TEXA at GutBrainIE 2026: A Multi-Stage Approach for Entity Recognition, Linking, and Relation Extraction in Gut-Brain Biomedical Abstracts

Danil Smirnov[1,*], Bita Khashechian[1,*] and Giorgio Maria Di Nunzio[1]

[1]Department of Information Engineering, University of Padova, Via Gradenigo 6/b, 35131 Padova, Italy

Abstract

This paper describes the TEXA system submitted to GutBrainIE 2026, Task #6 of the BioASQ Lab at CLEF 2026. The task addresses information extraction from biomedical abstracts concerning the gut microbiota and its links to neurological and mental-health conditions, including Alzheimer's disease, Parkinson's disease, multiple sclerosis, amyotrophic lateral sclerosis, and mental health. We participated in the four subtasks: named entity recognition (NER), named entity recognition and disambiguation (NERD), mention-level relation extraction (M-RE), and concept-level relation extraction (C-RE). Our system consists of multiple stages in which entity mentions are first detected using a fine-tuned GLiNER model, linked to biomedical concepts through a hybrid entity-linking component based on observed mention–URI mappings and URI definitions, and then used as input to a fine-tuned ATLOP relation extraction model. The outputs of these components are converted into the official JSON formats and packaged for the four required submission folders.

Keywords

GutBrainIE, Biomedical Information Extraction, Named Entity Recognition, Entity Linking, Relation Extraction, GLiNER, ATLOP

1. Introduction

Biomedical literature contains a rapidly growing body of research on the interaction between the gut microbiota and the central nervous system. This interaction, commonly referred to as the gut-brain axis, has become increasingly relevant in the study of neurological and mental-health conditions. PubMed abstracts in this area mention a wide range of biomedical entities, including diseases, microorganisms, chemicals, genes, anatomical locations, human participant groups such as patients and controls, and experimental techniques. They also describe semantic relations among these entities, such as associations between bacteria and diseases, effects of chemicals on biological processes, or links between microbiome-related concepts and anatomical regions.

Extracting this information manually from the biomedical literature is costly and difficult to scale. Automating this process therefore requires more than identifying relevant spans in text. A complete information extraction system must also normalize entity mentions to biomedical concept identifiers and identify meaningful relations between them. This is particularly difficult in biomedical texts, where entities are often expressed through abbreviations, synonyms, long noun phrases, or context-dependent surface forms. In addition, relations may be expressed indirectly or supported by evidence distributed across the document rather than through simple local patterns.

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

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

*Corresponding author.

danil.smirnov@studenti.unipd.it (D. Smirnov); bita.khashechian@studenti.unipd.it (B. Khashechian); giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

0000-0001-7116-9338 (G. M. Di Nunzio)

recognition and disambiguation (NERD), and relation extraction (RE), which is evaluated at two levels: mention-level relation extraction (M-RE) and concept-level relation extraction (C-RE). These subtasks correspond to increasingly demanding levels of biomedical information extraction. 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 TEXA system for GutBrainIE 2026. The system consists of three connected stages: entity recognition, entity linking, and relation extraction. In the first stage, a GLiNER model initialized from the `numind/NuNerZero` checkpoint is fine-tuned to recognize the official GutBrainIE entity categories. In the second stage, predicted entity mentions are linked to biomedical URIs using a hybrid strategy that combines exact matching over training annotations with similarity-based retrieval over URI definitions. In the third stage, relation extraction is performed using an ATLOP-style document-level model [6], after converting predicted entities into a DocRED-compatible representation.

This design was chosen for both practical and methodological reasons. Each component addresses a distinct part of the task and can therefore be trained, evaluated, and debugged independently. At the same time, the staged structure reflects the dependency pattern of the shared task: relation extraction depends on entity recognition, and concept-level relation extraction additionally depends on correct URI assignment. This makes it possible to analyze not only the final predictions, but also the propagation of errors across the different stages of the system.

The remainder of the paper is organized as follows. Section 2 reviews related work on biomedical named entity recognition, entity linking, and relation extraction. Section 3 describes the GutBrainIE task and dataset. Section 4 presents the system architecture. Section 5 describes the experimental setup and implementation details. Section 6 reports and analyzes the results. Section 7 discusses the strengths and limitations of the system. Section 8 concludes the paper, and Section 9 outlines directions for future work.

2. Related Work

Biomedical information extraction aims to convert unstructured scientific text into structured representations that support search, analysis, and knowledge discovery. In the setting of GutBrainIE, this process involves three closely connected tasks: identifying biomedical entity mentions, normalizing these mentions to concept identifiers, and extracting semantic relations between them. Our system is therefore situated at the intersection of research on biomedical named entity recognition, entity linking, and relation extraction.

The GutBrainIE benchmark itself is closely related to recent work on curated biomedical information extraction resources for the gut-brain axis. Martinelli et al. [7] 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. [8] analyze 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 specialized biomedical domain.

