER [8] in a zero-shot setting, aiming to speed up and facilitate the annotation process. We decided not to pre-annotate documents for RE since, in a zero-shot setting, the likelihood of introducing noise was

Table 1
Overview of the 13 entity labels used in the GutBrainIE-2025 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.

significantly higher than that of adding valid relations. Excessive noise in pre-annotations could lead to biases among annotators, ultimately impacting the quality of the final annotated dataset [9].

Articles were then distributed between expert and student annotators. In total, 7 experts and 26

Table 2
Overview of the relations used in the GutBrainIE-2025 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

students annotated documents. Documents from the first retrieval were annotated exclusively by experts, while those from the second retrieval were assigned to students.

The annotation process was conducted in two phases, each followed by iterative refinement. At the end of each phase, expert annotators conducted a meeting to review progress, discuss critical challenges noted during the annotation phase, and make any necessary adjustments to the guidelines. These guidelines, publicly available at https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/files/GutBrainIE_2025_Annotation_Guidelines.pdf, were also shared with task participants so they could better tailor and tune their systems.

Once manual annotation was completed, we fine-tuned GLiNER [8] for NER and ATLOP [10] for RE using the annotated entities and relations and used them to annotate the remaining unannotated documents from both batches of the original retrieval. More detailed information about the fine-tuning of these models can be found in Section 4.4.

3.2. Dataset Folds

The training set is divided into four parts:

1. **Platinum Collection**: highest-quality annotations, expert-curated and revised internally by a subgroup of annotators to ensure consistency, uniformity, and alignment with the final annotation guidelines;
2. **Gold Collection**: high-quality annotations, expert-curated and produced after the finalization of the annotation guidelines. No subsequent revision performed;
3. **Silver Collection**: mid-quality annotations, created 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.
4. **Bronze Collection**: automatically generated annotations obtained using fine-tuned GLiNER (for NER) [8] and fine-tuned ATLOP (for RE) [10]. No manual revision was performed on this subset.

The development and test sets are held-out selections of documents from the gold and platinum collections, selected to ensure full representativeness and coverage of all entity and relation types.

Table 3
Dataset statistics for GutBrainIE-2025.

Collection	# Docs	# Entities	Ents/Doc	# Rels	Rels/Doc
Train Platinum	111	3638	32.77	1455	13.11
Train Gold	208	5192	24.96	1994	9.59
Train Silver	499	15275	30.61	10616	21.27
Train Bronze	749	21357	28.51	8165	11.90
Development Set	40	1117	27.93	623	15.58
Test Set	40	1237	30.92	777	19.42

3.3. Dataset Format

Annotations are provided in JSON format. Each entry corresponds 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_7`, `student_A`, `student_B`, or `distant` (automatically generated). Participants may decide to filter or weight examples differently based on the annotator.

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

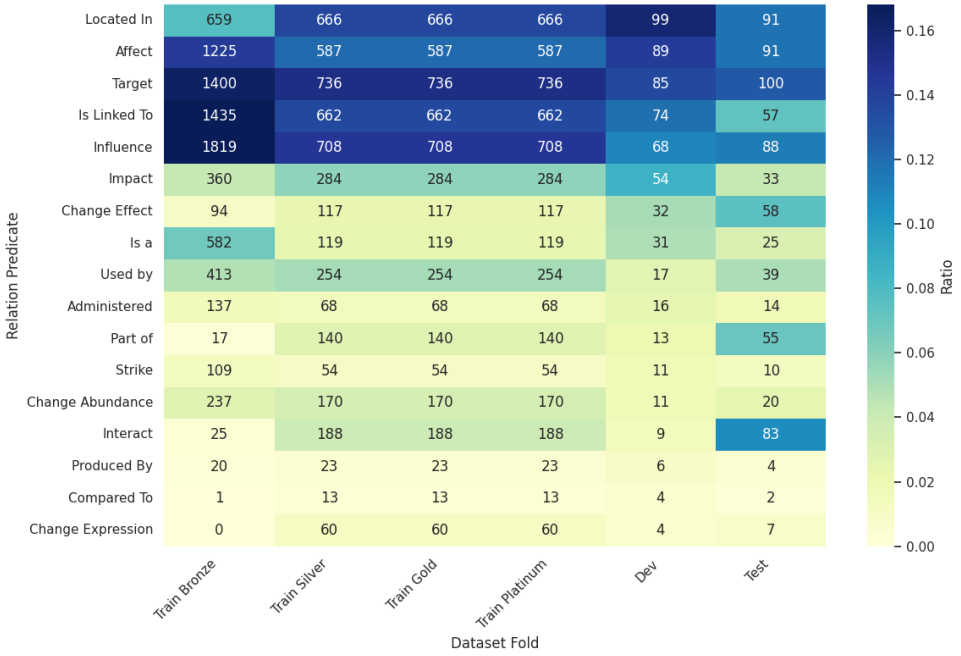

Figure 7: Distribution of annotated relations across dataset folds and relation predicates.

- **Entities:** An array of objects, each with:
  - *start*, *end*: character offsets of the text span associated to the entity mention;
  - *location*: "title" or "abstract";
  - *text_span*: the actual text span of the mention;
  - *label*: the annotated entity label (such as *bacteria*, *microbiome*).
- **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.

- **Binary Tag-based Relations:** Derived from the `Relations` array by extracting pairs ⟨`subject_label`, `object_label`⟩, omitting relation predicate and mention-level details.
- **Ternary Tag-based Relations:** Extracted from the `Relations` array as triples ⟨`subject_label`, `predicate`, `object_label`⟩, including relation predicate but leaving out mention-level details.
- **Ternary Mention-based 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.

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`
- `binary_tag_relations.csv` — `binary_tag_relations.tsv`
- `ternary_tag_relations.csv` — `ternary_tag_relations.tsv`
- `ternary_mention_relations.csv` — `ternary_mention_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-2025. A comprehensive description of the submitted systems can be found in Section 6 and in the participants' individual papers reported in Table 5.

Teams could participate in any of the four subtasks independently and submit up to 25 runs per subtask.

Although 85 teams from 29 different countries registered for the challenge, the final number of teams submitting at least one run was 17, resulting in 395 submitted runs. Among these, 15 teams also submitted a participant paper describing their methodologies, approaches, and systems. However, the discussion presented in Section 6 includes all 17 teams that submitted at least one run. Table 4 summarizes participation across the various subtasks.

The task began on February 3, 2025, with the release of the training and development sets. The test set was made available on April 28, and final submissions were due by May 10.

4.1. Guidelines

Participating teams were required to satisfy the following guidelines:

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

4.1.1. Subtask 1 (NER) Run Format

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

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	BT-RE	TT-RE	TM-RE
ata2425ds	3	3	—	—	—
ataupd2425-gainer	33	3	12	9	9
ataupd2425-pam	37	10	9	9	9
BIU-ONLP	17	5	4	4	4
DS@GT-bioasq-task6	1	1	—	—	—
DS@GT-BioNER	3	3	—	—	—
Graphwise-1	68	23	15	15	15
greenday	1	1	—	—	—
Gut-Instincts	28	16	4	4	4
GutUZH	4	4	—	—	—
ICUE	95	23	24	24	24
lasigeBioTM	10	2	—	4	4
LYX-DMIIP-FDU	4	1	1	1	1
NLPatVCU	40	5	15	10	10
ONTUG	4	—	2	1	1
Schemalink	4	1	1	1	1
ToGS	39	—	13	13	13

```
{
    "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.

Submitted runs must include all required fields for each document and entity and adhere strictly to valid JSON syntax. No specific ordering for documents or predictions is required.

4.1.2. Subtask 2 (BT-RE) Run Format

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

```
{
    "34870091": {
        "binary_tag_based_relations": [
            {
                "subject_label": "microbiome",
                "object_label": "human"
            }
        ]
    }
}
```

where:

- The top-level key (e.g. "34870091") is the PubMed ID of the document.
- `binary_tag_based_relations` is a list of relation objects.
- Each relation object represents a predicted binary tag-based relation and contains:
  - `subject_label`: the entity type of the relation's subject,
  - `object_label`: the entity type of the relation's object.

Submitted runs must include all required fields for each document and entity and adhere strictly to valid JSON syntax. No specific ordering for documents or predictions is required.

4.1.3. Subtask 3 (TT-RE) Annotation Format

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

```
{
    "34870091": {
        "ternary_tag_based_relations": [
            {
                "subject_label": "microbiome",
                "predicate": "located in",
                "object_label": "human"
            }
        ]
    }
}
```

where:

- The top-level key (e.g. "34870091") is the PubMed ID of the document.
- `ternary_tag_based_relations` is a list of relation objects.
- Each relation object represents a predicted ternary tag-based relation and contains:
  - `subject_label`: the entity type of the relation's subject,
  - `predicate`: the relation type between the subject and object,
  - `object_label`: the entity type of the relation's object.

Submitted runs must include all required fields for each document and entity and adhere strictly to valid JSON syntax. No specific ordering for documents or predictions is required.

4.1.4. Subtask 4 (TM-RE) Annotation Format

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

```
{
    "34870091": {
        "ternary_mention_based_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.
- ternary_mention_based_relations is a list of relation objects.
- Each relation object represents a predicted ternary mention-based 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.

Submitted runs must include all required fields for each document and entity and adhere strictly to valid JSON syntax. No specific ordering for documents or predictions is required.

4.1.5. Submission Upload

All runs must be submitted as a single ZIP archive named <teamID>_GutBrainIE_2025.zip. Within this archive, each run has to 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 T61 for NER, T621 for BT-RE, T622 for TT-RE, or T623 for TM-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 85 teams registered for the GutBrainIE2025 task, of which 17 submitted at least one run and thus participated in the evaluation.

In total, 391 runs were submitted: 101 for NER, 100 for BT-RE, and 95 each for TT-RE and TM-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-2025.

Team ID	Affiliation	Country	Repository	Paper
ata2425ds	University of Padua	Italy	https://github.com/andreastocco01/ATA2425	NA
ataupd2425-gainer	University of Padua	Italy	https://github.com/Vezzero/ataupd2425-gainer	[11]
ataupd2425-pam	University of Padua	Italy	https://github.com/Pami01/ataupd2425-pam	[12]
BIU-ONLP	Bar Ilan University	Israel	https://github.com/Ronke21/GutBrainIE_CLEF_2025_BIU_ONLP	[13]
DS@GT-bioasq-task6	NA	United States	https://github.com/rjmcoder/GutBrainIE_2025_Baseline	[14]
DS@GT-BioNER	NA	Canada	NA	NA
Graphwise-1	Graphwise	Bulgaria	Available upon request to the authors	[15]
greenday	Stony Brook University	United States	https://github.com/hpgupt/GutBrainIE-CLEF25	[16]
Gut-Instincts	Aalborg University	Denmark	https://github.com/P10-Natural-Language-Processing/Gut-Instincts	[17]
GutUZH	University of Zurich	Switzerland	https://github.com/VirginiaPoe/GutBrainIE_2025_PubMedBERTcrf	[18]
ICUE	University of Edinburgh	United Kingdom	https://github.com/chaeeunlee-io/bioasq2025	[19]
lasigeBioTM	LASIGE, Faculdade de Ciências, Universidade de Lisboa	Portugal	https://github.com/lasigeBioTM/BioASQ25-GutBrainIE	[20]
LYX-DMIIP-FDU	Fudan University	China	https://github.com/droidlyx/Team-LYX_DMIIP_FDU-Solution-for-GutBrainIE	[21]
NLPatVCU	Virginia Commonwealth University	United States	https://github.com/NLPatVCU/GutBrainIE	[22]
ONTUG	University of Technology, Graz; Ontotext, Bulgaria	Austria	NA	[23, 15]
Schemalink	Dept. of Computer Science, University of Milan	Italy	https://github.com/NLPatVCU/GutBrainIE	NA
ToGS	University of Technology, Graz	Austria	https://github.com/Dakantz/CLEANR	[23]

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 subtask 1 (NER): the set of entity types;
- for subtask 2 (BT-RE): the set of pairs *(subject label, object label)*;
- for subtasks 3 and 4 (TT-RE and TM-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 for all four GutBrainIE subtasks. This system is the same one used to generate the automatic annotations included in the Bronze fold of the training set (see Section 3.2).

