lasigeBioTM at BioASQ25 Task GutBrainIE - Lean Large Language Models with Syntactic Features

Notebook for the BioASQ Task GutBrainIE on Gut-Brain Interplay Information Extraction at CLEF 2025

Sofia I. R. Conceição[1,*], Paulo R. C. Lopes[1] and Francisco M. Couto[1]

[1]LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, 1749–016 Lisboa, Portugal

Abstract

Capturing the relationships between gut microbiota and the brain remains an open challenge that needs to be addressed with multi-disciplinary approaches. To help fill this gap, information extraction tasks such as Named Entity Recognition (NER) and Relation Extraction (RE) can be used to identify relevant information within this domain. In the context of the BioASQ2025 Task 6—GutBrainIE: Gut-Brain Interplay Information Extraction challenge — our team developed systems that incorporated external knowledge sources, including ontologies, syntactic, and semantic features. Our objective was to evaluate whether integrating external semantic information could improve the performance of baseline large language models (LLMs). While overall performance was limited, key insights were gained regarding the utility of domain-specific pre-processing for NER and the challenges of noise introduction with other semantic features. However, in the NER task, the LLM showed improvement when the text was pre-processed using entities extracted by a specialized domain tool, BENT. This suggests that, for this task, a specialized domain-specific tool remains preferable to generic LLM approaches. The code is publicly available at https://github.com/lasigeBioTM/BioASQ25-GutBrainIE.

Keywords

constituency parsing, LLM, NLP, ontologies, treebanks

1. Introduction

Several studies have highlighted the role of gut microbiota in regulating neurotransmitters, thus affecting brain function and influencing the development of common neurological and psychiatric diseases [1], as also cognitive development, emotions and behaviors [2]. Conditions such as depression, anxiety, Parkinson's disease, bipolar disorder, autism, schizophrenia and Alzheimer's disease are associated to the gut-brain axis relationship [2, 1]. These relationships can only be addressed by a multidisciplinary approach, incorporating fields such as microbiology, neurobiology, and clinical practice.

One approach for integrating this multidisciplinary data is through the extraction of structured information from biomedical articles. This can advance the various biomedical disciplines by facilitating the acquisition of new knowledge, mainly in understanding the architecture of the gut-brain axis and identifying potential therapeutic solutions. Overall, to establish a simple pipeline that aims to extract structured information, two main tasks are necessary. First, Named Entity Recognition (NER) consists of identifying entities of interest in the text and then classifying them into defined categories [3], and, secondly, Relation Extraction (RE) aims to analyze the relations between the identified entities [4].

The BioASQ2025 Task6 -GutBrainIE: Gut-Brain interplay Information Extraction [5, 6], is a challenge that is focused on extracting structured information from biomedical abstracts and the linking of the gut microbiota and its relationships with Parkinson's disease and mental health. The first subtask of this challenge was to identify and categorize specific text spans into already established categories. This only had one subtask, subtask 6.1 - Named Entity Recognition, where PubMed abstracts with information about gut-brain interplay were provided so the systems could classify the entity mentions into one of

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

sconceicao@lasige.di.fc.ul.pt (S. I. R. Conceição); prlopes@lasige.di.fc.ul.pt (P. R. C. Lopes); fjcouto@ciencias.ulisboa.pt (F. M. Couto)

0000-0002-8891-3546 (S. I. R. Conceição); 0009-0001-8599-4432 (P. R. C. Lopes); 0000-0003-0627-1496 (F. M. Couto)

the thirteen categories. The second objective was to determine the validity of a specified relationship between two categories. This was divided into three subtasks. The first subtask, 6.2.1 - Binary Relation Extraction, involved systems utilizing the provided abstracts to identify which entities are related within a document. The second subtask, 6.2.2 -Ternary Tag-Based Relation Extraction, required systems to identify the entities in relation to one another and predict the nature of their relationship. Finally, the third subtask, 6.2.3 - Ternary Mention-Based Relation Extraction, focused on identifying the specific entities involved in a relationship and predicting the type of relationship that exists between them.