Named entity recognition is a core task in biomedical natural language processing. It aims to identify spans corresponding to relevant categories such as diseases, chemicals, animals, humans, genes, anatomical locations, and experimental techniques. Early neural approaches typically formulated NER as a sequence labeling problem [9], while more recent systems rely heavily on transformer-based contextual encoders such as BERT [10]. Biomedical terminology is often domain-specific because many

terms refer to specialized biological entities, processes, or experimental procedures. At the same time, ambiguity may arise from abbreviations, synonymy, lexical variation, and context-dependent surface forms that can refer to different biomedical concepts. In addition, entity boundaries in biomedical abstracts may be long, nested, or coordinated, which makes span detection more difficult than in simpler NER settings. In this work, entity recognition is performed with GLiNER [11], which is well suited to settings with heterogeneous entity schemas and can be adapted to the GutBrainIE label space through task-specific fine-tuning.

Entity linking, also referred to as normalization or disambiguation, maps recognized mentions to identifiers in biomedical knowledge resources. This step is especially important in biomedical literature because the same concept may be expressed through multiple surface forms, abbreviations, or near-synonymous expressions, while the same string may refer to different concepts depending on context. Biomedical linking systems commonly combine dictionary-based matching, ontology lookup, rule-based normalization, and representation-based retrieval. Exact matching is often highly precise when the normalized surface form of a mention has already been observed, but it typically suffers from limited coverage. Similarity-based approaches can improve recall by comparing mentions with concept names, definitions, or synonyms, although they may introduce additional ambiguity when related concepts are lexically or semantically close. The NERD component used in our system follows a hybrid strategy of this kind, combining exact mention–URI mappings derived from the training annotations with similarity-based retrieval over URI definitions.

Relation extraction aims to identify typed semantic associations between entity pairs. In biomedical texts, this problem has been studied in tasks such as chemical–disease, gene–disease, drug–drug, and protein–protein relation extraction, including benchmark resources such as CDR [12] and BioRED [13]. Traditional approaches relied on hand-crafted features, syntactic paths, and trigger patterns, whereas more recent methods employ neural encoders based on convolutional, recurrent, or transformer architectures. Despite these advances, biomedical relation extraction remains difficult because relevant evidence is often distributed across long spans, multiple mentions, or different parts of the document. This is particularly relevant in PubMed abstracts, where relations may be expressed indirectly or supported by broader discourse context rather than by a single local expression.

Document-level relation extraction extends this setting beyond sentence boundaries and is therefore especially relevant for tasks such as GutBrainIE. When entities appear multiple times across the title and abstract, relation prediction can benefit from aggregating information over the whole document rather than focusing only on local contexts. ATLOP is a document-level relation extraction architecture based on adaptive thresholding and localized context pooling [6]. In our system, an ATLOP-style model is used after converting GutBrainIE annotations and predicted entities into a DocRED-compatible representation.

A further challenge in GutBrainIE is the distinction between mention-level and concept-level relation extraction. Mention-level relation extraction evaluates whether the system identifies the correct relation between textual entity mentions, whereas concept-level relation extraction additionally requires correct normalization of both arguments to their URIs. As a result, concept-level relation extraction is not only a harder prediction problem, but also a stricter evaluation of end-to-end system consistency. This makes GutBrainIE particularly suitable for studying multi-stage biomedical information extraction systems, where downstream performance depends strongly on the quality of earlier predictions.

3. Task and Dataset Description

3.1. Shared Task Overview

GutBrainIE 2026 is a biomedical information extraction task centered on PubMed titles and abstracts related to the interaction between the gut microbiota and the brain. The task focuses on literature concerning the gut-brain axis and its links to neurological and mental-health conditions, including Alzheimer's disease, Parkinson's disease, multiple sclerosis, amyotrophic lateral sclerosis, and mental health disorders.

Table 1
GutBrainIE subtasks addressed by the submitted system.

Subtask	Name	Required output
6.1.1	NER	Entity mentions with spans and labels
6.1.2	NERD	Entity mentions with labels and normalized URIs
6.2.1	M-RE	Relations between textual entity mentions
6.2.2	C-RE	Relations between normalized biomedical concepts

The shared task is organized into four subtasks covering two broad information extraction problems. The first concerns entity extraction and normalization: Subtask 6.1.1 targets named entity recognition (NER), while Subtask 6.1.2 targets named entity recognition and disambiguation (NERD). The second concerns relation extraction: Subtask 6.2.1 evaluates mention-level relation extraction, and Subtask 6.2.2 evaluates concept-level relation extraction. Table 1 summarizes the four subtasks addressed in this work.

The four subtasks are closely connected but evaluate different levels of structure. NER focuses on detecting entity spans and assigning entity labels. NERD extends this requirement by adding concept normalization through URIs. M-RE evaluates relations between textual mentions, whereas C-RE evaluates relations between normalized concepts. As a result, the later subtasks are not only more difficult in isolation, but also more sensitive to errors produced upstream.

3.2. Dataset Description

The GutBrainIE data are provided primarily in JSON format and consist of PubMed titles and abstracts together with structured entity and relation annotations. Each article is indexed by PMID and contains the textual fields used for prediction, along with annotations when available.