The system consists of two independent modules: a NER module based on GLiNER [8], and a RE module based on ATLOP [10]. The NER module employs GLiNER, a bidirectional transformer encoder trained for instruction-based named entity recognition [8]. We used the NuNERZero checkpoint [24] and fine-tuned the model on the Platinum, 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 RE module uses ATLOP, a document-level relation extraction model that employs localized context pooling and adaptive thresholding [10]. 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, the Platinum, 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_2025_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. Complete scores for every submitted run can be found in the appendix.

5.1. Subtask 1 (NER) Results

Most participating teams in the NER subtask adopted supervised fine-tuning or transformer-based models pre-trained on large-scale biomedical corpora, with the most employed ones being PubMedBERT [25], BioBERT [26], BioLinkBERT [27], and ELECTRA [28]. In addition to these, several teams employed

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_id	total_runs	NER	BT-RE	TM-RE	TT-RE
ata2425ds	1/3 (33%)	1/3 (33%)	—	—	—
ataupd2425-gainer	0/33 (0%)	0/3 (0%)	0/12 (0%)	0/9 (0%)	0/9 (0%)
ataupd2425-pam	18/37 (49%)	0/10 (0%)	9/9 (100%)	9/9 (100%)	0/9 (0%)
BIU-ONLP	2/17 (12%)	0/5 (0%)	1/4 (25%)	1/4 (25%)	0/4 (0%)
DSGT-bioasq-task6	0/1 (0%)	0/1 (0%)	—	—	—
DSGT-BioNER	0/3 (0%)	0/3 (0%)	—	—	—
Graphwise-1	19/68 (28%)	1/23 (4%)	6/15 (40%)	6/15 (40%)	6/15 (40%)
greenday	1/1 (100%)	1/1 (100%)	—	—	—
Gut-Instincts	28/28 (100%)	16/16 (100%)	4/4 (100%)	4/4 (100%)	4/4 (100%)
GutUZH	4/4 (100%)	4/4 (100%)	—	—	—
ICUE	40/95 (42%)	9/23 (39%)	0/24 (0%)	19/24 (79%)	12/24 (50%)
lasigeBioTM	0/10 (0%)	0/2 (0%)	—	0/4 (0%)	0/4 (0%)
LYX-DMIIP-FDU	1/4 (25%)	1/1 (100%)	0/1 (0%)	0/1 (0%)	1/1 (100%)
NLPatVCU	5/40 (13%)	5/5 (100%)	0/15 (0%)	0/10 (0%)	0/10 (0%)
ONTUG	3/4 (75%)	—	1/2 (50%)	1/1 (100%)	1/1 (100%)
Schemalink	0/4 (0%)	0/1 (0%)	0/1 (0%)	0/1 (0%)	0/1 (0%)
ToGS	0/39 (0%)	—	0/13 (0%)	0/13 (0%)	0/13 (0%)

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	system_desc	Precision	Recall	F1	Precision	Recall	F1
GutUZH	2	AugEnsemble	0.7950	0.7736	**0.7613**	**0.8384**	0.8432	**0.8408**
Gut-Instincts	5eedev		0.7619	0.7813	0.7591	0.8286	0.8480	0.8382
NLPatVCU	ensemble1	ensemble1	0.8139	0.7161	0.7169	0.8255	0.8488	0.8370
ICUE	ensemble5	th10	**0.8216**	0.7451	0.7546	0.8369	0.8294	0.8331
LYX-DMIIP-FDU	run1	EnsembleBERT	0.7605	**0.7910**	0.7347	0.8020	**0.8513**	0.8259
ata2425ds	trf	tranformer	0.7199	0.7546	0.7217	0.7914	0.8432	0.8164
greenday	1	llmner	0.7368	0.7682	0.7471	0.7956	0.8278	0.8114
Graphwise-1	13	NERWise	0.7691	0.7398	0.7185	0.8066	0.7955	0.8010
BASELINE	Organizers	NuNerZero-Finetuned	0.6883	0.7690	0.7047	0.7639	0.8238	0.7927
ataupd2425-gainer	ma	trainplatinumandgold	0.5808	0.5322	0.5281	0.8333	0.7397	0.7837
DS@GT-bioasq-task6	1	glinerbiomed	0.6342	0.7849	0.6872	0.7337	0.8197	0.7743
DS@GT-BioNER	run2	pubmedbert	0.6731	0.6497	0.6469	0.7783	0.7437	0.7606
ataupd2425-pam	3	biosyn-sapbert-bc2gn-12	0.6400	0.7435	0.6763	0.6809	0.7745	0.7247
Schemalink	1	SchemaBasedMultiPrompt	0.4813	0.5038	0.4650	0.5547	0.5659	0.5602
BIU-ONLP	3	3_gliner_large_bio-v0.1	0.4393	0.3585	0.3711	0.4916	0.4721	0.4816
lasigeBioTM	R1	BENTMistral	0.2206	0.1034	0.0863	0.3471	0.1964	0.2509

GLiNER [8] fine-tuned on the training data. Ensemble approaches were widely utilized to improve effectiveness, often combining models trained with different data, seeds, and configurations.

While the majority of teams used the platinum, gold, and silver folds, a few also included the noisier bronze data, applying cleaning or re-weighting strategies. Some systems also incorporated additional knowledge from external corpora or pseudo-labeled texts to enhance training coverage.

A smaller number of teams experimented with prompt-based or zero-shot methods using Large Language Models (LLMs). These approaches avoided traditional supervised learning and relied on structured prompting and schema-guided extraction.

Overall, systems that combined strong biomedical backbones with fine-tuning and ensemble strategies tended to outperform others.

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

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Gut-Instincts	6219eedev3re		**0.5166**	0.6315	**0.5386**	0.6304	0.7532	**0.6864**
ONTUG	union	ElectraCLEANR	0.4185	0.4073	0.4057	0.7121	0.6104	0.6573
Graphwise-1	104	AtlopOnto	0.4043	0.3748	0.3832	0.7418	0.5844	0.6538
ataupd2425-pam	A7	RE-BiomedNLP-3NoRel-1epoch-COMPLETE_DATASET	0.4807	0.6091	0.4993	0.5671	0.7316	0.6389
BIU-ONLP	4	RobertaLarge	0.4632	0.3379	0.3713	0.7453	0.5195	0.6122
BASELINE	Organizers	Atlop-Finetuned	0.4650	0.3564	0.3864	**0.7584**	0.4892	0.5947
LYX-DMIIP-FDU	run1	BioLinkBERT	0.3637	0.4269	0.3688	0.6168	0.5714	0.5933
NLPatVCU	C18	mixedCNNWLabModel4Preds	0.3975	0.8419	0.5082	0.4381	0.8571	0.5798
Schemalink	1	gpt4re	0.3758	0.6573	0.4421	0.4531	0.7532	0.5659
ataupd2425-gainer	bp	trainplatinumandgold	0.3171	0.3254	0.2968	0.6150	0.4978	0.5502
ICUE	run17	biolinkbertl_pp	0.3558	**0.8790**	0.4751	0.3894	**0.9221**	0.5476
ToGS	hermes8bragreorder	CLEANR	0.2211	0.1304	0.1451	0.5701	0.2641	0.3609

5.2. Subtask 2 (BT-RE) Results

A discussion of the systems and methodologies employed for BT-RE is provided in the section dedicated to the TM-RE subtask (see Section 5.4), which offers an overview valid across all RE subtasks.

5.3. Subtask 3 (TT-RE) Results

Table 9
Performance metrics of each team's top run for TT-RE. For each evaluation metric, the best result is in bold, the second-best is underlined.

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Gut-Instincts	6229eedev3re		0.4663	0.6445	**0.5184**	0.6280	0.7572	**0.6866**
ataupd2425-pam	B7	RE-BiomedNLP-3NoRel-1epoch-COMPLETE_DATASET	0.4409	0.5704	0.4694	0.5853	0.7202	0.6458
ONTUG	union	ElectraCLEANR	0.4254	0.4025	0.4058	0.7059	0.5926	0.6443
Graphwise-1	105	AtlopOnto	0.4119	0.3709	0.3840	0.7326	0.5638	0.6372
ICUE	run22	biolinkbertl_pp	0.4011	0.7123	0.4879	0.4974	0.7860	0.6093
BIU-ONLP	4	RobertaLarge	0.4725	0.3288	0.3630	0.7362	0.4938	0.5911
BASELINE	Organizers	Atlop-Finetuned	**0.4729**	0.3421	0.3745	**0.7533**	0.4650	0.5751
NLPatVCU	C19	mixedCNNWLabModel4Preds	0.3810	**0.8005**	0.4868	0.4362	**0.8436**	0.5750
LYX-DMIIP-FDU	run1	BioLinkBERT	0.3625	0.4171	0.3549	0.5973	0.5432	0.5690
Schemalink	1	gpt4re	0.3756	0.6592	0.4437	0.4523	0.7613	0.5675
ataupd2425-gainer	td	trainplatinumandgold	0.3167	0.2315	0.2528	0.7405	0.3992	0.5187
ToGS	hermes8bragreorder	CLEANR	0.2261	0.1267	0.1414	0.5556	0.2469	0.3419
lasigeBioTM	R1	ConstParsing	0.0797	0.0622	0.0646	0.3929	0.0453	0.0812

A discussion of the systems and methodologies employed for TT-RE is provided in the section dedicated to the TM-RE subtask (see Section 5.4), which offers an overview valid across all RE subtasks.

5.4. Subtask 4 (TM-RE) Results

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

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Gut-Instincts	6239eedev3re		0.3310	0.4303	**0.3497**	0.4215	0.5147	**0.4635**
Graphwise-1	107	AtlopOnto	0.3323	0.2369	0.2603	0.4686	0.3097	0.3729
ICUE	run23	biolinkbertl_pp	0.2509	0.4239	0.2825	0.2858	0.5054	0.3651
LYX-DMIIP-FDU	run1	BioLinkBERT	0.2106	0.2418	0.1990	0.3682	0.3257	0.3457
ONTUG	union	ElectraCLEANR	0.2589	0.2293	0.2266	0.3529	0.3231	0.3373
BASELINE	Organizers	Atlop-Finetuned	**0.3514**	0.1829	0.2123	**0.4986**	0.2453	0.3288
Schemalink	1	gpt4re	0.2265	0.4088	0.2546	0.1948	0.4665	0.2749
ataupd2425-pam	C7	RE-BiomedNLP-3NoRel-1epoch-COMPLETE_DATASET	0.1940	0.2764	0.1982	0.2278	0.3432	0.2738
ataupd2425-gainer	tms	trainplatinumandgold	0.2203	0.1384	0.1538	0.4272	0.1810	0.2542
NLPatVCU	C11	ensembleWLabModel4Preds	0.1522	**0.5041**	0.2163	0.1423	**0.6005**	0.2300
BIU-ONLP	4	RobertaLarge	0.1171	0.0854	0.0879	0.2339	0.1461	0.1799
ToGS	hermes3bloraragreorder	CLEANR	0.0249	0.0180	0.0203	0.1702	0.0536	0.0815
lasigeBioTM	R1	ConstParsing	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Most participating teams approached RE as a supervised classification task, using fine-tuned biomedical transformers such as BioBERT [26], BioLinkBERT [27], PubMedBERT [25], and BioMedElectra [28].

Entity pairs were detected via upstream NER modules, explicitly marked in input texts and used to generate relation-specific training instances.

Some teams tackled RE at the document level, incorporating sampling strategies (e.g., negative sampling, class-weighted losses) and architectural enhancements (e.g., query-based encoders, hypergraph neural networks) to better capture long-tail relations. Data augmentation, input filtering, and relation predicate-based constraints were also employed to refine candidate relation sets.

Ensemble techniques, including majority voting and model fusion, were used by several top-performing teams to improve systems' effectiveness across the three RE subtasks.