This work describes the participation of our team lasigeBioTM, at the BioASQ2025 Task6 -GutBrainIE: Gut-Brain interplay Information Extraction @CLEF 2025 for subtask 6.1 - Named Entity Recognition, subtask 6.2.2 - Ternary Tag-Based Relation Extraction and subtask 6.2.3 - Ternary Mention-Based Relation Extraction. Our goal was to assess whether the use of syntactic and semantic features could enhance the performance of the baseline large language model. A relevant consideration for lean models when applied to complex tasks such as this one. Our approach mainly focuses on improving a baseline large language model (LLM) by providing additional syntactic structure information. LLMs have had a great impact on advancing the field of natural language processing and, by consequence, in the domain of information extraction (IE). These models are capable of dealing with the various tasks encompassing IE, since they have the ability to generalize [7]. We used `Mistral-7B-Instruct-v0.3` [1] as our baseline LLM. This model is an instructive version of Mistral-7B-v0.3 [8][2], which is a generative text model fine-tuned using a variety of publicly available conversation datasets. It is designed to follow instructions and complete requests effectively. Although this is not the most recent state-of-the-art LLM, it was chosen since it is a lightweight but consistent model capable of efficient performance using 4-bit quantization while using fewer computational resources [9].

2. Related Work

Several works use semantic and syntactic features to perform Relation Extraction. The incorporation of this information helps to maintain the accuracy when dealing with distant dependencies and complex sentence structures and increases robustness to noise interference [10, 11, 12, 13, 14]. Lima et al. [10], explores the impact of semantic linguistic features in relation extraction by using a logical and relational learning approach. They use linguistic features such as lexical, syntactic, and deep semantic features. The work showed that the classifiers that used richer features had an improvement in contrast to those that only used morphosyntactic features (i.e., only morphology and syntax).

Another approach, by Jiang et al. [11], using only a rule-based network method that leverages constituency parsing, achieves comparable performance with state-of-the-art learning-based methods in the two tested benchmarks. They construct a network where nodes represent text-based entities (noun phrases) extracted from constituency parsing trees, and edges represent relationships captured by connecting verb or preposition phrases. Another work also using constituency parsing, applying it to the Named Entity Recognition task. Constituency parsing identifies and structures phrases (constituents) like noun phrases, verb phrases, and prepositional phrases [15]. The parse tree shows the hierarchical relationships between these constituents and the words within them (e.g., which words belong to which phrase, and which phrases combine to form larger phrases or clauses). This structured representation makes explicit how different parts of the sentence relate to one another [15]. The linguistic relationships between words and phrases in a sentence can be inferred from a straightforward description of constituent borders, the hierarchical structure that demonstrates how constituents combine, and the identification of heads within constituents. This structured information is relevant for tasks like relation extraction that aim to model relationships between elements in text.

In the work of Zhu et al. [12], to perform document-level relation extraction, they used dependency graphs and constituency tree to provide extra syntax information along with the information from the constituency tree to enhance representation of the dependency graph along with BERT-base models.

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[2] https://huggingface.co/mistralai/Mistral-7B-v0.3

The method utilizes dependency graphs to model interactions and identify paths between entities across the document. Constituency trees organize different words of a single sentence hierarchically and reasonably, thus aggregating better sentence-level information.

The BioGSF framework [14], is a biomedical relation extraction system that integrates shortest dependency paths between entities with entity-pair graphs. At the end, the entity-pair graph's topological information, obtained using a graph attention network, merges with SDP representation and sentence semantic properties. The semantic information contributes to the final classification decision alongside the structural information from the graph. In contrast to approaches that might rely more on direct entity information alone, the used case study indicates that models that rely on richer entity and contextual information (semantic information derived from dependency paths) perform better when dealing with entities that are under-represented in the literature.

Recently, in the work of Wang et al. [13], they used a Self-Attention-based Graph Convolutional Network to capture long-distance dependencies by filtering irrelevant nodes. Their method employed contextual semantic representations from BERT with syntactic features from the dependency syntax graph.

Many relation extraction systems still use pre-trained language models, which strive for high performance. However, with the advent of LLMs, these models have already been incorporated in some relation extraction systems since they can circumvent existing problems, such as dealing with longer contexts and zero- and few-shot being capable of generalizing and dealing with new unseen relations [16].