The training material is divided into multiple collections, including Gold, Silver, Silver 2025, and Bronze. These collections differ in annotation quality and provenance and are used differently across the stages of the system. In our local setup, the article files used for training contain 639 Gold documents, 811 Silver documents, 499 Silver 2025 documents, and 2,972 Bronze documents. The development and test article files each contain 80 documents.

In our experiments, the Gold, Silver, and Silver 2025 collections are used as the main supervised data for entity recognition and as the annotated component for relation extraction. The Bronze collection is used as an additional distant or weakly supervised source for relation extraction. This distinction reflects the different sensitivity of the stages to annotation quality: entity recognition benefits most directly from cleaner span-level supervision, whereas relation extraction can also benefit from broader but noisier relation evidence.

The official test set contains only article identifiers, titles, and abstracts, without gold annotations. Entity mentions, URI assignments, and relations must therefore be predicted automatically and exported in the official JSON format. Local experimentation is carried out on the development split, while the test split is reserved for final submission generation.

3.3. Annotation Format

In the JSON representation, each article contains the title and abstract together with optional annotation fields for entities and relations. Entity annotations include character offsets, the text span, the location of the mention in the title or abstract, the entity label, and the associated URI. These fields support both NER and NERD evaluation.

Relation annotations are represented at two levels. Mention-level relations are defined between textual mentions and include the subject text span, subject label, predicate, object text span, and object label. Concept-level relations are defined between normalized concepts and include the subject URI,

subject label, predicate, object URI, and object label. This design makes it possible to evaluate both surface-level relation extraction and normalized knowledge extraction within the same task framework.

One important consequence of this annotation scheme is that errors at different levels have different effects. A system may correctly identify a mention span but fail to assign the correct URI. Similarly, it may predict a plausible textual relation while still failing concept-level evaluation if one or both arguments are normalized incorrectly. The task therefore evaluates not only isolated component accuracy, but also cross-stage consistency.

3.4. Entity and Relation Schema

The entity schema contains 13 labels: anatomical location, animal, bacteria, biomedical technique, chemical, DDF (disease, disorder, or finding), dietary supplement, drug, food, gene, human, microbiome, and statistical technique. These categories reflect the biomedical scope of the task and include both broad biomedical types, such as chemical, gene, and drug, and more domain-specific categories relevant to the gut-brain literature, such as microbiome and bacteria.

The relation schema contains 17 predicates: administered, affect, change abundance, change effect, change expression, compared to, impact, influence, interact, is a, is linked to, located in, part of, produced by, strike, target, and used by. These predicates represent a range of semantic associations between biomedical entities, from relatively general relations such as affect and influence to more structurally specific ones such as located in, part of, and produced by.

The distinction between mention-level and concept-level relations is central to the task definition. Mention-level relations are represented using textual entity mentions and their labels, whereas concept-level relations require normalized URIs for both subject and object. Consequently, concept-level relation extraction depends on the combined correctness of mention detection, entity labeling, URI assignment, and relation prediction.

3.5. Evaluation Perspective

The official evaluation reports macro- and micro-averaged precision, recall, and F_1 across the four subtasks. Macro-averaged scores compute performance per class and then average across classes, giving equal importance to frequent and rare labels. Micro-averaged scores aggregate predictions globally before computing the metrics and therefore reflect overall performance weighted by class frequency.

For NER, a prediction is correct only when the span, location, text span, and entity label match the reference annotation. For NERD, the URI must also match. For mention-level relation extraction, correctness depends on the subject mention, subject label, predicate, object mention, and object label. For concept-level relation extraction, correctness depends on the subject URI, subject label, predicate, object URI, and object label.

This evaluation design makes the later subtasks particularly demanding. Concept-level relation extraction is the strictest setting because it requires a chain of correct decisions: the subject and object must be detected correctly, labeled correctly, normalized correctly, and connected by the correct predicate. For this reason, GutBrainIE provides a useful benchmark for studying multi-stage biomedical information extraction systems and the propagation of errors across their components.

4. System Description

4.1. System Overview

The submitted system follows a multi-stage architecture designed to address all four GutBrainIE subtasks. It consists of three main components: named entity recognition, entity linking, and relation extraction. Each component produces an intermediate JSON output that is either evaluated directly or used as input to the following stage.

Table 2
Overview of the system architecture.

Stage	Input	Output
Named Entity Recognition	Article title and abstract	Entity mentions with spans, labels, and scores
Entity Linking	Predicted entity mentions	Entity mentions enriched with URIs
Relation Extraction	Predicted entities in DocRED-style format	Predicted relations between entity pairs
Submission Formatting	NER, linking, and RE outputs	Official JSON files for all subtasks

Table 2 summarizes the inputs and outputs of each stage. The first stage detects biomedical entity mentions in PubMed titles and abstracts. This stage is implemented with a GLiNER model initialized from the `numind/NuNerZero` checkpoint and fine-tuned on the GutBrainIE training data. The second stage assigns normalized biomedical URIs to the predicted entity mentions. This is implemented as a hybrid entity-linking module that combines exact matching over observed mention–URI mappings with similarity-based retrieval over URI definitions. The third stage predicts relations between entity pairs using an ATLOP-style document-level relation extraction model. For this stage, predicted entities are converted into a DocRED-compatible representation before being passed to the relation extraction component.

