SMTE – GutBrainIE@CLEF 2026: Biomedical Named Entity Recognition and Mention-Level Relation Extraction for the Gut-Brain Axis

Sofia Maule[1,*], Giorgio Maria Di Nunzio[1]

[1]Department of Information Engineering, University of Padova, Italy

Abstract

We present the participation of team SMTE in the GutBrainIE shared task at CLEF 2026, which targets structured information extraction from PubMed abstracts on the gut-brain axis. We address two subtasks: Named Entity Recognition (NER, Subtask 6.1.1) and Mention-Level Relation Extraction (M-RE, Subtask 6.2.1).

For NER, we fine-tune BioBERT and PubMedBERT on heterogeneous training data (Gold, Silver, and Bronze annotations) and combine them in a hybrid ensemble. The ensemble achieves Micro-F1 0.8324 on the development set and 0.8019 on the official test set.

For M-RE, we adopt a pairwise classification approach with entity markers, mention-mean pooling, and hard negative sampling over automatically predicted entity mentions. Our best run, based on PubMedBERT-large with a local context window, achieves Micro-F1 0.4223 on the official test set.

Keywords

Biomedical NLP, Named Entity Recognition, Relation Extraction, Gut-Brain Axis, Transformer Models

1. Introduction

The gut-brain axis, the bidirectional communication network between the gastrointestinal tract and the central nervous system, has recently emerged as a major research focus in biomedicine. A growing body of studies highlights the role of gut microbiota dysbiosis in neurological and psychiatric conditions, including Alzheimer's disease, Parkinson's disease, Multiple Sclerosis, Amyotrophic Lateral Sclerosis, and various mental health disorders.

However, the rapid increase in biomedical publications makes manual analysis and curation difficult to scale, creating the need for automatic information extraction (IE) systems capable of structuring this knowledge. Such systems must identify relevant biomedical entity mentions and, in more advanced settings, extract semantic relations among them.

GutBrainIE 2026 addresses these challenges through a shared task on structured information extraction from PubMed titles and abstracts related to the gut-brain axis, organized within the BioASQ Lab at CLEF 2026 [1, 2]. The task builds on the first edition of GutBrainIE at CLEF 2025 [3], where gut-brain interplay information extraction was introduced as one of the new BioASQ tasks [4, 5]. In its 2026 edition, the task is organized into four subtasks: named entity recognition (NER), named entity recognition and disambiguation (NERD), and relation extraction (RE), evaluated at two levels, namely mention-level relation extraction (M-RE) and concept-level relation extraction (C-RE). NER focuses on identifying entity spans and assigning entity labels. NERD extends this requirement by assigning a normalized URI to each entity mention. M-RE identifies relations between textual entity mentions, whereas C-RE identifies relations between normalized biomedical concepts represented by URIs.

In this paper, we present the SMTE system for GutBrainIE 2026. We participate in two subtasks: Subtask 6.1.1, named entity recognition (NER), and Subtask 6.2.1, mention-level relation extraction (M-RE). The first requires the identification and classification of entity mentions into 13 predefined biomedical categories, while the second requires the identification of semantic relations between entity mentions, expressed as subject–predicate–object triples.

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

sofia.maule@studenti.unipd.it (S. Maule); giorgiomaria.dinunzio@unipd.it (G. M. D. Nunzio)

Both subtasks present significant challenges. In NER, several categories are semantically close (e.g., `chemical`, `drug`, and `dietary supplement`), making classification difficult even for domain experts. Additionally, the dataset is highly imbalanced, with frequent categories such as `human` and `microbiome` dominating rare ones like `food` or `gene`. In M-RE, the main challenges include the large number of possible entity pairs, the sparsity of positive relations, and the propagation of errors from the NER stage.

Another key difficulty lies in the heterogeneity of the training data. The dataset combines annotations of different quality levels: Gold, Silver, Silver 2025, and Bronze. These collections include expert annotations, student annotations, partially automatic annotations from the previous edition, and fully automatic annotations. This heterogeneity requires careful strategies to effectively exploit all available data while limiting the impact of annotation noise.

The SMTE system addresses both subtasks using transformer-based models fine-tuned on the GutBrainIE dataset. For NER, we explore multiple training strategies, including priority-based merging, curriculum learning, and annotator-weighted training, and combine the strongest models into a hybrid ensemble. For M-RE, we adopt a pairwise classification approach with entity markers and type-based candidate filtering. Concretely, this work contributes: (i) a comparison of curriculum, annotator-weighted, and priority-merge strategies for exploiting the heterogeneous Gold/Silver/Bronze annotations; (ii) a hybrid BioBERT–PubMedBERT ensemble for NER; (iii) a type-filtered, hard-negative-trained relation extraction pipeline; and (iv) an analysis of threshold-calibration strategies and their generalisation from the development to the test set.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset and task setup. Section 4 presents the NER system. Section 5 describes the relation extraction approach. Section 6 concludes the paper.