Few teams experimented with prompt-based or zero-shot approaches using LLMs guided by structured templates or relation schemas, without any form of supervised training or fine-tuning.

Overall, the most effective submissions combined strong biomedical encoders with supervised fine-tuning and ensemble mechanisms.

6. Discussion

This section provides an overview of the approaches adopted by participating teams in the GutBrainIE-2025 task. We organize the discussion into two subsections: one dedicated to NER (subtask 1) and another covering the RE subtasks (subtasks 2, 3, and 4).

6.1. Subtask 1 (NER) Discussion

Han et al. [18] (Team GutUZH) fine-tuned a BioMedBERT model [29] augmented with a Conditional Random Field (CRF) layer to improve label dependency modeling [30]. Titles and abstracts were processed separately, with special tokens (`[TITLE]`, `[ABSTRACT]`) used to mark structural components of the documents.

The team experimented with multiple runs involving data augmentation and model ensembling. In one setup, they pseudo-labeled 500 additional abstracts using ensemble predictions, integrating them into a second training phase. Another variant trained on the full labeled set, including also bronze-quality annotations, while a final run retrained the top-performing model using only manual annotations (platinum, gold, silver sets) to reinforce the patterns learned from the most reliable examples.

Training employed weighted loss functions for class imbalance, mixed-precision optimization, and early stopping based on entity-level F1 score [31, 32]. Inference relied on Viterbi decoding [33], with evaluation using the seqeval library [34].

Andersen et al. [17] (Team Gut-Instincts) built a large ensemble system integrating multiple biomedical transformers, including BioLinkBERT [27], BioMedBERT [29], and BioMedElectra [28], with different decoding heads (dense layers, CRFs, LSTM-CRFs). In their runs, they combined from 3 to 17 models.

All available training data were used, including a cleaned version of the silver and bronze sets and, in some runs, also the development set.

Preprocessing included boundary corrections using manually crafted dictionaries, while training involved class-weighted losses to give more importance to high-quality data during optimization and a custom learning rate scheduler. Post-processing rules were used to merge overlapping or adjacent entities.

Taylor et al. [22] (Team NLPatVCU) submitted ensembles of fine-tuned GLiNER models specialized for biomedical NER [35]. These models differed in pretraining sources, training subsets, and configuration parameters.

Training data included all annotation tiers, and some models were additionally pretrained on external corpora such as BC5CDR [36]. To improve training stability, the team adopted GLiNER's probabilistic masking mechanism [8], selectively ignoring potentially mislabeled non-entity spans during training. In addition, focal loss was used to emphasize harder examples and counter class imbalance.

Ensemble predictions were constructed by combining the outputs of the three models. Model 1, based on GLiNER-BioMed [35], was trained on all annotation tiers and served as the primary model; Model 2 introduced a two-stage training pipeline with initial fine-tuning on BC5CDR [36] to improve

performance on disease-like entities; Model 3 reused the same training data as Model 1 but employed different focal loss parameters to adjust class sensitivity. Post-processing involved per-entity confidence thresholds and merging rules derived heuristically from the development set.

Lee et al. [19] (Team ICUE) explored both token classification and span-based approaches. Their primary models were transformer-based classifiers using IOB2 tagging [37] and ensembled predictions across 11 models trained separately with variations in architectural choices and span manipulation strategies.

Training data comprised platinum, gold, and silver sets, with preprocessing involving token alignment, label assignment, and filtering based on entity presence.

The team employed BioLinkBERT [27] and PubMedBERT [25] as models, while span strategies included union-span and bigger-span [38]. Some configurations further integrated PubTator annotations as training data [39].

Liu [21] (Team LYX-DMIIP-FDU) used a majority-vote ensemble of BioMedBERT [29], BioLinkBERT [27], and a clinical variant of XLM-RoBERTa [40]. Each model was fine-tuned in a multi-task learning setup, treating each entity class as a distinct prediction objective.

Input annotations were converted to PubTator format before training [39]. Models were trained on platinum, gold, silver, and development sets. During inference, span-level voting was applied to determine final entity labels. Specifically, after separate inference using each model, they used the average predicted probability of each token as the probability of each entity span, and filtered the predicted entity spans based on the total probability across all models

Team ata2425ds trained spaCy-based NER models using both static word embeddings and transformer backbones [41].

Two main pipelines were implemented: one using `en_core_web_lg` with tok2vec + NER layers, the other based on `en_core_web_trf` with RoBERTa as the underlying encoder [42, 43]. Models were trained on the full dataset, including bronze-quality annotations, with different tokenization and input cleaning configurations.

Preprocessing involved HTML tag removal using BeautifulSoup [44] and tokenization adjustments to preserve annotated spans.

Gupta et al. [16] (Team greenday) proposed a generation-based NER model by fine-tuning GPT-4.1-mini [45] to perform entity annotation using inline text markers, following the approach adopted by the GPT-NER framework [46].

Training was conducted via OpenAI's API on platinum and gold subsets, using specific prompts that directly instructed entity tagging. The team experimented with zero- and few-shot settings, utilizing a FAISS-based vector database of training examples for retrieval-augmented few-shot prompting [47, 48].

Post-processing involved recovering token-level entity spans from the annotated output by resolving discrepancies and misalignments introduced by inline annotations and hallucinations.

Datseris et al. [15] (Team Graphwise-1) developed an ensemble approach combining fine-tuned biomedical transformers, GLiNER [8], and data-augmentation strategies. Their pipeline integrates BioBERT [26], ELECTRA-based models [28], and GLiNER [8] fine-tuned on the full annotated dataset.

To mitigate data imbalance in low-resource categories, they applied a data augmentation strategy based on distant supervision. Specifically, they queried the PubMed API using MeSH-based queries tailored to each entity type. Retrieved abstracts were then annotated using multiple NER systems, including GLiNER [8] and BioBERT [26], and incorporated into an expanded bronze-quality collection.

To further improve system robustness, the team experimented with spaCy pipelines enhanced with domain-specific gazetteers [49].

Ensemble predictions were constructed by selecting the best-performing model for each entity type. Post-processing rules were applied to adjust entity boundaries based on systematic validation error analysis.

Piron et al. [11] (Team ataupd2425-gainer) trained GliNER-based models initialized from the NuNer_Zero checkpoint [24]. Training variants explored different dataset combinations: platinum+gold, platinum+gold+dev, and platinum+gold+silver.

Preprocessing involved concatenating titles and abstracts, applying the DeBERTa-v3-large tokenizer [50], and mapping entity offsets across fields. Training used cosine learning rate scheduling, fixed batch size (2), and variable training steps (6k-12k depending on the setting).

Mehta [14] (Team DS@GT-bioasq-task6) submitted a single run using the GLiNER-biomed checkpoint fine-tuned on platinum, gold, and silver annotations [35].

Post-processing involved a dictionary-based refinement using external biomedical lexicons to correct low-confidence or invalid predictions.

Team DS@GT-BioNER submitted three runs based on BioBERT [26] and PubMedBERT [25] models fine-tuned on platinum, gold, and silver folds. All annotations were converted to BIO format before training [51].

The first and second runs used BioBERT and PubMedBERT individually, while the third run ensembled their outputs. Models were trained with HuggingFace's default settings.

Pamio et al. [12] (Team ataupd2425-pam) explored CRF- and transformer-based models across ten runs. Transformer models included BioBERT [26], BioMedBERT [29], NuNER [24], SapBERT [52], and SciBERT [53].

Some models were trained with class-weighted loss functions to address label imbalance. CRF-based models used custom F1/F2 loss weighting strategies. For most of their submitted runs, models were trained on the full dataset (all training and development sets), with data preprocessed by parsing entities into token-label sequences.

Team Schemalink applied a schema-driven in-context learning approach using OpenAI's GPT-4o [45]. No supervised training was employed.

A LinkML schema derived from the ontology provided in the challenge materials was used to guide the LLM [54], along with the incorporation of few-shot examples in the prompt. For each entity class, they generated a separate prompt and used OpenAI's `response_format` field to enforce structured extraction. UTF-8 normalization was applied as a preprocessing step to improve model input compatibility.

Keinan et al. [13] (Team BIU-ONLP) fine-tuned five variants of GLiNER [8] on platinum, gold, and silver tiers. Preprocessing included lowercasing and space normalization.

Models differed by GLiNER backbone (e.g., domain-specific or multilingual). All were trained with the same hyperparameters: 384-tokens input, learning rate of 5e-5, batch size of 8, and 3k training epochs. The confidence threshold has been fixed to 0.9 to retain only highly reliable predictions.

Conceição et al. [20] (Team LasigeBioTM) submitted two zero-shot runs using Mistal-7B [55].

The first run used the BENT tool [56] to insert inline entity annotations with unique IDs and label types, which were then passed to Mistral for processing. The second run applied Mistral directly to raw texts without tagging. No fine-tuning or labeled data was used in either run.

6.2. Subtasks 2,3, and 4 (RE) Discussion

Andersen et al. [17] (Team Gut-Instincts) extended their ensemble-based approach to all three RE subtasks. Their approach combined fine-tuned transformers (BioLinkBERT [27], BioMedBERT [29], BioMedElectra [28]) with specific adaptations to accommodate task-specific output structures.

To improve training quality, they cleaned the silver and bronze datasets by correcting or removing entity spans with misalignments and filtering out documents having more than 100 relations annotated. Candidate entity pairs were marked in input texts, and a 10:1 negative sampling ratio was used to balance the training data.

Training used class-weighted loss and a custom learning rate schedule. Final predictions were generated via ensemble voting across the three top-performing models per configuration.

Kantz et al. [23] (Team ToGS) submitted runs for all RE subtasks using a hybrid system combining retrieval-augmented generation (RAG) [57], LoRA fine-tuning [58], transformer-based models such as BioMedElectra [28] and Hermes-3 (`LLaMA-3.2 3B` and `LLaMA-3.1 8B` variants) [59], and prompting with GPT-4o-mini [45].

Prompts were dynamically built using training examples retrieved from a VectorDB and reordered to prioritize high-quality (platinum and gold) annotations. In addition to prompting, LoRA-based

fine-tuning was applied to improve models' specialization efficiently [58].

Furthermore, Teams Togs [23] and Graphwise-1 [15] submitted collaborative runs as Team ONTUG. Here, the BiomedElectra [28] model was first fine-tuned on the binary relation extraction task (BT-RE) and subsequently adapted for the mention-level task (TM-RE), leveraging shared entity representations across subtasks.

All models were trained for 100 epochs with the same hyperparameters: batch size of 2, gradient accumulation of 2 steps, learning rate of 5e-5, and a warm-up ratio of 0.06. Two different output fusion strategies (union and intersection) were evaluated to assess the impact of conservative and inclusive inference fusion.

Datseris et al. [15] (Team Graphwise-1) participated in all three RE subtasks, exploring transformer- and encoder-based classifiers.

For encoder models, BioMedElectra [28] and XLM-RoBERTa [40] were fine-tuned sequentially for BT-RE, TT-RE, and TM-RE using consistent settings (up to 200 epochs, learning rate of 5e-5). Some variants experimented with masked language modeling pre-training [60].

They also employed fine-tuned REBEL-large [61] to perform end-to-end relation generation.

Pamio et al.[12] (Team ataupd2425-pam) submitted models for all RE subtasks using transformer classifiers trained on relation-centric instances.

Entity mentions were extracted via upstream NER (e.g., SapBERT [52], NuNER [24]) and injected into text using marker tokens. These instances were then used to fine-tune multiple RE models, trained for one epoch on the full dataset (platinum, gold, silver, bronze, and development sets). No significant run-specific modifications or hyperparameter variations were reported across submissions.

Keinan et al. [13] (Team BIU-ONLP) submitted twelve runs across all RE subtasks based on fine-tuning ATLOP [10] on different language models, including SapBERT [52], BioBERT [26], and RoBERTa [43].

Each model was trained using a standardized configuration (learning rate of 5e-5, batch size of 4, 500 training epochs, warmup ratio of 0.06).

Only the platinum, gold, and silver folds were used. Preprocessing involved lowercasing and whitespace normalization. No ensemble, augmentation, or post-processing strategies were applied.