BioGPT [17] is one of the initial works that combines the generation abilities of GPT models with the biomedical domain. This model was fine-tuned on 15M PubMed abstract corpus and can be applied to six biomedical downstream tasks, including relation extraction. BioGPT achieved state-of-the-art results on three end-to-end relation extraction tasks.

In the work of Zhang et al. [18], they proposed a Retrieval-Augmented Semantic Parsing (RASP) framework that integrates external lexical knowledge into LLMs. Their results showed that using the RASP framework consistently improved the performance of different LLM models.

Another work that suggests that incorporating structural representations can enhance the capabilities of LLMs is the work of Zhang et al. [19]. They proposed a framework called SR-LLM that has two key components, using Abstract Meaning Representation (AMR). The first component uses AMR to reinforce the input into a more semantically rich and structured format. The second component fine-tunes the LLM on structured datasets for various natural language processing tasks. The results showed improvements in the LLM's performance on downstream tasks, revealing the potential for using structured representations to help LLMs generate novel and implicit information.

Not only in RE, but also in Named Entity Recognition, some models already can achieve comparable performance to fine-tuning and pre-trained models using reduced human and hardware expenses [20].

Since relation extraction systems seem to improve from more detailed semantic features, in this work, we incorporated syntactic features into a large language model to understand their impact.

3. Methodology

This section describes the methodologies employed by our systems. Figure 1 illustrates the overall workflow used by our team. For all systems, we utilized the Large Language Model `Mistral-7B-Instruct-v0.3`[3], which will be referred to as Mistral from now on in the paper.

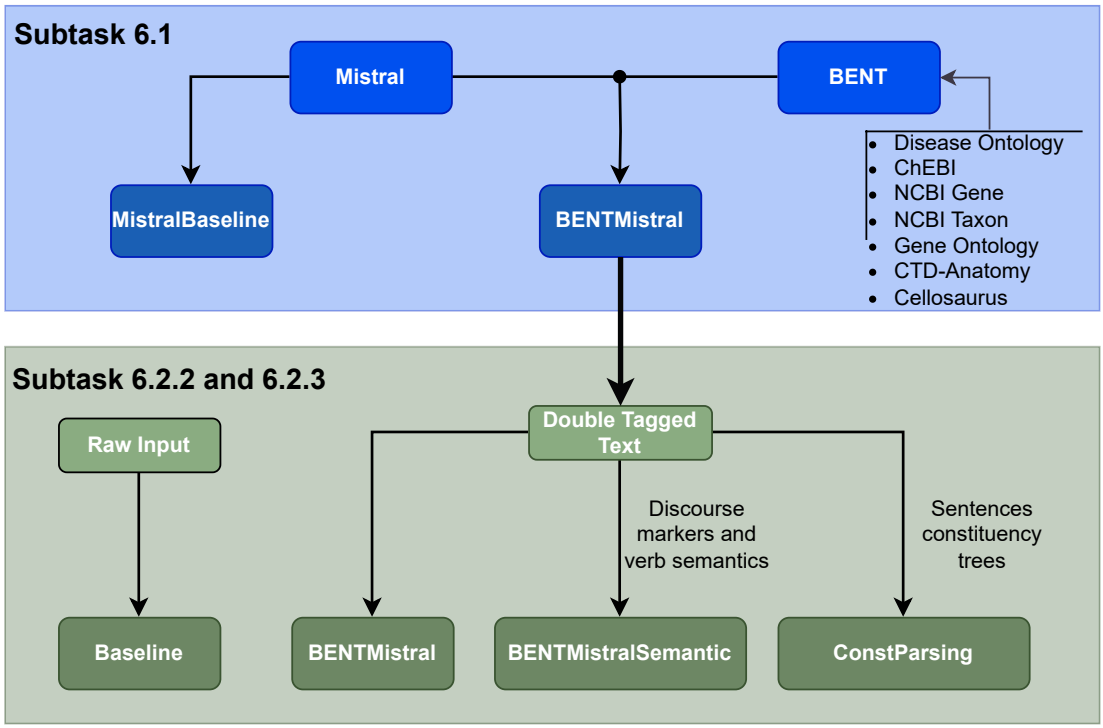


Figure 1: Overview of the used systems workflow for each subtask. **Subtask 6.1** - Pipeline for both **MistralBaseline** and **BENTMistral** where the latter receives annotated entities and the baseline doesn't. **Subtasks 6.2.2 and 6.2.3** - Pipeline showcasing the models applied and its main components. **BENTMistral** utilized the named entities recognized from **subtask 6.1**. **BENTMistralSemantic** utilized discourse markers and verb semantics. **ConstParsing** rearranges each sentence into treebanks.