2. Related Work

The SMTE system builds on prior work in biomedical information extraction and on recent resources developed for the gut-brain axis domain. The following paragraphs review transformer-based approaches to biomedical named entity recognition, then discuss relation extraction methods based on entity-pair classification and document-level modelling, and finally situate the system within the GutBrainIE benchmark and its previous edition.

Biomedical NER. Named entity recognition in the biomedical domain is most commonly framed as sequence labelling over BIO-tagged tokens on top of pre-trained transformer encoders [6]. Domain-adapted encoders substantially outperform general-domain BERT on biomedical text: BioBERT [7] continues pre-training on PubMed abstracts and PMC full texts, while PubMedBERT [8] shows that pre-training from scratch on in-domain text, together with an in-domain vocabulary, yields further gains. An alternative to fine-tuned taggers is GLiNER [9], a span-based generalist recogniser that the organisers adopt as the NER baseline for this task. Our NER system follows the fine-tuned-tagger paradigm and combines BioBERT and PubMedBERT in an ensemble.

Biomedical Relation Extraction. Relation extraction is typically cast either as pairwise mention classification or as document-level extraction. The entity-marker formulation, in which special tokens are inserted around the candidate arguments before classification, was popularised by Baldini Soares et al. [10] and remains a strong baseline; we adopt it here, pairing entity markers with mention pooling on top of biomedical encoders. Besides PubMedBERT, we also experiment with BioLinkBERT [11], which augments in-domain pre-training with document link structure. At the document level, ATLOP [12] introduces adaptive thresholding and localised context pooling and serves as the organiser baseline for the mention-level RE subtask. A recurring difficulty in biomedical RE is the extreme imbalance between the few positive pairs and the many type-compatible negatives, which motivates the candidate filtering and hard-negative sampling that are central to our approach.

The GutBrainIE task and benchmark. The GutBrainIE shared task was introduced in its first edition at CLEF 2025 [13] as part of the BioASQ Lab, targeting entity and relation extraction for the gut-brain axis over a corpus annotated in different quality tiers. The benchmark is closely related to recent work on curated biomedical information extraction resources for this domain. Martinelli et al. [14] describe the construction of a domain-specific benchmark for entity extraction, concept linking, and document-level relation extraction, based on PubMed abstracts and annotated with fine-grained entities, concept-level links, and relations. This work is particularly relevant to the present paper because it defines the data setting in which GutBrainIE systems are evaluated and highlights the importance of combining highly curated annotations with broader weakly supervised material.

From a terminological perspective, Bonato et al. [15] analyse the conceptual dimension of gut-brain terminology, with particular attention to intensional definitions and associative relations between biomedical concepts. Their study provides complementary motivation for treating relation extraction not only as a text-mining task, but also as a way to recover structured conceptual associations in a specialised biomedical domain.

The 2025 GutBrainIE participants explored a range of approaches: the GutUZH team [16] obtained the best results on the NER subtask with fine-tuned biomedical encoders, SCIRE [17] investigated LLM-driven biomedical NER, Graphwise [18] applied deep learning to both NER and relation extraction, and other teams built end-to-end information-extraction pipelines over the GutBrainIE corpus [19]. The 2026 edition [2] extends the task with additional annotation tiers and the mention- and concept-level RE subtasks addressed here. Our work continues the supervised fine-tuning and ensembling line found effective in 2025, and adds a systematic study of how different annotation tiers, threshold calibration, context-window selection, and NER-to-RE error propagation affect generalisation to the hidden test set.

3. Task and Data Description

The GutBrainIE shared task [2] focuses on extracting structured information from PubMed abstracts describing the interaction between gut microbiota and neurological or mental health conditions.

3.1. Subtask 6.1.1 – Named Entity Recognition

In this subtask, systems identify and classify entity mentions into 13 predefined biomedical categories: `anatomical location`, `animal`, `bacteria`, `biomedical technique`, `chemical`, `DDF` (disease, disorder, or finding), `dietary supplement`, `drug`, `food`, `gene`, `human`, `microbiome`, and `statistical technique`.

Each entity mention is represented as a tuple:

$$(\textit{entityCategory}\,;\ \textit{entityLocation}\,;\ \textit{startOffset}\,;\ \textit{endOffset})$$

where the location specifies whether the entity appears in the title or the abstract.

3.2. Subtask 6.2.1 – Mention-Level Relation Extraction

In this subtask, systems identify semantic relations between entity mentions within the same document. Each relation is expressed as a triple:

$$(\textit{subjectMention}\,;\ \textit{relationPredicate}\,;\ \textit{objectMention})$$

The relation schema defines directed predicates between specific pairs of entity types. Examples include:

- `bacteria` → `DDF`: `influence`
- `chemical` → `microbiome`: `impact`, `produced by`
- `microbiome` → `DDF`: `is linked to`
- `DDF` → `bacteria`: `change abundance`

3.3. Dataset

The training data is composed of four collections with different levels of annotation quality:

- Gold: high-quality annotations created by expert annotators;
- Silver: annotations produced by trained students, divided into two groups based on annotation consistency;
- Silver 2025: annotations from a previous edition, partially enriched with automatically generated concept-level information;
- Bronze: distantly supervised annotations generated automatically without manual validation.