Liu [21] (Team LYX-DMIIP-FDU) used a unified binary classification approach across all RE subtasks. Entity pairs were filtered by type compatibility and distance (<200 characters) and formatted in PubTator style with markers and contextual windows [39].

BioLinkBERT [27] was employed as the backbone, fine-tuned using platinum, gold, silver, and development sets. The same model and pipeline were reused across all RE subtasks, with no task-specific variation or augmentation.

Taylor et al. [22] (Team NLPatVCU) explored two families of models: sentence-level CNN classifiers [62] and document-level Hypergraph Neural Networks (HGNN) [63].

CNNs were trained on sentences labeled with relations and sampled sentences with no relation, using platinum, gold, and silver training datasets. Entity spans were derived from prior NER submissions, and final outputs were aggregated via ensemble logic.

HGNNs modeled entities and their interactions as nodes and hyperedges, using BioBERT embeddings [26] and a hypergraph convolution layer [64]. The outputs obtained with these approaches supported BT-RE and TT-RE predictions, but did not address TM-RE predictions.

Team Schemalink used prompting-based approaches via OpenAI's GPT-4o [45], operating in a fully zero-shot setting.

Entity mentions identified by GLiNER [8] were inserted into sentence-level prompts using custom tags. Prompts included few-shot examples from the platinum set and targeted predefined relation patterns (e.g., `[bacteria] LOCATED IN [host]`). The same system was applied to all subtasks with no fine-tuning or augmentations.

Piron et al. [11] (Team ataupd2425-gainer) submitted runs for all three RE subtasks using PubMedBERT [25] and BioBERT [26] trained via HuggingFace's classification pipeline.

Entity spans were marked using `[E1]` and `[E2]` tokens. Sentences were tokenized to a max length of 256 or 356, and negative sampling (0.2 or 0.3) was applied.

Across runs, models were fine-tuned for 5 to 8 epochs on stratified 80/20 train–validation splits with batch sizes between 8 and 12, and learning rates ranging from 1e-5 to 2e-5. Training data spanned different combinations of the platinum, gold, silver, and dev sets. No ensembles or post-processing were used.

Lee et al. [19] (Team ICUE) participated in all RE subtasks by framing RE as binary classification over entity combinations using a query-based BioLinkBERT model [27]. Inputs were constructed by inserting tagged entities and a natural language query representing the candidate relations.

Balanced sampling was used to mitigate class imbalance. Some runs included second-stage reasoning with a distilled LLM trained on synthetic binary-choice prompts: given a candidate relation and supporting context, the LLM is asked whether the relation holds, choosing between a positive or negative restatement. The LLM confidence was then fused with classified logits.

Ensemble strategies were also explored, with final outputs selected based on majority voting across models trained on distinct splits or using different sampling thresholds.

Conceição et al. [20] (Team LasigeBioTM) participated in TT-RE and TM-RE using a zero-shot approach combining BENT for entity tagging [56] and Mistral-7B for relation extraction [55].

Tagged inputs contained nested entity labels and IDs. In some runs, syntactic features (dependency paths, constituency parses) were added using spaCy [49]. All configurations relied uniquely on prompting and required no model fine-tuning or training data.

7. Conclusions and Future Works

GutBrainIE-2025 marked the first 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 first edition saw 85 teams registering and 17 teams submitting a total of 395 runs. Participants tackled a diverse set of subtasks, from Named Entity Recognition to increasingly fine-grained Relation Extraction, with results highlighting the effectiveness of ensemble-based methods and biomedical transformers fine-tuned on domain-specific data.

The released dataset, including over 1600 annotated PubMed abstracts and stratified into annotation quality tiers, represents a valuable resource for training and evaluating biomedical NLP systems.

As future work, we plan to further improve the overall quality of the dataset by manually reviewing and annotating the current bronze fold, currently composed of fully automatic and not revised annotations. Additionally, we will leverage the pool of submitted predictions to identify possible annotation errors, such as wrongly annotated entities or relations as well as missing annotations that may have been overlooked during the annotation process. Finally, we aim to extend the task by incorporating entity linking. This will enable the inclusion of two additional subtasks: entity linking itself, and the classical NLP task of Relation Extraction framed at the concept level rather than at the mention level.

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-4o and Grammarly in order to: Grammar and spelling check. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, N. V. Loukachevitch, A. Sakhovskiy, E. Tutubalina, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, L. Menotti, G. Silvello, G. Paliouras, BioASQ at CLEF2025: The Thirteenth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V, volume 15576 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 407–415. URL: https://doi.org/10.1007/978-3-031-88720-8_61. doi:10.1007/978-3-031-88720-8_61.

[7] N. Bach, S. Badaskar, A review of relation extraction, Literature review for Language and Statistics II 2 (2007) 1–15.

[8] 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.

[9] 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).

[10] 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.

[11] S. Piron, G. M. Di Nunzio, Named Entity Recognition with GLiNER and Relation Extraction with LLMs, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[12] L. Pamio, G. M. Di Nunzio, BioASQ task GutBrainIE 2025 Task 6: Comparing CRF vs BERT Models for Named Entity Recognition and Relation Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[13] R. Keinan, A. D. N. Cohen, R. Tsarfaty, From Named Entities to Relations: End-to-End Biomedical Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[14] R. Mehta, Enhancing Biomedical Named Entity Recognition using GLiNER-BioMed with Targeted

Dictionary-Based Post-processing for BioASQ 2025 task 6, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[15] A. Datseris, M. Kuzmanov, I. Nikolova-Koleva, D. Taskov, S. Boytcheva, Graphwise @ CLEF-2025 GutBrainIE: Towards Automated Discovery of Gut-Brain Interactions: Deep Learning for NER and Relation Extraction from PubMed Abstracts, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[16] H. P. Gupta, R. Banerjee, LLMs for Biomedical NER, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[17] L. R. Andersen, M. I. Gardshodn, M. H. Dolmer, 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 CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[18] J. Han, Y. Liu, GutUZH at CLEF2025 BioASQ Task 6: a method of SOTA performance with the best results at GutBrainIE NER subtask 1, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[19] C. Lee, S. Doneva, M. Rodriguez-Cubillos, E. Castagnari, A. Lain, J. Posma, T. I. Simpson, Understanding Gut-Brain Interplay in Scientific Literature: A Hybrid Approach from Classification to Generative LLM Reasoning, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[20] S. I. R. Conceição, P. R. C. Lopes, F. M. Couto, lasigeBioTM at BioASQ25 Task GutBrainIE - Lean Large language models with syntactic features, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[21] Y. Liu, LYX_DMIIP_FDU at BioASQ 2025: Utilizing BERT embeddings for biomedical text mining, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[22] S. Taylor, C. Dil, A. Shah, Jannat, C. Oldham, A. Upadhyay, J. Varughese, N. Yazbeck, B. T. McInnes, NLP@VCU at BioASQ2025: Information Extraction on the GutBrainIE dataset, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[23] B. Kantz, P. Waldert, S. Lengauer, T. Schreck, Constrained Linked Entity ANnotation using RAG (CLEANR), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.

[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] 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.

[26] 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).

[27] M. Yasunaga, J. Leskovec, P. Liang, LinkBERT: Pretraining Language Models with Document Links, in: Association for Computational Linguistics (ACL), 2022.

[28] 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.

[29] 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.
[30] C. Sutton, A. McCallum, et al., An introduction to conditional random fields, Foundations and Trends® in Machine Learning 4 (2012) 267–373.
[31] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, et al., Mixed precision training, arXiv preprint arXiv:1710.03740 (2017).
[32] Z. Ji, J. Li, M. Telgarsky, Early-stopped neural networks are consistent, Advances in Neural Information Processing Systems 34 (2021) 1805–1817.
[33] A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE transactions on Information Theory 13 (2003) 260–269.
[34] H. Nakayama, seqeval: A Python framework for sequence labeling evaluation, 2018. URL: https://github.com/chakki-works/seqeval, software available from https://github.com/chakki-works/seqeval.
[35] 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.
[36] J. Li, Y. Sun, R. J. Johnson, D. Sciaky, C.-H. Wei, R. Leaman, A. P. Davis, C. J. Mattingly, T. C. Wiegers, Z. Lu, BioCreative V CDR task corpus: a resource for chemical disease relation extraction, Database 2016 (2016).
[37] L. A. Ramshaw, M. P. Marcus, Text chunking using transformation-based learning, in: Natural language processing using very large corpora, Springer, 1999, pp. 157–176.
[38] C. Liu, H. Fan, J. Liu, Span-based nested named entity recognition with pretrained language model, in: Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II 26, Springer, 2021, pp. 620–628.
[39] C.-H. Wei, H.-Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration, Nucleic acids research 41 (2013) W518–W522.
[40] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).
[41] H. Shelar, G. Kaur, N. Heda, P. Agrawal, Named entity recognition approaches and their comparison for custom ner model, Science & Technology Libraries 39 (2020) 324–337.
[42] R. Weischedel, M. Palmer, M. Marcus, E. Hovy, S. Pradhan, L. Ramshaw, N. Xue, A. Taylor, J. Kaufman, M. Franchini, et al., Ontonotes release 5.0 ldc2013t19, Linguistic Data Consortium, Philadelphia, PA 23 (2013) 20.
[43] 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).
[44] L. Richardson, Beautiful soup documentation, 2007.
[45] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[46] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-ner: Named entity recognition via large language models, arXiv preprint arXiv:2304.10428 (2023).
[47] 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).
[48] M. Dong, Z. Cheng, C. Luo, T. He, Retrieval-Augmented Generation for Large Language Model based Few-shot Chinese Spell Checking, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 10767–10780.
[49] Y. Vasiliev, Natural language processing with Python and spaCy: A practical introduction, No Starch Press, 2020.
[50] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021. arXiv:2111.09543.
[51] L. Ramshaw, M. Marcus, Text Chunking using Transformation-Based Learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107/.
[52] 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.

[53] I. Beltagy, K. Lo, A. Cohan, SciBERT: Pretrained Language Model for Scientific Text, in: EMNLP, 2019. arXiv:arXiv:1903.10676.

[54] S. A. Moxon, H. Solbrig, D. R. Unni, D. Jiao, R. M. Bruskiewich, J. P. Balhoff, G. Vaidya, W. D. Duncan, H. Hegde, M. Miller, et al., The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics, ICBO 3073 (2021) 148–151.

[55] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, 2023. URL: https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

[56] P. Ruas, F. M. Couto, NILINKER: attention-based approach to NIL entity linking, Journal of Biomedical Informatics 132 (2022) 104137.

[57] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[58] 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.

[59] R. Teknium, J. Quesnelle, C. Guang, Hermes 3 technical report, arXiv preprint arXiv:2408.11857 (2024).

[60] K. Sinha, R. Jia, D. Hupkes, J. Pineau, A. Williams, D. Kiela, Masked language modeling and the distributional hypothesis: Order word matters pre-training for little, arXiv preprint arXiv:2104.06644 (2021).

[61] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204.

[62] A. Jarrahi, R. Mousa, L. Safari, SLCNN: Sentence-level convolutional neural network for text classification, arXiv preprint arXiv:2301.11696 (2023).

[63] Y. Feng, H. You, Z. Zhang, R. Ji, Y. Gao, Hypergraph neural networks, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 3558–3565.

[64] S. Bai, F. Zhang, P. H. Torr, Hypergraph convolution and hypergraph attention, Pattern Recognition 110 (2021) 107637.