This design allows the system to produce outputs for all four subtasks. The NER output is used for Subtask 6.1.1, the linked entity output is used for Subtask 6.1.2, and the relation extraction output is converted into mention-level and concept-level relations for Subtasks 6.2.1 and 6.2.2, respectively.

4.2. Named Entity Recognition

The named entity recognition component is based on GLiNER, initialized from the `numind/NuNerZero` checkpoint. Its objective is to identify entity mentions in article titles and abstracts and assign one of the 13 official GutBrainIE entity labels.

The original GutBrainIE annotations are converted into GLiNER-compatible training instances through a preprocessing step that preserves character positions while mapping character-level entity annotations to token-level spans. The NER model is trained on the Gold, Silver, and Silver 2025 collections, which provide the main supervised data for span detection and entity classification.

At inference time, titles and abstracts are processed separately. This simplifies offset handling during prediction. After inference, abstract offsets are shifted by the title length plus one character in order to align all predictions with the combined article-level representation used by the official format. The final inference configuration uses a confidence threshold of 0.65, with flat NER enabled and multi-label prediction disabled.

Duplicate predictions are removed during post-processing using span, text, label, and score information. Each final entity prediction contains the start and end offsets, text span, entity label, confidence score, and an indicator of whether the entity occurs in the title or in the abstract.

4.3. Entity Linking and Named Entity Recognition with Disambiguation

The entity-linking component assigns a normalized URI to each entity mention predicted by the NER model. This stage corresponds to the named entity recognition and disambiguation subtask and is implemented as a hybrid strategy combining exact matching and similarity-based retrieval.

The first source of evidence consists of mention–URI mappings extracted from the annotated training data. Two exact-match mappings are constructed: one based on normalized text spans alone and one based on normalized text-span–label pairs. During inference, the text-span–label mapping is used first because it is more specific. If no URI is found, the text-only mapping is applied.

For mentions that cannot be resolved through exact matching, the system falls back to similarity-based retrieval over URI definitions. The definition database is built from two main sources: text spans associated with URIs in the GutBrainIE annotations, and labels, names, or synonyms collected from external biomedical resources. These include resources such as UMLS, MeSH, CHEBI, STATO, and OBO-style ontologies [14, 15, 16, 17]. The collected definitions are normalized, deduplicated, and indexed using `txtai` with PubMedBERT-based embeddings [18]. When exact matching fails, the predicted text span is used as a query against this index, and the top retrieved entry is mapped back to its corresponding URI.

If no suitable match is found, the entity is assigned NA. The final linked output stores both the predicted URI and metadata indicating whether the assignment was obtained through text-label exact matching, text-only exact matching, similarity-based retrieval, or no match.

This hybrid design aims to balance precision and coverage. Exact matching provides reliable assignments for entity mentions already observed in the training annotations, while similarity-based retrieval improves robustness for unseen or lexically variable forms.

4.4. Relation Extraction

The relation extraction component is based on an ATLOP-style document-level relation extraction model. Its purpose is to predict one of the official GutBrainIE relation predicates for candidate entity pairs. The model is used for both mention-level and concept-level relation extraction, with the final output format determined during the submission-generation stage. ATLOP was selected because it supports document-level relation extraction with localized context pooling, which is particularly suitable for long biomedical abstracts.

Before relation extraction, the GutBrainIE data are converted into a DocRED-compatible representation. This representation contains the tokenized document, sentence structure, entity clusters, entity mention positions, candidate head–tail pairs, and relation labels. During training, Gold, Silver, and Silver 2025 are combined into an annotated training set, while Bronze is stored separately as a distant supervision set.

The preprocessing for relation extraction includes tokenization of titles and abstracts, alignment of entity offsets to token positions, sentence segmentation, and conversion of relation annotations into DocRED-style labels. The resulting structure contains the fields `vertexSet`, `labels`, `title`, and `sents`. At inference time, the relation extraction model operates on predicted entities rather than gold annotations, making the setting consistent with the full test-time scenario.

The ATLOP-style model uses a transformer document encoder configured with a maximum sequence length of 1024 tokens. This setting was chosen to preserve as much title-and-abstract context as possible while remaining compatible with the available computational resources. No explicit chunking strategy was applied. This is useful for PubMed abstracts, where relation evidence may be distributed across the title and abstract rather than expressed in a single local context. In fact, many biomedical relations require integrating evidence distributed across multiple sentences that may be far apart.

For each candidate entity pair, the model constructs head, tail, and localized context representations. Entity representations are derived from token embeddings at entity positions, and mention information is aggregated when an entity appears multiple times. Let $\mathbf{h}_i$ and $\mathbf{h}_j$ denote the projected representations of the head and tail entities for a candidate pair (i, j), and let $\mathbf{c}_{ij}$ denote the localized context representation for the same pair. Relation logits are computed with a bilinear classifier, implemented in block form in our model:

$$\mathbf{s}_{ij} = \mathbf{h}_i^\top \mathbf{W} \mathbf{h}_j + \mathbf{U}\mathbf{c}_{ij} + \mathbf{b},$$

where $\mathbf{s}_{ij}$ contains one score for each relation label, including the threshold class.