3.1. Subtask 6.1 - Named Entity Recognition

We employed two systems for this subtask, MistralBaseline and BENTMistral. In **MistralBaseline** system, entities were extracted by giving the test set raw input to Mistral. To do so, we provide to the model a prompt with a simplified description of the entities provided at the challenge website[4].

As for the **BENTMistral** system, we used BENT[5] [21] which is a biomedical entity annotator that provides several pre-trained models fine-tuned on PubMedBERT [22] with respective knowledge organization systems. The extracted entities were then marked with double tags on the text. Since BENT does not cover all the entities types required for this challenge, these tagged entities worked as seed entities to assist Mistral in extracting new entities or correcting previously identified ones. The setup and used knowledge organization systems are described in Section 4 Experimental Setup.

[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[4] Entity and Relation Labels https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/#three
[5] https://github.com/lasigeBioTM/BENT

3.2. Subtask 6.2.2 - Ternary Tag-Based Relation Extraction and Subtask 6.2.3 - Ternary Mention-Based Relation Extraction

For tasks 6.2.2 and 6.2.3, we used the following four systems: Baseline, BENTMistral, BENTMistralSemantic and ConstParsing. With exception to the Baseline system, we used the named entity recognition output from BENTMistral system generated in subtask 6.1. As pre-processing, these models had the double tagged text as input. This input consisted of an external tag and an inner tag arrangement, which allowed for simultaneous entity identification and classification. For a detailed explanation, please refer to Section 4. A brief summary of the systems used in these subtasks is shown in Table 1.

Table 1
Systems Summary

System	Type of Input	Syntactic Information
Baseline	Raw input text	None
BENTMistral	Double tagged text	None
BENTMistralSemantic	Double tagged text	Discourse markers and verb semantics
ConstParsing	Double tagged text	Sentences constituency trees in Penn Treebank format

The first, the **Baseline** system was provided with the raw test set input text into Mistral to extract the relations using the defined set of relations provided at the challenge website as a guide.

The **BENTMistral** system simply uses the tagged entities that are then used for relation extraction, guided by the defined set of relations.

In the **BENTMistralSemantic** system, we used spaCy library [23][6], which is used for advanced natural language processing tasks. This system tries to extract relationships between the named entities across two adjacent sentences using discourse markers (e.g., "therefore", "nevertheless") and semantic cues from verbs (e.g., "increase", "affect"). This method tries to reduce the noise caused by complex sentences by using discourse markers and semantic clues to simplify grammar and make connections between the entities clearer. Then this information is provided to Mistral prompt. Example of the provided information:

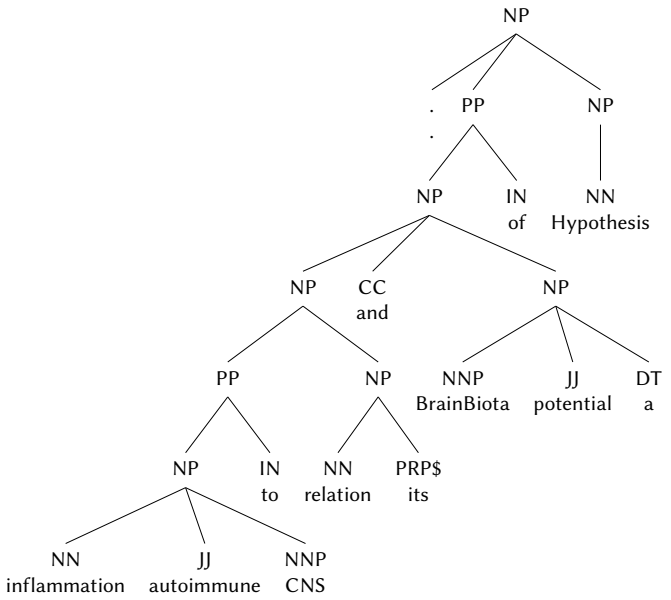
> ('Schizophrenia (DDF) → Is Linked To → neuropsychiatric disorder (DDF)' 'Context: Alteration of Gut Microbiome in Patients With Schizophrenia ' 'Indicates Links Between Bacterial Tyrosine Biosynthesis and Cognitive ' 'Dysfunction. ||| Schizophrenia (SCZ) is a heterogeneous neuropsychiatric ' 'disorder for which current treatment has insufficient efficacy and severe ' 'adverse effects.')
> ('choline (chemical) → Is Linked To → schizophrenia. (DDF)' 'Context: Furthermore, this was confirmed by an increase of choline levels, a ' 'brain imaging marker of membrane dysfunction, which is also significantly ' 'elevated in UHR subjects compared to the HR and HC groups. ||| Both gut ' 'microbiome and imaging studies of UHR subjects suggest the membrane ' 'dysfunction in the brain and hence might support the membrane hypothesis of ' 'schizophrenia.)