A. Subtask 6.1 (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**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BASELINE	Organizers	NuNerZero-Finetuned	0.6883	0.7690	0.7047	0.7639	0.8238	0.7927
BIU-ONLP	1	1_multi_pii-v1	0.4627	0.3687	0.3846	0.4908	0.4721	0.4813
BIU-ONLP	2	2_gliner_medium-v2.1	0.4049	0.3717	0.3864	0.4866	0.4568	0.4712
BIU-ONLP	3	3_gliner_large_bio-v0.1	0.4393	0.3585	0.3711	0.4916	0.4721	0.4816
BIU-ONLP	4	4_gliner_large-v2.1	0.4029	0.3710	0.3842	0.4893	0.4632	0.4759
BIU-ONLP	5	5_gliner_large_bio-v0.2	0.4488	0.3633	0.3775	0.4961	0.4632	0.4791
DS@GT-BioNER	run1	biobert	0.6438	0.5901	0.6020	0.7392	0.7057	0.7221
DS@GT-BioNER	run2	pubmedbert	0.6731	0.6497	0.6469	0.7783	0.7437	0.7606
DS@GT-BioNER	run3	ensemble	0.6203	0.6467	0.6226	0.7341	0.7478	0.7409
DS@GT-bioasq-task6	1	glinerbiomed	0.6342	0.7849	0.6872	0.7337	0.8197	0.7743
Graphwise-1	10	NERWise	0.6964	0.7200	0.7005	0.7859	0.7922	0.7890
Graphwise-1	11	NERWise	0.6913	0.7173	0.6988	0.7533	0.7777	0.7653
Graphwise-1	12	NERWise	0.6267	0.6175	0.5822	0.6437	0.6572	0.6504
Graphwise-1	13	NERWise	0.7691	0.7398	0.7185	0.8066	0.7955	0.8010
Graphwise-1	14	NERWise	0.7159	0.6657	0.6739	0.7900	0.7631	0.7763
Graphwise-1	15	NERWise	0.7083	0.6051	0.6185	0.7004	0.6823	0.6912
Graphwise-1	16	NERWise	0.6930	0.6181	0.6165	0.6738	0.6880	0.6808
Graphwise-1	17	NERWise	0.6940	0.6942	0.6912	0.7690	0.7777	0.7733
Graphwise-1	18	NERWise	0.5646	0.6265	0.5832	0.6770	0.7065	0.6915
Graphwise-1	19	NERWise	0.6938	0.7363	0.6978	0.7565	0.7939	0.7748
Graphwise-1	1	NERWise	0.7290	0.6432	0.6691	0.7886	0.7389	0.7629
Graphwise-1	20	NERWise	0.7364	0.6264	0.6473	0.7663	0.7316	0.7486
Graphwise-1	21	ONTO-Bio-GPT	0.4071	0.5702	0.4598	0.4593	0.6015	0.5208
Graphwise-1	22	ONTO-Bio-GPT	0.3763	0.5657	0.4328	0.4528	0.6281	0.5262
Graphwise-1	23	ONTO-Bio-GPT	0.5678	0.4371	0.4634	0.6387	0.4559	0.5321
Graphwise-1	2	NERWise	0.6217	0.5839	0.5967	0.7257	0.7316	0.7287
Graphwise-1	3	NERWise	0.6612	0.5664	0.5774	0.6797	0.6621	0.6708
Graphwise-1	4	NERWise	0.6964	0.7200	0.7005	0.7859	0.7922	0.7890
Graphwise-1	5	NERWise	0.6554	0.6850	0.6666	0.7314	0.7704	0.7504
Graphwise-1	6	NERWise	0.6056	0.6128	0.5694	0.6158	0.6492	0.6320
Graphwise-1	7	NERWise	0.7556	0.6326	0.6596	0.7976	0.7486	0.7723
Graphwise-1	8	NERWise	0.6800	0.6172	0.6386	0.7512	0.7470	0.7491
Graphwise-1	9	NERWise	0.6800	0.6172	0.6386	0.7512	0.7470	0.7491
Gut-Instincts	11ee		0.7877	0.7681	0.7466	0.8207	0.8327	0.8266
Gut-Instincts	11eedev		0.7860	0.7713	0.7459	0.8255	0.8416	0.8335
Gut-Instincts	13ee		0.7503	0.7691	0.7464	0.8243	0.8343	0.8292
Gut-Instincts	13eedev		0.7885	0.7691	0.7469	0.8261	0.8407	0.8333
Gut-Instincts	15ee		0.7493	0.7660	0.7442	0.8217	0.8310	0.8264
Gut-Instincts	15eedev		0.7843	0.7738	0.7452	0.8303	0.8424	0.8363
Gut-Instincts	17ee		0.7503	0.7633	0.7433	0.8225	0.8319	0.8272
Gut-Instincts	17eedev		0.7869	0.7698	0.7464	0.8266	0.8399	0.8332
Gut-Instincts	3ee		0.7308	0.7626	0.7393	0.8121	0.8351	0.8234
Gut-Instincts	3eedev		0.7412	0.7669	0.7387	0.8182	0.8367	0.8273
Gut-Instincts	5ee		0.7854	0.7669	0.7453	0.8192	0.8351	0.8271
Gut-Instincts	5eedev		0.7619	0.7813	0.7591	0.8286	0.8480	0.8382
Gut-Instincts	7ee		0.7525	0.7663	0.7460	0.8240	0.8367	0.8303
Gut-Instincts	7eedev		0.7774	0.7719	0.7408	0.8218	0.8424	0.8319
Gut-Instincts	9ee		0.7958	0.7700	0.7522	0.8288	0.8375	0.8331
Gut-Instincts	9eedev		0.7756	0.7696	0.7382	0.8219	0.8432	0.8324
GutUZH	1	PubMedBERTcrf	0.6877	0.7404	0.7100	0.8021	0.8222	0.8120
GutUZH	2	AugEnsemble	0.7950	0.7736	0.7613	0.8384	0.8432	**0.8408**
GutUZH	3	Ensemble	0.8023	0.7678	0.7634	0.8281	0.8375	0.8328
GutUZH	4	EnsembleContGoodQuality	0.8029	0.7760	**0.7686**	0.8269	0.8456	0.8361
ICUE	ensemble10	th7pubtator	0.7698	0.7756	0.7350	0.7966	0.8456	0.8204
ICUE	ensemble11	th10pubtator	0.8138	0.7449	0.7471	0.8274	0.8254	0.8264
ICUE	ensemble1	biggerspan	0.5483	0.7590	0.6253	0.6776	0.8529	0.7552
ICUE	ensemble2	unionspan	0.5483	0.7590	0.6253	0.6776	0.8529	0.7552
ICUE	ensemble3	th2	0.6386	**0.7951**	0.6848	0.7178	**0.8658**	0.7849
ICUE	ensemble4	th4	0.7253	0.7916	0.7170	0.7629	0.8634	0.8100
ICUE	ensemble5	th10	**0.8216**	0.7451	0.7546	0.8369	0.8294	0.8331
ICUE	ensemble6	biggerspanpubtator	0.5445	0.7566	0.6208	0.6742	0.8496	0.7518
ICUE	ensemble7	unionspanpubtator	0.5445	0.7566	0.6208	0.6742	0.8496	0.7518
ICUE	ensemble8	th2pubtator	0.6336	0.7943	0.6789	0.7138	0.8610	0.7805
ICUE	ensemble9	th4pubtator	0.7204	0.7914	0.7108	0.7582	0.8593	0.8056
ICUE	gliner1	v1	0.6620	0.7688	0.6869	0.7431	0.8278	0.7832
ICUE	gliner2	v2	0.7262	0.7668	0.7046	0.7656	0.8238	0.7936
ICUE	gliner3	v3	0.7691	0.7398	0.7185	0.8066	0.7955	0.8010
ICUE	gliner4	v4	0.8067	0.7052	0.7247	0.8383	0.7672	0.8012
ICUE	gliner5	v5pubtator	0.7206	0.7672	0.6988	0.7592	0.8205	0.7887
ICUE	gliner6	v6pubtator	0.7355	0.7535	0.7007	0.7764	0.8084	0.7921
ICUE	single1	pubmedbertl	0.7458	0.7546	0.7157	0.7688	0.8278	0.7972
ICUE	single2	pubmedbertl	0.7095	0.7697	0.6946	0.7522	0.8367	0.7922

(Table 11 continued)

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ICUE	single3	biolinkbert	0.6408	0.7215	0.6700	0.7695	0.8124	0.7904
ICUE	single4	pubmedbertb	0.6164	0.6851	0.6445	0.7477	0.8027	0.7743
ICUE	single5	pubmedbertb	0.6130	0.7460	0.6653	0.7453	0.8351	0.7876
ICUE	single6	biolinkbertpubtator	0.6265	0.7439	0.6672	0.7480	0.8230	0.7837
LYX-DMIIP-FDU	run1	EnsembleBERT	0.7605	0.7910	0.7347	0.8020	0.8513	0.8259
NLPatVCU	ensemble1	ensemble1	0.8139	0.7161	0.7169	0.8255	0.8488	0.8370
NLPatVCU	ensemble2	ensemble2	0.7747	0.7313	0.7464	0.8468	0.7955	0.8203
NLPatVCU	ensemble3	ensemble3	0.7773	0.7395	0.7541	0.8480	0.8027	0.8248
NLPatVCU	model4	model4	0.7149	0.6441	0.6725	**0.8573**	0.7866	0.8204
NLPatVCU	model6	model6	0.7134	0.6359	0.6628	0.8566	0.7777	0.8153
Schemalink	1	SchemaBasedMultiPrompt	0.4813	0.5038	0.4650	0.5547	0.5659	0.5602
ata2425ds	HTMLremoval		0.5032	0.3939	0.4111	0.5680	0.5166	0.5411
ata2425ds	hyperparams		0.6552	0.5994	0.6125	0.7714	0.7203	0.7450
ata2425ds	trf	tranformer	0.7199	0.7546	0.7217	0.7914	0.8432	0.8164
ataupd2425-gainer	ma	trainplatinumandgold	0.5808	0.5322	0.5281	0.8333	0.7397	0.7837
ataupd2425-gainer	md	trainplatinumgolddev	0.4054	0.5416	0.4569	0.6397	0.7106	0.6733
ataupd2425-gainer	ms	trainplatinumgoldsilver	0.3889	0.5505	0.4511	0.6332	0.7243	0.6757
ataupd2425-pam	10	customCRF	0.4147	0.3380	0.3472	0.4098	0.4390	0.4239
ataupd2425-pam	1	biobert-base-cased-v1.2-14-CW-xtreme	0.6468	0.6804	0.6259	0.6720	0.7421	0.7053
ataupd2425-pam	2	biosyn-sapbert-bc2gn-8	0.6447	0.7383	0.6856	0.6778	0.7736	0.7225
ataupd2425-pam	3	biosyn-sapbert-bc2gn-12	0.6400	0.7435	0.6763	0.6809	0.7745	0.7247
ataupd2425-pam	4	BiomedNLP-BiomedBERT	0.6097	0.7079	0.6391	0.6571	0.7623	0.7058
ataupd2425-pam	5	NuNerv2.0-22-CW-xtreme	0.6256	0.7052	0.6186	0.6564	0.7567	0.7030
ataupd2425-pam	6	scibert-47	0.6554	0.6987	0.6406	0.6736	0.7607	0.7145
ataupd2425-pam	7	scibert-27	0.6350	0.6997	0.6256	0.6641	0.7607	0.7091
ataupd2425-pam	8	customCRF-LowF	0.3790	0.2997	0.3174	0.4935	0.3670	0.4210
ataupd2425-pam	9	customCRF-LowF	0.3605	0.3566	0.3470	0.4917	0.4527	0.4714
greenday	1	llmner	0.7368	0.7682	0.7471	0.7956	0.8278	0.8114
lasigeBioTM	R1	BENTMistral	0.2206	0.1034	0.0863	0.3471	0.1964	0.2509
lasigeBioTM	R1	MistralBaseline	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