Training uses adaptive threshold loss, following ATLOP. Let $\mathcal{P}_{ij}$ be the set of positive relation labels for the entity pair (i, j), let $\mathcal{N}_{ij}$ be the set of negative relation labels, and let TH denote the threshold

class, corresponding to class 0 in our implementation. The loss is defined as:

$$\mathcal{L}_{ij} = -\sum_{r \in \mathcal{P}_{ij}} \log \frac{\exp(s_{ij}^{r})}{\exp(s_{ij}^{r}) + \exp(s_{ij}^{\mathrm{TH}})} - \log \frac{\exp(s_{ij}^{\mathrm{TH}})}{\exp(s_{ij}^{\mathrm{TH}}) + \sum_{r \in \mathcal{N}_{ij}} \exp(s_{ij}^{r})}.$$

At inference time, relation labels whose logits exceed the threshold logit are selected, optionally subject to a limit on the number of predicted labels.

4.5. Output Conversion and Submission Generation

The final stage merges the outputs of the NER, entity-linking, and relation extraction components into the official submission format. Several utility notebooks are used to convert intermediate predictions into the structures required by the four subtasks.

Raw GLiNER outputs are first converted into the official entity format, including span adjustment and merging of consecutive fragments when necessary. The linked entity predictions are then used to generate the NERD output. For relation extraction, predicted ATLOP triples are mapped back to the corresponding predicted entities. Invalid relations are removed, and the remaining relations are converted into two forms: mention-level relations, represented with subject and object text spans and labels, and concept-level relations, represented with subject and object URIs and labels.

The final output consists of four task-specific JSON files, one for each subtask. These files, together with the submission folder structure, are validated before packaging to ensure compatibility with the official evaluation workflow.

4.6. Design Rationale

The system is designed as a staged architecture rather than a single end-to-end model. This choice reflects the structure of the shared task itself, where entity recognition, concept normalization, and relation extraction correspond to distinct but connected levels of analysis. Modular architectures facilitate independent debugging, targeted optimization, and error analysis across subtasks. In this sense, it also improves interpretability, since intermediate outputs can be inspected separately to determine whether an error originates from span detection, entity labeling, URI assignment, or relation prediction.

At the same time, this design introduces error propagation. A missed entity cannot be linked and cannot participate in relation extraction. An incorrect URI can invalidate a concept-level relation even when the relation predicate is otherwise plausible. The reported results should therefore be interpreted as the behavior of a full multi-stage system, rather than as isolated component-level scores.

5. Experimental Setup

5.1. Data Preparation

The original GutBrainIE data are provided in JSON format and contain article-level text together with entity, URI, and relation annotations. Since the three main components of the system require different input formats, the data preparation stage produces several intermediate representations.

For named entity recognition, the original annotations are converted into GLiNER-compatible training instances. This conversion preserves character offsets while mapping entity annotations from character-level spans to token-level spans. The resulting format contains tokenized text together with span-level entity labels.

For entity linking, mention–URI mappings are extracted from the annotated training collections. These mappings are used to construct exact-match dictionaries based on normalized mention strings and mention–label pairs. In addition, URI definitions are collected from both the task annotations and external biomedical resources. These definitions are normalized, deduplicated, and indexed for similarity-based retrieval.

Table 3
NER training and inference configuration.

Parameter	Value
Base model	`numind/NuNerZero`
Training collections	Gold, Silver, Silver 2025
Training steps	3000
Evaluation interval	500 steps
Batch size	8
Maximum sequence length	768
Warmup ratio	0.1
Encoder learning rate	1×10^{-5}
Task-specific learning rate	5×10^{-5}
Scheduler	Cosine schedule with warmup
Inference threshold	0.65
Flat NER	True
Multi-label prediction	False

For relation extraction, the annotations are converted into a DocRED-compatible representation. This conversion includes tokenization, sentence segmentation, alignment of entity spans to token positions, construction of entity clusters, and mapping of relation annotations to head–tail entity pairs. At inference time, predicted entities are converted into the same structure so that relation extraction operates on automatically detected entities rather than gold annotations.

5.2. Named Entity Recognition Configuration

The NER component uses GLiNER initialized from the `numind/NuNerZero` checkpoint. The model is fine-tuned on the Gold, Silver, and Silver 2025 collections. Table 3 reports the main training and inference parameters.

The maximum sequence length was selected based on the empirical distribution of tokenized input lengths in the training and development data. A limit of 768 tokens provides near-complete coverage of the training instances and full coverage of the development set while remaining computationally feasible for fine-tuning on the available hardware.

At inference time, titles and abstracts are processed separately, and abstract offsets are shifted after prediction to align them with the combined title–abstract document representation required by the official output format.

5.3. Threshold Selection for NER

In addition to fine-tuning the model parameters, we tuned the NER confidence threshold on the development set. Since GLiNER returns scored entity predictions, the threshold directly controls the precision–recall trade-off and therefore affects both the standalone NER task and the later stages of the system.

We evaluated thresholds in the range from 0.40 to 0.80 and selected 0.65 for the submitted system. This value provided the best downstream behavior in the complete pipeline, balancing the number of predicted mentions available to the linking and relation extraction components with the need to avoid excessive false positives.