The test set consists of a held-out subset of the Gold data, ensuring high-quality evaluation.

4. Named Entity Recognition

We frame NER as a token classification task over BIO-tagged sequences. All 13 entity categories are expanded into a BIO label set of 27 tags (one `B-`/`I-` pair per class, plus `O`). Each document is split into two independent segments — title and abstract — so that character offsets remain consistent with the annotated `location` field.

4.1. System Overview

We submitted three runs for Subtask 6.1.1, corresponding to three distinct model configurations:

- R1 (NERensemble): a hybrid ensemble of NB1b-BioBERT and NB1b-PubMedBERT — our primary submission;
- R2 (NERpubmedbert): single-model NB1b-PubMedBERT;
- R3 (NERbiobert): single-model NB1b-BioBERT.

All three configurations are trained on the same Gold + Silver + Bronze data with a priority-based merge and share the same inference pipeline. R1 is the strongest on the development set (Macro-F1 0.802, Micro-F1 0.8324), and was selected as the primary submission. R2 and R3 are included to isolate the contribution of each backbone independently.

The design of the ensemble was informed by a systematic comparison of seven training configurations (NB1–NB4b), all sharing a common inference pipeline. The following subsections describe training strategies, the unified inference procedure, and ensemble construction in detail.

4.2. Training Strategy

The training strategy is designed to address two main sources of variation: the choice of biomedical encoder and the heterogeneous quality of the available annotations. We therefore compare several configurations that differ in the amount and ordering of training data, while keeping the core model architecture and most hyperparameters fixed. This allows us to isolate the effect of adding noisier annotation tiers, using curriculum-style training, and applying annotator-based weighting.

Backbone models and shared hyperparameters. We use transformer-based biomedical encoders for all NER experiments. Our main backbone is BioBERT [7] (`dmis-lab/biobert-v1.1`), and we also evaluate a variant based on PubMedBERT [8] (`microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract`). In all settings, the encoder is fine-tuned end-to-end with a token-level classification head.

Unless otherwise stated, all configurations use the same training setup:

- learning rate 3×10^{-5} with linear scheduling and 10% warmup;
- weight decay 0.01;

- batch size 8 with gradient accumulation over 2 steps;
- best checkpoint selected on the development set;
- class-weighted cross-entropy loss, where class weights are computed from inverse label frequency$^{0.5}$ and clipped to the range $[0.5, 5.0]$.

Experimental settings. All model variants were first developed and compared on the development set. We explored several strategies for combining training data of different quality.

- NB1 – Single-stage training on Gold + Silver (BioBERT). Gold and Silver data are merged using a priority-based policy, keeping the higher-quality annotation when the same PubMed identifier (PMID) appears in more than one source. The model is trained in a single stage for 5 epochs.
- NB1b – Single-stage training on Gold + Silver + Bronze (BioBERT / PubMedBERT). This configuration extends NB1 by also including Bronze data after the same priority merge. We test this setup with two different backbones: BioBERT and PubMedBERT. This is the strongest single-stage setting.
- NB2 – Two-stage curriculum training (BioBERT). Training is divided into two sequential stages: the first stage uses Gold data only, while the second continues from the first checkpoint on a mixture of oversampled Gold and shuffled Silver data.
- NB3 – Three-stage cross-year curriculum (BioBERT). This setting adds the 2025 collection as an initial warm-up stage. Training starts from 2025 Platinum + Gold data, then continues on 2026 Gold, and finally on a 2026 Gold + Silver mixture.
- NB4 / NB4b – Annotator-weighted training (BioBERT). Instead of merging data with a hard priority rule, these variants assign different sample weights according to the annotation source. NB4 uses Gold + Silver + Silver 2025, while NB4b also includes Bronze.

Selected models for the final system. Among the single-model configurations, NB1b gives the strongest overall results, especially when using the PubMedBERT backbone. The final submitted NER run is obtained through a post-hoc hybrid ensemble combining NB1b-BioBERT and NB1b-PubMedBERT, described in Section 4.4.

4.3. Inference

All models are evaluated with the same inference pipeline, applied post-hoc to every checkpoint.

Step 1 – BIO decoding and confidence scoring. Each title or abstract is tokenised with offset mapping enabled. The predicted BIO sequence is decoded into entity spans, and each span is assigned a confidence score computed as the median token probability within the span:

$$\text{score}(e) = \text{median}_{t \in e}\, p(y_t \mid \mathbf{x})$$

We use the median rather than the mean as a design choice intended to reduce the influence of low-confidence boundary tokens on the span score.

Step 2 – Label-specific thresholding. We apply a two-pass thresholding strategy. In the first pass, entities are kept only if their confidence exceeds a high threshold specific to the predicted label. In the second pass, lower thresholds are used only for a small set of recall-critical labels (`chemical`, `bacteria`, `food`, and `dietary supplement`). This design improves recall for difficult labels while keeping precision under control.