B. Subtask 6.2.1 (BT-RE) Overall Results

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

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BASELINE	Organizers	Atlop-Finetuned	0.4650	0.3564	0.3864	0.7584	0.4892	0.5947
BIU-ONLP	1	SapBERT	0.3846	0.3598	0.3545	0.6554	0.5022	0.5686
BIU-ONLP	2	BertBaseCased	0.4383	0.2912	0.3273	0.7955	0.4545	0.5785
BIU-ONLP	3	BioBert1.1PubMed	0.4309	0.2965	0.3293	0.7519	0.4199	0.5389
BIU-ONLP	4	RobertaLarge	0.4632	0.3379	0.3713	0.7453	0.5195	0.6122
Graphwise-1	1	ONTO-REBEL	0.3626	0.3022	0.3091	0.6303	0.4502	0.5253
Graphwise-1	2	ONTO-REBEL	0.4226	0.3767	0.3680	0.6859	0.5671	0.6209
Graphwise-1	100	AtlopOnto	0.1581	0.0842	0.1063	0.5000	0.0779	0.1348
Graphwise-1	101	AtlopOnto	0.1581	0.0842	0.1063	0.5000	0.0779	0.1348
Graphwise-1	102	AtlopOnto	0.0581	0.0233	0.0326	0.4286	0.0130	0.0252
Graphwise-1	103	AtlopOnto	0.2205	0.1524	0.1744	0.6122	0.1299	0.2143
Graphwise-1	104	AtlopOnto	0.4043	0.3748	0.3832	0.7418	0.5844	0.6538
Graphwise-1	105	AtlopOnto	0.3526	0.2547	0.2832	0.7237	0.4762	0.5744
Graphwise-1	106	AtlopOnto	0.4424	0.3987	0.4088	0.6842	0.5628	0.6176
Graphwise-1	107	AtlopOnto	0.4292	0.3529	0.3699	0.7062	0.5411	0.6127
Graphwise-1	108	AtlopOnto	0.3902	0.3011	0.3238	0.7375	0.5108	0.6036
Graphwise-1	109	AtlopOnto	0.4115	0.3244	0.3498	0.7256	0.5152	0.6025
Graphwise-1	12	ONTO-Bio-GPT	0.3151	0.4174	0.3359	0.4689	0.5541	0.5079
Graphwise-1	13	ONTO-Bio-GPT	0.3644	0.4725	0.3819	0.4750	0.5758	0.5205
Graphwise-1	14	ONTO-Bio-GPT	0.4723	0.5229	0.4494	0.5316	0.5455	0.5385
Gut-Instincts	6219ee3re		**0.5336**	0.6131	0.5380	0.6450	0.7316	0.6856
Gut-Instincts	6219ee3redev		0.5226	0.5896	0.5163	0.6350	0.7229	0.6761
Gut-Instincts	6219eedev3re		0.5166	0.6315	**0.5386**	0.6304	0.7532	**0.6864**
Gut-Instincts	6219eedev3redev		0.5061	0.5927	0.5102	0.6236	0.7316	0.6733
ICUE	run10	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run11	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run12	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run13	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run14	biolinkbertl_pp	0.3402	**0.9932**	0.4659	0.3207	**0.9870**	0.4841
ICUE	run15	biolinkbertl_pp	0.3461	0.8973	0.4667	0.3731	0.9351	0.5333
ICUE	run16	biolinkbertl_pp	0.3461	0.8973	0.4667	0.3731	0.9351	0.5333
ICUE	run17	biolinkbertl_pp	0.3558	0.8790	0.4751	0.3894	0.9221	0.5476
ICUE	run18	biolinkbertl_pp	0.3558	0.8790	0.4751	0.3894	0.9221	0.5476
ICUE	run19	biolinkbertl_pp	0.3402	**0.9932**	0.4659	0.3207	**0.9870**	0.4841
ICUE	run1	biolinkbertl	0.3402	**0.9932**	0.4659	0.3207	**0.9870**	0.4841
ICUE	run20	biolinkbertl_pp	0.3402	**0.9932**	0.4659	0.3207	**0.9870**	0.4841
ICUE	run21	biolinkbertl_pp	0.3402	**0.9932**	0.4659	0.3207	**0.9870**	0.4841
ICUE	run22	biolinkbertl_pp	0.3443	**0.9932**	0.4686	0.3225	**0.9870**	0.4861
ICUE	run23	biolinkbertl_pp	0.3443	**0.9932**	0.4686	0.3225	**0.9870**	0.4861
ICUE	run24	biolinkbertl_pp	0.3443	**0.9932**	0.4686	0.3225	**0.9870**	0.4861
ICUE	run2	biolinkbertl	0.3443	**0.9932**	0.4686	0.3225	**0.9870**	0.4861
ICUE	run3	biolinkbertl	0.3402	**0.9932**	0.4659	0.3207	**0.9870**	0.4841
ICUE	run4	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run5	biolinkbertl	0.3402	**0.9932**	0.4659	0.3207	**0.9870**	0.4841
ICUE	run6	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run7	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run8	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
ICUE	run9	biolinkbertl	0.3352	0.8558	0.4521	0.3682	0.9307	0.5276
LYX-DMIIP-FDU	run1	BioLinkBERT	0.3637	0.4269	0.3688	0.6168	0.5714	0.5933
NLPatVCU	C0	ensembleWLabEnsemble1Preds	0.4062	0.8537	0.5198	0.4274	0.8788	0.5751
NLPatVCU	C12	mixedCNNWLabEnsemble1Preds	0.4021	0.8611	0.5132	0.4184	0.8874	0.5687
NLPatVCU	C15	mixedCNNWLabEnsemble2Preds	0.3934	0.8496	0.5049	0.4301	0.8658	0.5747
NLPatVCU	C18	mixedCNNWLabModel4Preds	0.3975	0.8419	0.5082	0.4381	0.8571	0.5798
NLPatVCU	C21	platPlusCNNWLabEnsemble1Preds	0.3333	0.6866	0.4247	0.4244	0.8139	0.5579
NLPatVCU	C24	platPlusCNNWLabEnsemble2Preds	0.3343	0.6703	0.4226	0.4312	0.8009	0.5606
NLPatVCU	C27	platPlusCNNWLabModel4Preds	0.3392	0.6626	0.4263	0.4410	0.7922	0.5666
NLPatVCU	C3	ensembleWLabEnsemble2Preds	0.3727	0.8151	0.4858	0.4264	0.8528	0.5685
NLPatVCU	C6	ensembleWLabEnsemble3Preds	0.3654	0.8151	0.4793	0.4246	0.8528	0.5669
NLPatVCU	C9	ensembleWLabModel4Preds	0.3758	0.8073	0.4876	0.4324	0.8442	0.5718
NLPatVCU	HG0	HGensemble2	0.3050	0.6032	0.3792	0.3679	0.6450	0.4686
NLPatVCU	HG1	HGensemble1	0.2816	0.5922	0.3553	0.3568	0.6580	0.4627
NLPatVCU	HG2	HGmodel4	0.2969	0.5701	0.3658	0.3661	0.6450	0.4671
NLPatVCU	HG3	HGmodel6	0.2904	0.7180	0.3884	0.3427	0.7359	0.4677
NLPatVCU	HG4	HGensemble3	0.2690	0.7113	0.3652	0.3400	0.7359	0.4651
ONTUG	intersection	ElectraCLEANR	0.1721	0.0558	0.0799	**0.8857**	0.1342	0.2331
ONTUG	union	ElectraCLEANR	0.4185	0.4073	0.4057	0.7121	0.6104	0.6573
Schemalink	1	gpt4re	0.3758	0.6573	0.4421	0.4531	0.7532	0.5659
ToGS	hermes3b	CLEANR	0.0352	0.0143	0.0132	0.1176	0.0173	0.0302
ToGS	hermes3blora	CLEANR	0.1440	0.0521	0.0742	0.7368	0.1212	0.2082
ToGS	hermes3bloraragreorder	CLEANR	0.1601	0.0882	0.1058	0.7255	0.1602	0.2624
ToGS	hermes3blorareorder	CLEANR	0.1440	0.0521	0.0742	0.7368	0.1212	0.2082
ToGS	hermes3bragreorder	CLEANR	0.1847	0.0896	0.1098	0.5652	0.1688	0.2600

(Table 12 continued)

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ToGS	hermes3breorder	CLEANR	0.0352	0.0143	0.0132	0.1176	0.0173	0.0302
ToGS	hermes8b	CLEANR	0.1031	0.0397	0.0519	0.5000	0.0866	0.1476
ToGS	hermes8bragreorder	CLEANR	0.2211	0.1304	0.1451	0.5701	0.2641	0.3609
ToGS	hermes8breorder	CLEANR	0.1031	0.0397	0.0519	0.5000	0.0866	0.1476
ToGS	openai4omini	CLEANR	0.1027	0.0958	0.0841	0.2275	0.1861	0.2048
ToGS	openai4ominirag	CLEANR	0.1723	0.1402	0.1457	0.3693	0.2814	0.3194
ToGS	openai4ominiragreorder	CLEANR	0.1960	0.1420	0.1475	0.3943	0.2987	0.3399
ToGS	openai4ominireorder	CLEANR	0.0930	0.0816	0.0685	0.2353	0.1732	0.1995
ataupd2425-gainer	ba1	trainplatinumgolddev	0.1998	0.3983	0.2500	0.4119	0.5671	0.4772
ataupd2425-gainer	ba2	trainplatinumgoldsilver	0.1932	0.4029	0.2458	0.4036	0.5801	0.4760
ataupd2425-gainer	ba	trainplatinumandgold	0.2381	0.3680	0.2699	0.4809	0.5455	0.5112
ataupd2425-gainer	bd1	trainplatinumgolddev	0.1751	0.5400	0.2497	0.3232	0.6926	0.4408
ataupd2425-gainer	bd2	trainplatinumgoldsilver	0.1698	0.5460	0.2449	0.3126	0.6970	0.4316
ataupd2425-gainer	bd	trainplatinumandgold	0.2159	0.5401	0.2902	0.3852	0.6970	0.4961
ataupd2425-gainer	bp1	trainplatinumgolddev	0.2811	0.3652	0.2960	0.5000	0.5195	0.5096
ataupd2425-gainer	bp2	trainplatinumgoldsilver	0.2519	0.3787	0.2854	0.4773	0.5455	0.5091
ataupd2425-gainer	bp	trainplatinumandgold	0.3171	0.3254	0.2968	0.6150	0.4978	0.5502
ataupd2425-gainer	bs1	trainplatinumgolddev	0.1730	0.5430	0.2469	0.3126	0.6970	0.4316
ataupd2425-gainer	bs2	trainplatinumgoldsilver	0.1663	0.5482	0.2407	0.3011	0.7013	0.4213
ataupd2425-gainer	bs	trainplatinumandgold	0.2135	0.5386	0.2875	0.3791	0.6926	0.4900
ataupd2425-pam	A0	RE-BiomedNLP-1NoRel-1epoch	0.4908	0.7080	0.5272	0.5082	0.8052	0.6231
ataupd2425-pam	A1	RE-BiomedNLP-1NoRel-1epoch	0.4580	0.7118	0.5176	0.5068	0.8095	0.6233
ataupd2425-pam	A2	RE-BiomedNLP-1NoRel-1epoch	0.4413	0.7187	0.5165	0.4987	0.8095	0.6172
ataupd2425-pam	A3	RE-BiomedNLP-2NoRel-1epoch	0.4528	0.6209	0.4805	0.5449	0.7359	0.6262
ataupd2425-pam	A4	RE-BiomedNLP-2NoRel-1epoch	0.4338	0.6065	0.4740	0.5463	0.7403	0.6287
ataupd2425-pam	A5	RE-BiomedNLP-2NoRel-1epoch	0.4386	0.6078	0.4741	0.5414	0.7359	0.6239
ataupd2425-pam	A6	RE-BiomedNLP-3NoRel-1epoch	0.5020	0.5929	0.5003	0.5619	0.7273	0.6340
ataupd2425-pam	A7	RE-BiomedNLP-3NoRel-1epoch	0.4807	0.6091	0.4993	0.5671	0.7316	0.6389
ataupd2425-pam	A8	RE-BiomedNLP-3NoRel-1epoch	0.4682	0.6066	0.4952	0.5567	0.7229	0.6290