The final system, **ConstParsing**, involves extracting constituency trees from each sentence using spaCy with benepar (Berkeley Neural Parser)[7] [24] to extract the treebanks. Treebanks are corpora that have been annotated with syntactic structure [15]. The focus is on the tagged entities, and the results are provided in the Penn Treebank Part-of-Speech tagset format [25] to the Mistral prompt.

An example of a syntactic parse tree in Penn Treebank format, for the sentence "Hypothesis of a potential <e1>@microbiome$ BrainBiota @/microbiome$</e1> and its relation to <e2>@DDF$ CNS autoimmune inflammation @/DDF$</e2>.", is shown in Figure 2. In the example, words are attributed part-of-speech tags, such as, for example, NNP (Proper Noun, Singular) for "BrainBiota". The phrase "a potential BrainBiota and its relation to CNS autoimmune inflammation" is a noun phrase (NP). It

[6] https://spacy.io/
[7] https://spacy.io/universe/project/self-attentive-parser/

is divided into two organized sections: "a potential BrainBiota" and "its relation to CNS autoimmune inflammation." The prepositional phrase (IN) "of [...]" modifies the noun (NN) "Hypothesis" by combining these two elements, which are connected by the conjunction (CC) "and." So, the hypothesis concerns both the possibility of a BrainBiota and that it could be connected to autoimmune inflammation in the central nervous system.



"Hypothesis of a potential microbiome BrainBiota and its relation to CNS autoimmune inflammation."

Figure 2: Syntactic parse tree in Penn Treebank format. The tree shows the decomposition of a sentence into its constituents.

In syntactic constituency, the main idea is that a group of words may exist as single units (constituents) within a sentence's grammatical structure [15]. Because biomedical texts often contain complex language, this is highly advantageous. It reduces confusion caused by specialized language by simplifying grammar and revealing the relationships between biomedical concepts through the hierarchical organization of sentences. It can also help resolve ambiguity by generating alternative parses, providing additional insights into different grammatical structures, and enabling more accurate relationship extraction in situations where meanings are unclear. Since it can better handle internal and nested phrase patterns that are frequently seen in biomedical text, this method is especially useful for relation extraction. Given that the GutBrainIE task includes lengthy and syntactically complex entities, for instance, the biomedical technique entity "amplification and sequencing of the V4 region of the 16S rDNA gene", consistency parsing provides a suitable approach for addressing the challenges presented by this workshop task.

4. Experimental Setup

All code developed for this work is available at https://github.com/lasigeBioTM/BioASQ25-GutBrainIE. The repository is organized into several directories to enhance usability and clarity. The `data` directory contains all datasets, specifically within the GutBrainIE_Full_Collection_2025 directory, and is further divided into an `intermediate` directory that has pre-processed files not yet in their final format, and a `processed` directory containing data in its finalized state. Additionally, a `prompts` directory is included, which contains all prompt templates utilized throughout the project. The source code directory (`src`) contains the primary source code for the systems developed. To facilitate reproducibility, an environment file is also provided, ensuring that users can replicate the setup.

Our systems used `Mistral-7B-Instruct-v0.3` [8] as the baseline LLM model. The inference was

[8] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

done on an NVIDIA A30 GPU. The models were loaded in 4-bit precision using the BitsAndBytes library[9], with nested double quantization and Normal Float 4 as the quantization type, and the computation dtype was set to bfloat16.