Step 3 – Label-aware post-processing. We apply a small set of deterministic label-specific heuristics. Obvious false positives are removed (very short spans, markup fragments, overly generic terms). Systematic confusions between semantically close categories are corrected: gene-like mentions predicted as `chemical` are remapped to `gene`, and vice versa; similarly for `food` vs. `dietary supplement`.

Table 1
NER results on the development set. Best value per column in bold.

Model	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1
NB1	0.8264	0.7215	0.7648	0.8638	0.7572	0.8070
NB1b	0.8091	0.7587	0.7749	0.8480	0.7925	0.8194
$NB1b_{pubmed}$	**0.8305**	0.7728	0.7973	**0.8693**	0.7886	0.8270
NB2	0.8117	0.7307	0.7638	0.8496	0.7616	0.8032
NB3	0.8000	0.6910	0.7381	0.8546	0.7322	0.7887
NB4	0.8027	0.7130	0.7506	0.8548	0.7449	0.7961
NB4b	0.8202	0.7352	0.7685	0.8645	0.7644	0.8114
Ensemble	0.804	**0.8068**	**0.802**	0.8403	**0.8247**	**0.8324**

Step 4 – Deduplication and overlap handling. Duplicate predictions are removed and overlapping spans are resolved with the same logic used by the official evaluator.

4.4. Ensemble

The final NER system combines the predictions of two models: NB1b-BioBERT and NB1b-PubMedBERT.

Motivation. Although PubMedBERT achieves the best overall performance as a single model, the two models show complementary behavior across entity types. BioBERT recovers additional correct mentions for some labels where PubMedBERT is more conservative.

Merge procedure. The final merge is performed as follows: (1) all spans predicted by PubMedBERT are kept as the base set; (2) spans predicted by BioBERT are added only if they do not overlap with any span in the base set; (3) the resulting set is deduplicated based on character offsets; (4) overlapping spans are resolved by keeping the longer span.

4.5. Experiments and Results

This subsection evaluates the NER component from three complementary perspectives. First, we compare the development-set performance of the different training configurations in order to assess the impact of data selection, backbone choice, and ensembling. Second, we analyse per-label behaviour to identify which entity categories benefit most from the proposed strategy and which remain difficult. Finally, we report the official hidden test-set results and compare the submitted runs with the organiser baseline.

Evaluation metrics. We follow the official evaluation protocol, which requires exact span match: a prediction is correct only if both label and character offsets match the gold annotation. We report Macro-F1 and Micro-F1, along with Precision and Recall.

Development set results. Table 1 reports the results of all configurations on the development set. Among single models, $NB1b_{pubmed}$ (PubMedBERT) achieves the best performance. The final ensemble further improves the results, reaching Macro-F1 $= 0.802$ and Micro-F1 $= 0.8324$, outperforming all individual models.

Effect of training data. Adding Bronze annotations (NB1b vs. NB1) improves recall and overall performance, showing that noisy data can still provide useful signal. Curriculum-based strategies (NB2, NB3) do not outperform single-stage training, likely due to distribution differences across datasets. Annotator-based weighting (NB4) increases precision but reduces recall.

Table 2
Per-label F1 on the development set for selected configurations.

Label	NB1	NB1b	$\text{NB1b}_{\text{pubmed}}$	Ensemble
DDF	0.872	0.893	0.878	0.893
anatomical location	0.768	0.758	0.832	0.832
animal	0.852	0.852	0.823	0.864
bacteria	0.715	0.748	0.742	0.748
biomedical technique	0.672	0.698	0.700	0.700
chemical	0.696	0.722	0.738	0.738
dietary supplement	0.741	0.767	0.786	0.786
drug	0.857	0.855	0.850	0.868
food	0.591	0.614	0.746	0.746
gene	0.627	0.606	0.652	0.652
human	0.917	0.904	0.923	0.938
microbiome	0.921	0.910	0.919	0.934
statistical technique	0.714	0.747	0.778	0.778

Table 3
NER results on the official test set (Subtask 6.1.1). The organizer baseline is GLiNER.

Run	System	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1
Baseline	GLiNER	0.7114	0.7480	0.7267	0.7782	0.8221	0.7996
R1	NERensemble	0.7367	0.7311	0.7302	0.8109	0.7932	**0.8019**
R2	NERpubmedbert	0.7667	0.6985	0.7270	0.8415	0.7554	0.7961
R3	NERbiobert	0.7290	0.6983	0.7101	0.8175	0.7739	0.7951

Effect of backbone. Replacing BioBERT with PubMedBERT yields the largest improvement among single models, consistent with prior findings that in-domain pre-training and vocabulary benefit biomedical NLP tasks [8].

Ensemble. The ensemble combines PubMedBERT predictions with non-overlapping spans from BioBERT. This simple strategy improves recall while preserving precision, consistently outperforming both individual models.