C. Subtask 6.2.2 (TT-RE) Overall Results

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

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BASELINE	Organizers	Atlop-Finetuned	0.4729	0.3421	0.3745	0.7533	0.4650	0.5751
BIU-ONLP	1	SapBERT	0.3734	0.3454	0.3430	0.6497	0.4733	0.5476
BIU-ONLP	2	BertBaseCased	0.4467	0.2799	0.3182	**0.7803**	0.4239	0.5493
BIU-ONLP	3	BioBert1.1PubMed	0.4134	0.2866	0.3187	0.7519	0.3992	0.5215
BIU-ONLP	4	RobertaLarge	0.4725	0.3288	0.3630	0.7362	0.4938	0.5911
Graphwise-1	1	ONTO-REBEL	0.3735	0.2949	0.3052	0.6341	0.4280	0.5111
Graphwise-1	2	ONTO-REBEL	0.4333	0.3711	0.3693	0.6888	0.5556	0.6150
Graphwise-1	100	AtlopOnto	0.1478	0.0835	0.1044	0.5000	0.0741	0.1290
Graphwise-1	101	AtlopOnto	0.1478	0.0835	0.1044	0.5000	0.0741	0.1290
Graphwise-1	102	AtlopOnto	0.0543	0.0217	0.0304	0.4286	0.0123	0.0240
Graphwise-1	103	AtlopOnto	0.2188	0.1439	0.1667	0.5918	0.1193	0.1986
Graphwise-1	104	AtlopOnto	0.4142	0.3474	0.3686	0.7198	0.5391	0.6165
Graphwise-1	105	AtlopOnto	0.4119	0.3709	0.3840	0.7326	0.5638	0.6372
Graphwise-1	106	AtlopOnto	0.3455	0.2528	0.2807	0.7170	0.4691	0.5672
Graphwise-1	107	AtlopOnto	0.4466	0.3932	0.4077	0.6837	0.5514	0.6105
Graphwise-1	108	AtlopOnto	0.4352	0.3461	0.3667	0.7049	0.5309	0.6056
Graphwise-1	109	AtlopOnto	0.3717	0.2856	0.3070	0.7134	0.4815	0.5749
Graphwise-1	110	AtlopOnto	0.4151	0.3148	0.3436	0.7118	0.4979	0.5860
Graphwise-1	12	ONTO-Bio-GPT	0.3191	0.3620	0.3087	0.4939	0.5021	0.4980
Graphwise-1	13	ONTO-Bio-GPT	0.3387	0.3874	0.3304	0.4803	0.5021	0.4909
Gut-Instincts	6229ee3re		0.4826	0.6357	**0.5216**	0.6329	0.7449	0.6843
Gut-Instincts	6229ee3redev		**0.4869**	0.6149	0.5138	0.6177	0.7449	0.6754
Gut-Instincts	6229eedev3re		0.4663	0.6445	0.5184	0.6280	0.7572	**0.6866**
Gut-Instincts	6229eedev3redev		0.4736	0.6147	0.5094	0.6087	0.7490	0.6716
ICUE	run10	biolinkbertl	0.3571	0.6539	0.4435	0.4650	0.7654	0.5785
ICUE	run11	biolinkbertl	0.3525	0.6199	0.4320	0.4786	0.7366	0.5802
ICUE	run12	biolinkbertl	0.3480	0.6344	0.4326	0.4775	0.7407	0.5806
ICUE	run13	biolinkbertl	0.3449	0.6339	0.4293	0.4682	0.7572	0.5786
ICUE	run14	biolinkbertl_pp	0.3516	0.7186	0.4531	0.4402	0.8025	0.5685
ICUE	run15	biolinkbertl_pp	0.4034	0.6324	0.4639	0.5241	0.7160	0.6052
ICUE	run16	biolinkbertl_pp	0.4009	0.7172	0.4869	0.4848	0.7901	0.6009
ICUE	run17	biolinkbertl_pp	0.3990	0.6132	0.4562	0.5262	0.7037	0.6021
ICUE	run18	biolinkbertl_pp	0.3819	0.6618	0.4635	0.5028	0.7490	0.6017
ICUE	run19	biolinkbertl_pp	0.3539	0.6591	0.4406	0.4604	0.7654	0.5750
ICUE	run1	biolinkbertl	0.3407	0.6358	0.4221	0.4439	0.7490	0.5574
ICUE	run20	biolinkbertl_pp	0.3606	0.6611	0.4499	0.4794	0.7654	0.5895
ICUE	run21	biolinkbertl_pp	0.3709	0.6620	0.4569	0.4779	0.7572	0.5860
ICUE	run22	biolinkbertl_pp	0.4011	0.7123	0.4879	0.4974	0.7860	0.6093
ICUE	run23	biolinkbertl_pp	0.3945	0.6409	0.4589	0.5131	0.7243	0.6007
ICUE	run24	biolinkbertl_pp	0.3627	0.7227	0.4632	0.4722	0.8025	0.5945
ICUE	run2	biolinkbertl	0.3486	0.7263	0.4524	0.4569	0.8066	0.5833
ICUE	run3	biolinkbertl	0.3516	0.7186	0.4531	0.4402	0.8025	0.5685
ICUE	run4	biolinkbertl	0.3739	0.5769	0.4299	0.4985	0.6872	0.5779
ICUE	run5	biolinkbertl	0.3519	0.7150	0.4463	0.4273	0.8107	0.5597
ICUE	run6	biolinkbertl	0.3571	0.6539	0.4435	0.4650	0.7654	0.5785
ICUE	run7	biolinkbertl	0.3449	0.6339	0.4293	0.4682	0.7572	0.5786
ICUE	run8	biolinkbertl	0.3577	0.5440	0.4114	0.5156	0.6790	0.5861
ICUE	run9	biolinkbertl	0.3780	0.6371	0.4570	0.4958	0.7325	0.5914
LYX-DMIIP-FDU	run1	BioLinkBERT	0.3625	0.4171	0.3549	0.5973	0.5432	0.5690
NLPatVCU	C10	ensembleWLabModel4Preds	0.3552	0.7627	0.4593	0.4325	0.8313	0.5690
NLPatVCU	C13	mixedCNNWLabEnsemble1Preds	0.3840	**0.8202**	0.4905	0.4160	**0.8765**	0.5642
NLPatVCU	C16	mixedCNNWLabEnsemble2Preds	0.3772	0.8077	0.4837	0.4286	0.8519	0.5702
NLPatVCU	C19	mixedCNNWLabModel4Preds	0.3810	0.8005	0.4868	0.4362	0.8436	0.5750
NLPatVCU	C1	ensembleWLabEnsemble1Preds	0.3804	0.8078	0.4868	0.4280	0.8683	0.5734
NLPatVCU	C22	platPlusCNNWLabEnsemble1Preds	0.3088	0.6403	0.3922	0.4205	0.7942	0.5499
NLPatVCU	C25	platPlusCNNWLabEnsemble2Preds	0.3099	0.6233	0.3898	0.4257	0.7778	0.5502
NLPatVCU	C28	platPlusCNNWLabModel4Preds	0.3145	0.6161	0.3933	0.4349	0.7695	0.5557
NLPatVCU	C4	ensembleWLabEnsemble2Preds	0.3487	0.7699	0.4540	0.4259	0.8395	0.5651
NLPatVCU	C7	ensembleWLabEnsemble3Preds	0.3455	0.7699	0.4517	0.4250	0.8395	0.5643
ONTUG	union	ElectraCLEANR	0.4254	0.4025	0.4058	0.7059	0.5926	0.6443
Schemalink	1	gpt4re	0.3756	0.6592	0.4437	0.4523	0.7613	0.5675
ToGS	hermes3b	CLEANR	0.0329	0.0134	0.0123	0.1212	0.0165	0.0290
ToGS	hermes3blora	CLEANR	0.1315	0.0422	0.0621	0.6765	0.0947	0.1661
ToGS	hermes3bloraragreorder	CLEANR	0.1472	0.0828	0.0989	0.7059	0.1481	0.2449
ToGS	hermes3blorareorder	CLEANR	0.1315	0.0422	0.0621	0.6765	0.0947	0.1661
ToGS	hermes3bragreorder	CLEANR	0.1712	0.0834	0.1020	0.5507	0.1564	0.2436
ToGS	hermes3breorder	CLEANR	0.0329	0.0134	0.0123	0.1212	0.0165	0.0290
ToGS	hermes8b	CLEANR	0.0887	0.0391	0.0482	0.3590	0.0576	0.0993
ToGS	hermes8bragreorder	CLEANR	0.2261	0.1267	0.1414	0.5556	0.2469	0.3419
ToGS	hermes8breorder	CLEANR	0.0887	0.0391	0.0482	0.3590	0.0576	0.0993
ToGS	openai4omini	CLEANR	0.0988	0.0885	0.0775	0.2010	0.1605	0.1785
ToGS	openai4ominirag	CLEANR	0.1866	0.1365	0.1447	0.3652	0.2675	0.3088

(Table 13 continued)

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ToGS	openai4ominiragreorder	CLEANR	0.2066	0.1476	0.1531	0.3898	0.2840	0.3286
ToGS	openai4ominireorder	CLEANR	0.0895	0.0739	0.0639	0.2057	0.1481	0.1722
ataupd2425-gainer	ta1	trainplatinumgolddev	0.2095	0.1784	0.1685	0.5814	0.3086	0.4032
ataupd2425-gainer	ta2	trainplatinumgoldsilver	0.2097	0.1815	0.1680	0.5580	0.3169	0.4042
ataupd2425-gainer	ta	trainplatinumandgold	0.2543	0.1663	0.1818	0.7228	0.3004	0.4244
ataupd2425-gainer	td1	trainplatinumgolddev	0.2862	0.2690	0.2596	0.6074	0.4074	0.4877
ataupd2425-gainer	td2	trainplatinumgoldsilver	0.2718	0.2766	0.2525	0.5754	0.4239	0.4882
ataupd2425-gainer	td	trainplatinumandgold	0.3167	0.2315	0.2528	0.7405	0.3992	0.5187
ataupd2425-gainer	ts1	trainplatinumgolddev	0.2719	0.2829	0.2584	0.5424	0.3951	0.4571
ataupd2425-gainer	ts2	trainplatinumgoldsilver	0.2497	0.2863	0.2448	0.5000	0.3951	0.4414
ataupd2425-gainer	ts	trainplatinumandgold	0.3230	0.2578	0.2728	0.6812	0.3868	0.4934
ataupd2425-pam	B0	RE-BiomedNLP-1NoRel-1epoch	0.4419	0.6742	0.4980	0.5219	0.7860	0.6273
ataupd2425-pam	B1	RE-BiomedNLP-1NoRel-1epoch	0.4411	0.6805	0.5003	0.5257	0.7984	0.6340
ataupd2425-pam	B2	RE-BiomedNLP-1NoRel-1epoch	0.4290	0.6870	0.5017	0.5229	0.7984	0.6319
ataupd2425-pam	B3	RE-BiomedNLP-2NoRel-1epoch	0.4427	0.6097	0.4776	0.5570	0.7243	0.6297
ataupd2425-pam	B4	RE-BiomedNLP-2NoRel-1epoch	0.4398	0.6052	0.4812	0.5701	0.7366	0.6427
ataupd2425-pam	B5	RE-BiomedNLP-2NoRel-1epoch	0.4325	0.6064	0.4787	0.5651	0.7325	0.6380
ataupd2425-pam	B6	RE-BiomedNLP-3NoRel-1epoch	0.4554	0.5680	0.4729	0.5767	0.7119	0.6372
ataupd2425-pam	B7	RE-BiomedNLP-3NoRel-1epoch	0.4409	0.5704	0.4694	0.5853	0.7202	0.6458
ataupd2425-pam	B8	RE-BiomedNLP-3NoRel-1epoch	0.4278	0.5679	0.4638	0.5710	0.7119	0.6337
lasigeBioTM	R1	BENTMistral	0.1204	0.0655	0.0768	0.4571	0.0658	0.1151
lasigeBioTM	R1	BENTMistralSemantic	0.0109	0.0097	0.0102	0.3077	0.0165	0.0312
lasigeBioTM	R1	Baseline	0.1116	0.0479	0.0616	0.4091	0.0370	0.0679
lasigeBioTM	R1	ConstParsing	0.0797	0.0622	0.0646	0.3929	0.0453	0.0812