5.4. Entity Linking Configuration

The entity-linking component combines exact matching and similarity-based retrieval over URI definitions. Exact matching is applied first using dictionaries derived from the annotated training data. Two mappings are used: one based on normalized mention strings and one based on normalized mention–label pairs. The latter is applied first because it provides a more specific signal.

Table 4
Entity-linking assignment routes for development and test predictions.

Split	Text+label exact	Text-only exact	Similarity retrieval
Development	1,195 (45.09%)	664 (25.05%)	791 (29.84%)
Test	1,270 (45.24%)	664 (23.65%)	873 (31.10%)

Table 5
Relation extraction configuration.

Parameter	Value
Model type	ATLOP-style document-level RE
Transformer type	BERT
Base model	`bert-base-cased`
Maximum sequence length	1024
Train batch size	4
Test batch size	8
Learning rate	5×10^{-5}
Adam epsilon	1×10^{-6}
Warmup ratio	0.06
Maximum gradient norm	1.0
Training epochs	30
Seed	66
Loss function	Adaptive threshold loss
Evaluation mode	Micro by default

Table 4 reports the assignment routes used by the entity-linking component. Exact matching accounts for approximately 70.15% of development assignments and 68.89% of test assignments, while the remaining predictions rely on similarity retrieval. These figures show that the linking component depends heavily on mappings observed in the training annotations. However, coverage alone does not measure whether the assigned URI is correct; route-specific URI accuracy is left for future analysis.

For mentions not covered by exact matching, the system applies similarity-based retrieval over a database of URI definitions. These definitions are collected from the GutBrainIE annotations and external biomedical resources, normalized, deduplicated, and indexed using `txtai`. The similarity index is built with biomedical text representations motivated by PubMedBERT-style domain-specific pretraining [18]. If no suitable match is found, the mention is assigned NA.

This configuration is intended to balance precision and coverage: exact matching is reliable for previously observed forms, whereas similarity-based retrieval improves robustness for unseen or lexically variable mentions.

5.5. Relation Extraction Configuration

The relation extraction component is based on an ATLOP-style document-level relation extraction model. The input is represented in a DocRED-compatible format containing tokenized sentences, entity clusters, entity positions, candidate head–tail pairs, and relation labels.

Gold, Silver, and Silver 2025 are combined into an annotated training set, while Bronze is used as a distant supervision set. Table 5 reports the main training and inference parameters.

The RE component uses `bert-base-cased` as its encoder to remain compatible with the original ATLOP-style implementation and DocRED preprocessing pipeline. We acknowledge that this encoder is not biomedical-domain specific, unlike the representations used in the linking component. Replacing it with a biomedical encoder such as SciBERT, BioBERT, or PubMedBERT is therefore a natural direction for future work.

The model is configured with a maximum sequence length of 1024 tokens. This limit was chosen to

Table 6
Development-set results.

Task	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
NER	0.7531	0.7899	0.7688	0.8038	0.8449	0.8238
NERD	0.6044	0.6311	0.6154	0.6162	0.6478	0.6316
M-RE	0.2853	0.2844	0.2590	0.4244	0.3723	0.3966
C-RE	0.1818	0.2184	0.1812	0.2350	0.2283	0.2316

preserve as much title-and-abstract context as possible while remaining compatible with the transformer encoder and available GPU memory. No explicit chunking strategy was applied. This configuration is useful for PubMed abstracts, where relation evidence may be distributed across the title and abstract rather than expressed in a single local context.

5.6. Implementation and Hardware

The system was implemented through a combination of Python scripts and notebooks covering preprocessing, training, conversion, prediction, and validation. The three core stages are connected through JSON-based intermediate files, allowing each stage to be inspected independently and reused for the official submission format.

Fine-tuning of the GLiNER and ATLOP components was performed on an NVIDIA A100 GPU.

5.7. Evaluation Protocol

Local evaluation was performed on the development set. The official evaluator computes macro- and micro-averaged precision, recall, and F_1 for all four subtasks.

For NER, a prediction is considered correct only when the span, location, text span, and entity label match the reference annotation. For NERD, the URI must also match. For mention-level relation extraction, correctness depends on the subject mention, subject label, predicate, object mention, and object label. For concept-level relation extraction, correctness depends on the subject URI, subject label, predicate, object URI, and object label.

Duplicate entities and duplicate relations are removed before scoring. Overlapping entity predictions are resolved during post-processing so that evaluation is not distorted by repeated or conflicting outputs.

Official test results were released by the organizers after submission. The official leaderboards report the best-performing run according to micro-F_1, while the detailed evaluation spreadsheet provides macro- and micro-averaged precision, recall, and F_1 for each submitted run.

6. Results and Analysis

6.1. Development Results

Table 6 reports the development-set results of the submitted system configuration across the four subtasks. Scores are reported using macro- and micro-averaged precision, recall, and F_1.