Per-label analysis. Table 2 reports per-label F1 scores. Frequent labels such as DDF, human, and microbiome achieve high performance, while more challenging categories such as food and gene remain harder due to data scarcity and label ambiguity. These per-label scores are computed on the development set; the official evaluation releases only aggregate scores on the hidden test set, so per-label test F1 cannot be reported directly. The aggregate scores are nonetheless informative: from development to test, Micro-F1 drops by only 0.031 ($0.8324 \rightarrow 0.8019$), whereas Macro-F1 drops more than twice as much, by 0.072 ($0.802 \rightarrow 0.7302$). Because Macro-F1 weights every label equally, this gap indicates that the degradation is concentrated in the low-frequency categories rather than spread uniformly. The rare labels that are already weakest on the development set — `food`, `gene`, and `statistical technique` — are therefore the most likely to collapse on the hidden test set.

Official test-set results. Table 3 reports the final submission results on the official hidden test set. The ensemble (R1) achieves Micro-F1 0.8019, outperforming the organizer baseline (Micro-F1 0.7996) and both single-model submissions. Among all participating teams, R1 ranks 5th by Micro-F1.

Notably, the ensemble (R1) achieves a better balance of precision and recall compared to both single-model runs. The PubMedBERT model (R2) yields the highest precision on the test set (Mic-P 0.8415) but lower recall, while the ensemble recovers additional true positives from BioBERT at minimal precision cost.

Overall, the experiments suggest that increased data diversity outweighed the additional annotation noise introduced by Bronze examples. Rare labels remain difficult despite augmentation. Long-context architectures appear particularly beneficial for biomedical abstracts containing multiple entity mentions distributed across long discourse segments.

5. Mention-Level Relation Extraction

We frame M-RE as a pairwise classification task over pre-identified entity mentions. Given a document and its entity mentions, every candidate pair (subject, object) whose entity types are compatible with the official relation schema is classified into one of 17 semantic predicates or `no_relation`, yielding 18 classes in total. The schema defines 55 legal type-pair patterns across 52 unique subject–object type combinations, with predicates such as *influence*, *is linked to*, *change abundance*, and *impact*.

5.1. System Overview

We submitted two runs for Subtask 6.2.1, both based on PubMedBERT-large [8] fine-tuned with hard negative sampling and mention-mean pooling, differing only in the context window provided to the model:

- R1 (REfullctx): full title–abstract as context (A5-fullctx); dev Micro-F1 0.5961 — primary submission;
- R2 (RElargewindow): local ± 300-character window around the entity pair (A5-large); dev Micro-F1 0.5932.

In both cases, entity mentions at inference time are taken from the NER ensemble (R1 of Subtask 6.1.1), making the full system an end-to-end pipeline over automatically predicted entities. R1 was selected as primary submission based on its superior Micro-F1 on the development set.

5.2. Input Representation

Candidate generation. For each document, title and abstract are concatenated into a single string, with abstract character offsets shifted by $|title| + 1$. Candidate pairs are generated by enumerating all ordered mention pairs (subject, object) whose entity type combination appears in the legal-pair dictionary (52 unique type pairs, 55 legal patterns). Pairs whose subject and object start positions differ by more than 400 characters are discarded.

The training set contains 4,921 documents (639 Gold, 811 Silver, 499 Silver 2025, 2,972 Bronze), yielding 53,791 positive and 194,662 negative examples (ratio 3.6:1). At inference time, entity mentions are provided by the NER ensemble (SMTE T611 R1), making the M-RE submission a true end-to-end pipeline over automatically predicted entities. Development-set M-RE results, in contrast, are reported on *gold* entity mentions, in order to isolate relation-extraction quality from NER errors, since gold entities are not available for the test set.

Entity markers and context window. Each candidate pair is represented as a text window around the two mentions, with four special tokens inserted at the mention boundaries: `[E1]`…`[/E1]` wraps the subject and `[E2]`…`[/E2]` wraps the object [10]. The marked string is tokenised with a maximum length of 512 tokens. Configuration A5-fullctx uses the full concatenated title–abstract, while other variants use a local window of ± 300 characters.

Entity representation. The contextual representation of each entity is obtained by mention-mean pooling: the hidden states of all tokens between the opening and closing marker tokens are averaged.

Formally, let $\mathbf{h}_t \in \mathbb{R}^d$ be the encoder hidden state of token t and let S_i denote the set of token positions spanned by entity e_i (i.e. the tokens enclosed by its marker pair). The entity representation is

$$\mathbf{h}_{e_i} = \frac{1}{|S_i|} \sum_{t \in S_i} \mathbf{h}_t, \tag{1}$$

and the pair representation $\mathbf{z} = [\mathbf{h}_{e_1} ; \mathbf{h}_{e_2}] \in \mathbb{R}^{2d}$ is obtained by concatenating the two entity vectors. This is passed through a linear classification head to produce logits over the 18 relation classes.

5.3. Training Strategy

Backbone and shared hyperparameters. The shared training setup is: learning rate 1×10^{-5} with linear decay and 6% warmup; weight decay 0.01; gradient clipping at norm 1.0; per-device batch size 4 with gradient accumulation over 8 steps (effective batch size 32); 2 epochs; bf16 mixed precision with gradient checkpointing. The best checkpoint is selected on the development set using positive-class Micro-F1.

Experimental configurations.