D. Subtask 6.2.3 (TM-RE) Overall Results

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

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BASELINE	Organizers	Atlop-Finetuned	**0.3514**	0.1829	0.2123	**0.4986**	0.2453	0.3288
BIU-ONLP	1	SapBERT	0.0777	0.0807	0.0765	0.2033	0.1327	0.1606
BIU-ONLP	2	BertBaseCased	0.1274	0.0777	0.0899	0.2929	0.1166	0.1668
BIU-ONLP	3	BioBert1.1PubMed	0.0935	0.0682	0.0683	0.2459	0.1206	0.1619
BIU-ONLP	4	RobertaLarge	0.1171	0.0854	0.0879	0.2339	0.1461	0.1799
Graphwise-1	1	ONTO-REBEL	0.2074	0.1780	0.1712	0.3832	0.2507	0.3031
Graphwise-1	2	ONTO-REBEL	0.2792	0.2356	0.2262	0.4336	0.3110	0.3622
Graphwise-1	100	AtlopOnto	0.0681	0.0256	0.0341	0.1831	0.0174	0.0318
Graphwise-1	101	AtlopOnto	0.0681	0.0256	0.0341	0.1831	0.0174	0.0318
Graphwise-1	102	AtlopOnto	0.0217	0.0036	0.0062	0.0833	0.0013	0.0026
Graphwise-1	103	AtlopOnto	0.1440	0.0740	0.0920	0.3373	0.0375	0.0676
Graphwise-1	104	AtlopOnto	0.2641	0.1921	0.2031	0.4415	0.2681	0.3336
Graphwise-1	105	AtlopOnto	0.2816	0.2205	0.2359	0.4612	0.2949	0.3598
Graphwise-1	106	AtlopOnto	0.2173	0.1538	0.1608	0.3954	0.2534	0.3088
Graphwise-1	107	AtlopOnto	0.3323	0.2369	0.2603	0.4686	0.3097	0.3729
Graphwise-1	108	AtlopOnto	0.3106	0.1954	0.2167	0.4277	0.2775	0.3366
Graphwise-1	109	AtlopOnto	0.2602	0.1766	0.1860	0.4552	0.2654	0.3353
Graphwise-1	110	AtlopOnto	0.2795	0.1867	0.2090	0.4511	0.2413	0.3144
Graphwise-1	12	ONTO-Bio-GPT	0.1062	0.1113	0.0951	0.1784	0.1595	0.1684
Graphwise-1	13	ONTO-Bio-GPT	0.0816	0.1090	0.0775	0.1206	0.1595	0.1373
Gut-Instincts	6239ee3re		0.3255	0.4237	0.3409	0.4137	0.5013	0.4533
Gut-Instincts	6239ee3redev		0.3315	0.3599	0.3100	0.3893	0.4477	0.4165
Gut-Instincts	6239eedev3re		0.3310	0.4303	**0.3497**	0.4215	0.5147	**0.4635**
Gut-Instincts	6239eedev3redev		0.3277	0.3658	0.3145	0.3933	0.4598	0.4240
ICUE	run10	biolinkbertl	0.2017	0.4041	0.2452	0.2203	0.4853	0.3031
ICUE	run11	biolinkbertl	0.2249	0.4144	0.2685	0.2649	0.4946	0.3450
ICUE	run12	biolinkbertl	0.2182	0.3893	0.2549	0.2530	0.4544	0.3250
ICUE	run13	biolinkbertl	0.2180	0.4189	0.2634	0.2466	0.4933	0.3289
ICUE	run14	biolinkbertl_pp	0.1604	0.4538	0.2222	0.1970	0.5509	0.2903
ICUE	run15	biolinkbertl_pp	0.2581	0.4061	0.2821	0.2886	0.4759	0.3593
ICUE	run16	biolinkbertl_pp	0.2103	0.4436	0.2577	0.2340	0.5147	0.3217
ICUE	run17	biolinkbertl_pp	0.2681	0.3948	0.2878	0.2967	0.4450	0.3560
ICUE	run18	biolinkbertl_pp	0.2296	0.4230	0.2711	0.2642	0.4665	0.3374
ICUE	run19	biolinkbertl_pp	0.1951	0.4367	0.2526	0.2378	0.5335	0.3289
ICUE	run1	biolinkbertl	0.1885	0.4179	0.2402	0.2162	0.5214	0.3057
ICUE	run20	biolinkbertl_pp	0.1878	0.4006	0.2405	0.2285	0.5067	0.3150
ICUE	run21	biolinkbertl_pp	0.1892	0.4167	0.2446	0.2344	0.5094	0.3211
ICUE	run22	biolinkbertl_pp	0.2165	0.4563	0.2709	0.2473	0.5523	0.3416
ICUE	run23	biolinkbertl_pp	0.2509	0.4239	0.2825	0.2858	0.5054	0.3651
ICUE	run24	biolinkbertl_pp	0.1862	0.4846	0.2517	0.2316	0.5697	0.3293
ICUE	run2	biolinkbertl	0.1756	0.4992	0.2429	0.2153	0.5925	0.3158
ICUE	run3	biolinkbertl	0.1604	0.4538	0.2222	0.1970	0.5509	0.2903
ICUE	run4	biolinkbertl	0.2403	0.3677	0.2613	0.2661	0.4437	0.3327
ICUE	run5	biolinkbertl	0.1525	0.4666	0.2147	0.1881	0.5684	0.2827
ICUE	run6	biolinkbertl	0.2017	0.4041	0.2452	0.2203	0.4853	0.3031
ICUE	run7	biolinkbertl	0.2180	0.4189	0.2634	0.2466	0.4933	0.3289
ICUE	run8	biolinkbertl	0.2251	0.3551	0.2482	0.2705	0.4330	0.3330
ICUE	run9	biolinkbertl	0.2165	0.3790	0.2543	0.2545	0.4517	0.3256
LYX-DMIIP-FDU	run1	BioLinkBERT	0.2106	0.2418	0.1990	0.3682	0.3257	0.3457
NLPatVCU	C11	ensembleWLabModel4Preds	0.1522	0.5041	0.2163	0.1423	0.6005	0.2300
NLPatVCU	C14	mixedCNNWLabEnsemble1Preds	0.1278	0.5067	0.1864	0.0999	0.6300	0.1724
NLPatVCU	C17	mixedCNNWLabEnsemble2Preds	0.1291	0.4815	0.1885	0.1104	0.5858	0.1858
NLPatVCU	C20	mixedCNNWLabModel4Preds	0.1375	0.4812	0.1966	0.1114	0.5831	0.1870
NLPatVCU	C23	platPlusCNNWLabEnsemble1Preds	0.1312	0.4804	0.1853	0.1128	0.6180	0.1907
NLPatVCU	C26	platPlusCNNWLabEnsemble2Preds	0.1236	0.4500	0.1819	0.1199	0.5643	0.1978
NLPatVCU	C29	platPlusCNNWLabModel4Preds	0.1287	0.4532	0.1884	0.1222	0.5643	0.2009
NLPatVCU	C2	ensembleWLabEnsemble1Preds	0.1465	**0.5304**	0.2088	0.1293	**0.6488**	0.2156
NLPatVCU	C5	ensembleWLabEnsemble2Preds	0.1474	0.5022	0.2104	0.1405	0.6019	0.2279
NLPatVCU	C8	ensembleWLabEnsemble3Preds	0.1525	0.5025	0.2145	0.1403	0.6059	0.2278
ONTUG	union	ElectraCLEANR	0.2589	0.2293	0.2266	0.3529	0.3231	0.3373
Schemalink	1	gpt4re	0.2265	0.4088	0.2546	0.1948	0.4665	0.2749
ToGS	hermes3b	CLEANR	0.0001	0.0027	0.0003	0.0022	0.0013	0.0017
ToGS	hermes3blora	CLEANR	0.0262	0.0159	0.0193	0.1850	0.0429	0.0696
ToGS	hermes3bloraragreorder	CLEANR	0.0249	0.0180	0.0203	0.1702	0.0536	0.0815
ToGS	hermes3blorareorder	CLEANR	0.0262	0.0159	0.0193	0.1850	0.0429	0.0696
ToGS	hermes3bragreorder	CLEANR	0.0450	0.0338	0.0185	0.0397	0.0241	0.0300
ToGS	hermes3breorder	CLEANR	0.0001	0.0027	0.0003	0.0022	0.0013	0.0017
ToGS	hermes8b	CLEANR	0.0045	0.0108	0.0054	0.0163	0.0134	0.0147
ToGS	hermes8bragreorder	CLEANR	0.0277	0.0250	0.0248	0.0580	0.0375	0.0456
ToGS	hermes8breorder	CLEANR	0.0045	0.0108	0.0054	0.0163	0.0134	0.0147
ToGS	openai4omini	CLEANR	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ToGS	openai4ominirag	CLEANR	0.0041	0.0015	0.0017	0.0066	0.0040	0.0050

(Table 14 continued)

			Macro-averaging			Micro-averaging		
team_id	**run_id**	**system_desc**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
ToGS	openai4ominiragreorder	CLEANR	0.0033	0.0047	0.0038	0.0046	0.0027	0.0034
ToGS	openai4ominireorder	CLEANR	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ataupd2425-gainer	tma1	trainplatinumgolddev	0.1066	0.0640	0.0706	0.2637	0.1099	0.1552
ataupd2425-gainer	tma2	trainplatinumgoldsilver	0.0955	0.0630	0.0651	0.2382	0.1086	0.1492
ataupd2425-gainer	tma	trainplatinumandgold	0.1659	0.0719	0.0906	0.4226	0.1354	0.2051
ataupd2425-gainer	tmd1	trainplatinumgolddev	0.1462	0.1238	0.1275	0.2944	0.1555	0.2035
ataupd2425-gainer	tmd2	trainplatinumgoldsilver	0.1494	0.1229	0.1179	0.2629	0.1568	0.1965
ataupd2425-gainer	tmd	trainplatinumandgold	0.2127	0.1254	0.1419	0.4050	0.1743	0.2437
ataupd2425-gainer	tms1	trainplatinumgolddev	0.1600	0.1349	0.1327	0.2850	0.1635	0.2078
ataupd2425-gainer	tms2	trainplatinumgoldsilver	0.1336	0.1344	0.1197	0.2612	0.1635	0.2012
ataupd2425-gainer	tms	trainplatinumandgold	0.2203	0.1384	0.1538	0.4272	0.1810	0.2542
ataupd2425-pam	C0	RE-BiomedNLP-1NoRel-1epoch	0.1546	0.3223	0.1857	0.1786	0.3887	0.2447
ataupd2425-pam	C1	RE-BiomedNLP-1NoRel-1epoch	0.1538	0.3124	0.1802	0.1766	0.3941	0.2439
ataupd2425-pam	C2	RE-BiomedNLP-1NoRel-1epoch	0.1454	0.3050	0.1746	0.1734	0.3874	0.2395
ataupd2425-pam	C3	RE-BiomedNLP-2NoRel-1epoch	0.1796	0.3142	0.2020	0.2080	0.3472	0.2602
ataupd2425-pam	C4	RE-BiomedNLP-2NoRel-1epoch	0.1798	0.3063	0.1993	0.2064	0.3539	0.2607
ataupd2425-pam	C5	RE-BiomedNLP-2NoRel-1epoch	0.1755	0.3021	0.1955	0.2014	0.3485	0.2553
ataupd2425-pam	C6	RE-BiomedNLP-3NoRel-1epoch	0.2012	0.2872	0.2069	0.2270	0.3378	0.2716
ataupd2425-pam	C7	RE-BiomedNLP-3NoRel-1epoch	0.1940	0.2764	0.1982	0.2278	0.3432	0.2738
ataupd2425-pam	C8	RE-BiomedNLP-3NoRel-1epoch	0.1873	0.2718	0.1936	0.2179	0.3365	0.2645
lasigeBioTM	R1	BENTMistral	0.0268	0.0048	0.0078	0.0930	0.0054	0.0101
lasigeBioTM	R1	Baseline	0.0217	0.0008	0.0016	0.0667	0.0013	0.0026
lasigeBioTM	R1	ConstParsing	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
lasigeBioTM	R1	BENTMistralSemantic	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000