The development results show a clear progression in task difficulty. NER is the strongest component, reaching a micro-F_1 of 0.8238 and a macro-F_1 of 0.7688. This indicates that the fine-tuned GLiNER model is effective at recovering entity mentions and assigning the official GutBrainIE labels. NERD yields lower scores, with a micro-F_1 of 0.6316 and a macro-F_1 of 0.6154, reflecting the additional difficulty introduced by URI assignment.

The relation extraction subtasks are substantially more difficult. Mention-level RE reaches a micro-F_1 of 0.3966, while concept-level RE reaches 0.2316. This decrease is expected: relation extraction depends on the availability and quality of predicted entities, and concept-level RE additionally depends on correct

Table 7
Official test-set results for the TEXA run reported in this paper.

Subtask	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
NER	0.7353	0.7529	0.7413	0.7980	0.8302	0.8138
NERD	0.4039	0.4168	0.4090	0.4581	0.4766	0.4672
M-RE	0.3804	0.2761	0.2947	0.4548	0.3382	0.3880
C-RE	0.1063	0.0991	0.0981	0.1556	0.1383	0.1464

normalization of both relation arguments. In this sense, concept-level RE is the strictest evaluation of the full system, as it combines the uncertainty of entity recognition, linking, and relation prediction.

6.2. Official Test Results

After submission, the organizers released the official GutBrainIE 2026 evaluation results. Table 7 reports the official test-set scores for the TEXA run described in this paper.

The official test results follow the same general pattern observed on the development set. NER is the strongest subtask, achieving a micro-F_1 of 0.8138, which indicates that the GLiNER-based component generalizes well to the hidden test data. NERD remains substantially more difficult, with a micro-F_1 of 0.4672, confirming that entity normalization is a major source of error in the overall system.

M-RE achieves a micro-F_1 of 0.3880, which is lower than the entity-level scores but broadly consistent with the development-set behavior. C-RE is again the most difficult setting, with a micro-F_1 of 0.1464. This confirms that concept-level RE is highly sensitive to both linking quality and relation prediction quality.

The difference between development and test performance is most pronounced for NERD and C-RE. NER micro-F_1 decreases only slightly, from 0.8238 on development to 0.8138 on test, whereas NERD micro-F_1 decreases from 0.6316 to 0.4672. This suggests that the main degradation is more closely related to URI normalization than to mention detection. A likely explanation is that exact mention–URI mappings derived from the training data generalize less effectively to unseen surface forms, abbreviations, or ambiguous biomedical mentions in the hidden test set. Since C-RE requires correct URIs for both relation arguments, these normalization errors propagate directly into concept-level relation extraction.

6.3. Comparison Between Development and Test Results

A comparison between development and official test results suggests that the general ranking of subtasks remains stable across the two splits: NER is strongest, NERD is lower, mention-level RE is more difficult, and concept-level RE is the most challenging. This consistency indicates that the overall behavior of the system is robust across evaluation settings.

At the same time, the drop from development to test is more pronounced for NERD and concept-level RE than for NER. This suggests that normalization remains one of the main bottlenecks of the system. A relation may be correct at the mention level but still fail at the concept level if one or both argument URIs are incorrect. The test results therefore reinforce the view that improvements in entity linking are likely to benefit not only NERD but also downstream concept-level relation extraction.

6.4. Macro and Micro Behavior

Across the subtasks, micro-F_1 is consistently higher than macro-F_1. This indicates that the system performs better on frequent classes than on rare ones. The effect is especially visible in the relation extraction subtasks, where the label distribution is expected to be more imbalanced than in NER.

This gap between macro and micro performance is typical of biomedical information extraction tasks. Micro-level performance reflects the system's effectiveness on the most common patterns in the data,

whereas macro-level performance is more sensitive to rare entity types and relation predicates. The observed differences therefore suggest that the submitted system is comparatively stronger on frequent biomedical categories and relation types than on long-tail cases.

6.5. System-Level Interpretation

Taken together, the results illustrate the cumulative nature of the GutBrainIE task. NER provides the basis for the entire system; NERD adds the requirement of correct normalization; mention-level RE depends on predicted entities; and concept-level RE further depends on correct URI assignment. Each stage therefore inherits the uncertainty of the previous one.

This pattern explains why the largest performance drop occurs between the entity-level and relation-level subtasks, and why concept-level RE is the most difficult setting. In particular, the gap between mention-level and concept-level RE highlights the importance of accurate entity linking in end-to-end biomedical information extraction.

7. Discussion

The results highlight a clear pattern across all four subtasks. Entity recognition is the strongest component of the submitted system, entity linking introduces a substantial additional source of difficulty, and relation extraction remains the most challenging stage, especially at the concept level. This progression is consistent with the structure of the task itself: each downstream stage depends on the correctness of the preceding one.

One of the main strengths of the system is the use of specialized components for entity recognition, entity linking, and relation extraction. This design makes the system easier to inspect, train, and debug than a fully end-to-end approach. In particular, the strong NER results provide a reliable basis for the subsequent stages, while the hybrid linking strategy offers a practical compromise between precision-oriented exact matching and more flexible similarity-based retrieval.