- A0 – Baseline (marker-only). Entity representation is the hidden state of the `[E1]` start token. Training uses negative-to-positive multiplier 5 with uniform sampling. Micro-F1 0.5753.
- A3 – Mention-mean pooling. Switches entity representation to mention-mean pooling, improving recall for multi-token entities (Micro-F1 0.5780, $+0.0027$).
- A4 – Label smoothing. Extends A3 with label smoothing ($\varepsilon = 0.1$). Does not improve (Micro-F1 0.5748).
- A5 – Hard negative sampling. For each document, 70% of the negatives are drawn from hard negatives (pairs where at least one entity type appears among the positive pairs) and 30% from random easy negatives. With PubMedBERT-base: Micro-F1 0.5899; with PubMedBERT-large (A5-large): Micro-F1 0.5932.
- A5-BioLink. Same as A5 but using BioLinkBERT-base [11] as backbone. Micro-F1 0.5845.
- A5-fullctx – Full-context input (best configuration). Identical to A5-large but uses the full title–abstract string instead of the $\pm$300-character local window. Micro-F1 0.5961 ($+0.0208$ vs. A0).
- A6-typed – Typed entity markers (no hard negatives). Replaces the generic `[E1]`/`[E2]` tokens with type-specific markers (e.g., `[BACTERIA]`…`[/BACTERIA]`), adding 26 new special tokens, on top of uniform negative sampling. This is the worst configuration (Micro-F1 0.5005).
- A7 – Typed entity markers with hard negative sampling. Combines the type-specific markers of A6-typed with the 70/30 hard-negative scheme of A5. It still decreases performance relative to generic markers (Micro-F1 0.5444, -0.0309 vs. A0).

5.4. Inference

All models share the same inference pipeline.

Step 1 – Logit caching. All legal candidate pairs are enumerated and fed to the model in batches of 16. Raw logits are stored in a per-document cache on disk.

Step 2 – Legal-only decoding with temperature scaling. Decoding is restricted to the predicates legally allowed for the given subject–object type pair. Logits are divided by temperature $T = 1.25$ and re-normalised via softmax.

Step 3 – Distance-aware minimum probability. A variable probability floor is applied as a function of character distance: $p_{\min} = 0.10$ for pairs closer than 150 characters, $p_{\min} = 0.20$ for pairs between 150 and 300 characters, and $p_{\min} = 0.30$ for pairs further than 300 characters.

Table 4
M-RE results on the development set (official evaluator). Metric: Micro-F1.

Run	Backbone	Pooling	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1
A0	PubMedBERT-base	marker-only	**0.5383**	0.5028	0.4981	**0.6336**	0.5268	0.5753
A3	PubMedBERT-base	mention-mean	0.5146	0.4903	0.4782	0.6195	0.5417	0.5780
A4	PubMedBERT-base	mention-mean	0.5208	0.5166	0.4956	0.6144	0.5399	0.5748
A7	PubMedBERT-base	mention-mean	0.4763	0.5548	0.4796	0.5304	0.5593	0.5444
A6-typed	PubMedBERT-base	mention-mean	0.4780	0.3555	0.3828	0.6219	0.4188	0.5005
A5	PubMedBERT-base	mention-mean	0.5243	0.5550	0.5142	0.5907	0.5891	0.5899
A5-BioLink	BioLinkBERT-base	mention-mean	0.4554	0.4855	0.4519	0.5951	0.5742	0.5845
A5-large	PubMedBERT-large	mention-mean	0.4970	**0.5933**	**0.5155**	0.5652	**0.6242**	0.5932
A5-fullctx	PubMedBERT-large	mention-mean	0.5151	0.5751	0.5149	0.6041	0.5882	**0.5961**

Step 4 – Per-predicate threshold tuning. Starting from fixed global defaults (margin $= 0.25$, $p_{\min} = 0.10$, maximum pair distance $= 250$ characters), four thresholds are tuned independently for each of the 17 predicates on the development set by maximising per-predicate Micro-F1 in four sequential greedy steps: (i) *margin*, (ii) *min_prob*, (iii) *max_chars*, (iv) *top2_gap*.

Step 5 – Deduplication. When the same mention pair receives multiple predictions, only the one with the highest adjusted probability is retained.

5.5. Ensemble

Following the NER approach, we explored post-hoc ensemble strategies combining the two strongest M-RE models: A5-fullctx (Micro-F1 0.5961) and A5-large (Micro-F1 0.5932). We evaluated five combination strategies on the development set without additional training. Table 5 reports the results. None of the strategies improves over the single A5-fullctx model: the best ensemble (intersection) reaches Micro-F1 0.5885, still -0.0076 below. We therefore submitted both models independently as R1 (A5-fullctx) and R2 (A5-large), corresponding to the two strongest single-model configurations.

5.6. Experiments and Results

Evaluation metrics. We follow the official evaluation protocol described in the GutBrainIE task overview [2]: a relation triple is correct only if subject mention, predicate, and object mention all match the gold annotation exactly. We report Macro-F1 and Micro-F1; the official ranking metric is Micro-F1.

Development set results. Table 4 reports development-set results for all configurations. The baseline A0 achieves Micro-F1 0.5753. Hard negative sampling (A5) is the most effective single training intervention at fixed model size ($+0.0146$ over the A0 baseline with PubMedBERT-base); scaling the same configuration to PubMedBERT-large (A5-large) yields a further gain, reaching Micro-F1 0.5932. Extending the context window to the full abstract (A5-fullctx) yields the best result, Micro-F1 **0.5961** ($+0.0208$ vs. A0).

Analysis. Hard negative sampling (A5) is the most impactful single training intervention. Compared to the baseline (A0), it substantially improves recall (from 0.5268 to 0.5891) at a modest precision cost, yielding a net gain of $+0.0146$ Micro-F1. This confirms that the dominant error mode is missing genuine relations rather than predicting spurious ones.

Scaling the backbone from base to large (A5-large) further increases recall to 0.6242 — the highest across all configurations — while precision drops slightly. A5-large also achieves the best Macro-F1 (0.5155) and Macro-recall (0.5933), suggesting that the larger model is better at handling rare predicates.

Table 5
Ensemble results on the development set (A5-fullctx + A5-large).

Strategy	Mac-F1	Mic-F1	Mic-P	Mic-R
A5-fullctx (single)	0.5149	**0.5961**	0.6041	0.5882
A5-large (single)	**0.5155**	0.5932	0.5652	**0.6242**
Weighted	0.5125	0.5836	0.6154	0.5549
Hybrid A5+large (rare)	0.5029	0.5826	0.5850	0.5803
Intersection	0.5103	0.5885	**0.6321**	0.5505
Hybrid large+A5 (rare)	0.5072	0.5765	0.5464	0.6102
Union	0.4980	0.5760	0.5265	0.6356

Table 6
Contribution of each decoding component, measured as the change in development-set Micro-F1 when the component is removed from the full inference pipeline (A5-fullctx).

Component removed	Δ Micro-F1
Per-predicate thresholds	-0.014
Temperature scaling	-0.003
Distance-aware min-probability	0.000

Table 7
M-RE results on the official test set (Subtask 6.2.1). The organizer baseline is ATLOP.

Run	System	Mac-P	Mac-R	Mac-F1	Mic-P	Mic-R	Mic-F1
Baseline	ATLOP	0.3660	0.2862	0.3003	0.4444	0.3453	0.3886
R1	REfullctx	0.3301	0.3239	0.2973	0.4349	0.4042	0.4190
R2	RElargewindow	0.3459	0.3369	0.3084	0.4476	0.3997	**0.4223**

Extending the context window to the full abstract (A5-fullctx) reverses the precision–recall trade-off: precision recovers to 0.6041 while recall remains high at 0.5882, yielding the best Micro-F1 of **0.5961**. This indicates that wider context helps the model make more accurate predictions by resolving long-range and cross-sentence relations.

Typed entity markers hurt performance in both tested variants (A7 and A6-typed), with A6-typed producing the worst result (Micro-F1 0.5005). The low frequency of each typed token during training prevents the model from learning useful representations for the 26 additional special tokens.

Decoding ablation. To justify the decoding rules, we ablate each component of the inference pipeline on the development set, removing one component at a time from the full configuration and measuring the change in Micro-F1 (Table 6). Per-predicate threshold tuning is by far the dominant component: reverting to a single global threshold reduces Micro-F1 by 0.014. Temperature scaling contributes a smaller gain (0.003). The distance-aware minimum-probability floor has no measurable effect on the development set (0.000); it is retained in our submitted configuration but could be removed without loss. Overall, decoding performance is driven primarily by per-predicate calibration rather than by the auxiliary heuristics.

Official test-set results. Table 7 reports the final submission results on the official hidden test set. R1 (A5-fullctx) achieves Micro-F1 0.4190, outperforming the organizer baseline (Micro-F1 0.3886) by $+0.0304$. R2 (A5-large, $\pm$300-character local window) achieves Micro-F1 0.4223, the best of our two submissions and the 2nd highest Micro-F1 among all participating teams for this subtask.

Interestingly, R2 (A5-large, $\pm$300-character local window) outperforms R1 (A5-fullctx, full abstract) on the test set, reversing the relative ordering observed on the development set. This ordering does

not transfer directly to the test set; as our oracle analysis shows, most of the development-to-test gap reflects the change in entity source (gold entities on the development set, predicted entities on the test set) rather than overfitting. Both runs substantially improve over the baseline on Micro-F1, with R2 surpassing it by $+0.0337$.

6. Conclusion

We presented our system for biomedical Named Entity Recognition and Mention-Level Relation Extraction developed for the GutBrainIE shared task at CLEF 2026.

For NER, our hybrid ensemble of BioBERT and PubMedBERT, trained on heterogeneous Gold/Silver/Bronze data with priority-based merging, achieves Micro-F1 0.8019 on the official test set, ranking 5th among all participating teams and outperforming the GLiNER baseline. The results confirm that (i) including noisy Bronze data in single-stage training is beneficial, (ii) domain-specific pre-training (PubMedBERT) yields larger gains than curriculum or annotator-weighting strategies, and (iii) model ensembling provides consistent improvements by combining complementary predictions.