At the same time, the staged design makes the system highly sensitive to error propagation. Missed entity mentions cannot be linked and cannot participate in relation extraction. Incorrect boundaries or labels may reduce both linking accuracy and relation extraction precision. This cumulative effect is visible in the progression from NER to NERD and from mention-level to concept-level relation extraction. In particular, the much lower concept-level RE scores confirm that correct URI assignment remains a major bottleneck of the full system.

The results also show that design decisions made at one stage can affect the entire workflow. This is particularly clear in the threshold analysis for NER. The threshold that yields the best standalone NER result is not necessarily the one that produces the best downstream behavior, because the later stages depend directly on the set of predicted mentions. This confirms that the GutBrainIE subtasks should not be viewed as isolated prediction problems, but as connected levels of a broader biomedical information extraction setting.

A further limitation concerns the normalization stage itself. Even when entity mentions are detected correctly, assigning the correct URI remains difficult because biomedical expressions are often ambiguous, abbreviated, or lexically variable. This affects not only NERD, but also concept-level relation extraction, where an incorrect normalization can invalidate an otherwise plausible relation prediction.

As shown in Table 4, approximately 70% of URI assignments rely on exact matching, which suggests that the linking component remains strongly dependent on the coverage and generalization of mention–URI mappings observed in the training data.

Overall, the submitted system provides a complete and interpretable solution for all four subtasks, but the results also make clear that improvements in entity linking and robustness to upstream noise are likely to have the largest impact on end-to-end performance.

8. Conclusion

This paper presented the TEXA system for GutBrainIE 2026, covering all four shared-task subtasks: named entity recognition, named entity recognition and disambiguation, mention-level relation extraction, and concept-level relation extraction. The submitted system combines a fine-tuned GLiNER model for entity recognition, a hybrid linking component for URI assignment, and an ATLOP-style document-level model for relation extraction.

The results show that entity recognition is the strongest component of the system, while entity linking and relation extraction remain substantially more difficult, especially at the concept level. This pattern is consistent across both development and official test evaluation and reflects the cumulative nature of the task, where downstream predictions depend on the correctness of earlier stages. In particular, the gap between mention-level and concept-level relation extraction highlights the importance of accurate normalization in end-to-end biomedical information extraction.

Overall, the TEXA system provides a complete multi-stage solution for GutBrainIE 2026 and establishes a clear baseline for future improvements in entity linking, threshold calibration, and robustness of relation extraction under noisy predicted entities.

9. Future Work

Several directions could improve the submitted system. First, the NER component could benefit from more refined calibration strategies, such as label-specific thresholds or more selective filtering of overlapping spans. The development experiments already show that threshold choice affects not only standalone NER performance, but also the behavior of the downstream stages.

Second, the entity-linking stage remains a clear target for improvement. Future work could explore stronger use of ontology structure, improved handling of abbreviations and lexical variants, and more context-sensitive normalization strategies. Since concept-level relation extraction depends directly on URI assignment, improvements in linking are likely to benefit both NERD and the concept-level RE subtask.

Third, relation extraction could be made more robust to noisy predicted entities. This may involve stronger biomedical encoders, improved negative sampling, or training strategies better suited to long-tail relation labels. More generally, tighter interaction between entity recognition, linking, and relation extraction could help reduce the cascading errors observed in the current multi-stage design.

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.

Declaration on Generative AI

During the preparation of this work, Grammarly was used for grammar and spelling checking. ChatGPT and Gemini were used for rephrasing and sentence polishing to improve clarity and readability. ChatGPT was also used to convert parts of the manuscript into LaTeX code and to adapt the document to the CEUR-WS one-column CEUR-ART style as formatting assistance. In addition, Google Translate was used for text translation and subsequent language polishing. All content was reviewed and edited by the authors, who take full responsibility for the final manuscript.

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] 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.

[7] 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.

[8] 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.

[9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2016, pp. 260–270. URL: https://aclanthology.org/N16-1030/. doi:10.18653/v1/N16-1030.

[10] 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, Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[11] 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 (Volume 1: Long Papers), Association for Computational Linguistics, 2024, pp. 3998–4011. URL: https://aclanthology.org/2024.naacl-long.300/.

[12] 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) baw068. URL: https://doi.org/10.1093/database/baw068. doi:10.1093/database/baw068.

[13] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, BioRED: a rich biomedical relation extraction dataset, Briefings in Bioinformatics 23 (2022) bbac282. URL: https://doi.org/10.1093/bib/bbac282. doi:10.1093/bib/bbac282.

[14] National Library of Medicine, Unified Medical Language System (UMLS), https://www.nlm.nih.gov/research/umls/, 2026. Accessed 2026-05-13.

[15] National Library of Medicine, Medical Subject Headings (MeSH), https://www.ncbi.nlm.nih.gov/mesh/, 2026. Accessed 2026-05-13.

[16] European Bioinformatics Institute, Chemical Entities of Biological Interest (ChEBI), https://www.ebi.ac.uk/chebi/, 2026. Accessed 2026-05-13.

[17] European Bioinformatics Institute, Ontology Lookup Service, https://www.ebi.ac.uk/ols/, 2026. Accessed 2026-05-13.

[18] 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. URL: https://dl.acm.org/doi/10.1145/3458754. doi:10.1145/3458754.