For M-RE, our pairwise classification approach with mention-mean pooling and hard negative sampling achieves Micro-F1 0.4223 on the test set, ranking 2nd among all participating teams and improving over the ATLOP baseline by $+0.0337$. Hard negative sampling is the most impactful training intervention, confirming that exposing the model to plausible-but-negative pairs directly addresses the dominant error mode of missing genuine relations. The apparent reversal between development and test rankings is largely explained by this change in entity source rather than by threshold overfitting: once the entity source is held fixed, the relation model generalises well from development to test, as quantified by our oracle analysis below.

Both M-RE submissions use entity mentions predicted by the NER ensemble rather than gold annotations, making the system a realistic end-to-end pipeline. To quantify how much NER errors cost the relation stage, we run the relation model with an identical decoding configuration on gold versus predicted entities over the development set. Replacing gold with predicted entity mentions reduces Micro-F1 by 0.137 (from 0.566 to 0.429), identifying NER error propagation — rather than relation classification — as the dominant bottleneck of the end-to-end pipeline. This also explains the apparent development-to-test gap: because development scores are reported on gold entities and the test on predicted entities, most of the gap reflects the change in entity source rather than overfitting. Indeed, with the entity source held fixed the relation model scores 0.429 on the development set and 0.419 on the test set, indicating that it generalises well and that the residual reversal between the two context-window configurations is a comparatively small effect.

7. Code and Data Availability

The code and models used to produce all submitted runs, together with the training and inference configurations, are publicly available at https://github.com/sofiamaule/SMTE-GutBrainIE.git. Our system pipeline is built on the baseline code made available by the course tutor.[1]. The GutBrainIE datasets are provided by the task organisers upon registration [2].

Acknowledgments

This work has been partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074, and it is part of the initiatives of the Center for Studies in Computational Terminology (CENTRICO) of the University of Padua and in the research directions of the Italian Common Language Resources and Technology Infrastructure CLARIN-IT.

[1] https://github.com/MMartinelli-hub/ATA_Tutorship

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: drafting content, grammar and spelling check, paraphrase and reword and, peer review simulation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, I. Rozhkov, E. Tutubalina, D. Dimitriadis, V. Patsiou, 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 2026: The fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, volume Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026) of *Lecture Notes in Computer Science*, Springer, 2026.

[2] M. Martinelli, V. Bonato, G. M. Di Nunzio, G. Faggioli, N. Ferro, O. Irrera, B. Kantz, S. Marchesin, L. Menotti, S. Merlo, R. Michelotto, G. Tussardi, F. Vezzani, P. Waldert, G. Silvello, Overview of GutBrainIE@CLEF 2026: Gut-Brain Interplay Information Extraction, in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[3] M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, CEUR, Madrid, Spain, 2025, pp. 65–98.

[4] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. R. Ortega, N. 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, Springer Nature Switzerland, Cham, 2025, pp. 407–415. doi:10.1007/978-3-031-88720-8_61.

[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. D. 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, in: J. Carrillo-de-Albornoz, A. García Seco de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2026, pp. 173–198. doi:10.1007/978-3-032-04354-2_12.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019, pp. 4171–4186.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240. doi:10.1093/bioinformatics/btz682.

[8] 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 3 (2021) 1–23. doi:10.1145/3458754.

[9] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist model for named entity recognition using bidirectional transformer, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2024, pp. 5364–5376. URL: https://aclanthology.org/2024.naacl-long.300/.

[10] L. Baldini Soares, N. FitzGerald, J. Ling, T. Kwiatkowski, Matching the blanks: Distributional similarity for relation learning, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2895–2905. URL: https://aclanthology.org/P19-1279/. doi:10.18653/v1/P19-1279.

[11] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 8003–8016. doi:10.18653/v1/2022.acl-long.551.

[12] 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, volume 35, 2021, pp. 14612–14620. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17717. doi:10.1609/aaai.v35i16.17717.

[13] M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 65–98. URL: https://ceur-ws.org/Vol-4038/paper_5.pdf.

[14] M. Martinelli, S. Marchesin, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, L. Menotti, F. Vezzani, G. Silvello, A domain-specific curated benchmark for entity and document-level relation extraction, in: V. Demberg, K. Inui, L. Marquez (Eds.), Findings of the Association for Computational Linguistics: EACL 2026, Association for Computational Linguistics, Rabat, Morocco, 2026, pp. 5693–5711. doi:10.18653/v1/2026.findings-eacl.301.

[15] V. Bonato, L. Menotti, F. Vezzani, G. M. D. Nunzio, G. Silvello, Defining and Relating Concepts in Medical Terminology: A Case Study on the Gut–Brain Interplay, Journal of Digital Terminology and Lexicography 2 (2026) 37–55.

[16] 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.), CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 302–316.

[17] H. P. Gupta, R. Banerjee, SCIRE at BioASQ 2025: LLM Driven Biomedical Named Entity Recognition for GutBrainIE 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 281–291.

[18] 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.), CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 226–242.

[19] 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.), CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 359–372.