As pre-processing for subtask 6.1, we used BENT by setting it up using the docker and instructions available at its GitHub page [10]. In this work we extracted the entities types of interest using the following knowledge bases available on BENT: disease ontology[11] for disease entities, Chemical Entities of Biological Interest (ChEBI)[12], an ontology for drugs and chemicals, NCBI gene[13] for genes and gene products, NCBI Taxon[14] for organisms, Gene Ontology[15] - Biological Process and Cellular Component for bioprocesses and cellular components respectively, Comparative Toxicogenomics Database-Anatomy[16] for anatomical entities and Cellosaurus[17] for cell line information.

Afterwards, we applied a double-tag annotation system to label entities in the text. This system consists of two components: an external tag and an inner tag, which provide both identification and classification of each entity.

The external tag assigns a unique identifier to every entity instance, following the format `<eN> EntitySpanText </eN>`, where `N` denotes the entity's numerical ID. For example, `<e1> BrainBiota </e1>` uniquely marks the entity "BrainBiota" as the first in the document.

The inner tag, encodes the type of the entity. It uses the format `@entityType$ EntitySpanText @/entityType$`, where `entityType` corresponds to the specific category assigned to the entity. An example would be `@microbiome$ BrainBiota @/microbiome$`, indicating that "BrainBiota" belongs to the `microbiome` category.

Finally, the combined or full tag integrates both the unique ID and the type information into a single annotation. This is represented as `<eN>@entityType$ EntitySpanText @/entityType$ </eN>`. An example of two sentences employing this tagging scheme:

> "The association between <e1>@microbiome$ oral and gut microbiota @/microbiome$</e1> in <e2>@human$ male patients @/human$</e2> with <e3>@DDF$ alcohol dependence @/DDF$</e3>."
>
> <e1>@anatomical location$ Fecal @/anatomical location$</e1> microbiota transplantation ameliorates <e2>@anatomical location$ gut @/anatomical location$</e2> microbiota imbalance and <e3>@anatomical location$ intestinal barrier @/anatomical location$</e3> damage in rats with stress-induced <e4>@DDF$ depressive-like behavior @/DDF$</e4>.

5. Results

Precision, recall, and F1-score, standard information extraction metrics, were used to assess the submitted runs. The challenge considers both the macro- and micro-average method for each. The macro-average approach deals with aggregates and examines the measure as a whole, calculating precision and recall as the arithmetic mean of individual classes. Regarding the micro, the precision and recall score is determined by calculating the scores for each individual class. The average metric is then computed by combining the contributions of all classes and weights each sample equally. The micro-average F1-score is used as the reference metric for the final leaderboard for each subtask, since it more accurately takes class imbalances into account.

[9] https://huggingface.co/docs/bitsandbytes/main/en/index.
[10] https://github.com/lasigeBioTM/BENT
[11] https://disease-ontology.org/
[12] https://www.ebi.ac.uk/chebi/
[13] https://www.ncbi.nlm.nih.gov/gene/
[14] https://www.ncbi.nlm.nih.gov/taxonomy
[15] http://geneontology.org/
[16] http://ctdbase.org/
[17] https://www.cellosaurus.org/

The performance results of the systems evaluated on the NER task (6.1) demonstrate several limitations, as it is shown in Table 2- subtask T61. The BENTMistral system achieved a Macro F1 score of 0.086 and a Micro F1 score of 0.251, which are low values and suggest the system has inconsistent performance. The MistralBaseline system had a score of 0 across all evaluation metrics. These results indicate that Mistral alone is not enough to extract the entities and benefits from the pre-processing using BENT. MistralBaseline system tended to extract longer and less precise entities, such as "neurodegenerative disorders such as Alzheimer's and Parkinson's diseases". In contrast, the BENTMistral system produced entities with more clearly defined and accurate boundaries, showing that the addition of the BENT annotations helps the system to better define the entities and improve the precision of entity recognition. However, the BENTMistral system also occasionally identified irrelevant entities, such as "decline".

For subtasks 6.2.2 and 6.2.3, the systems also showed low performance, suggesting a dependency with the output of subtask 6.1. In both subtasks, models incorporating external information-BENTMistralSemantic and ConstParsing— achieved lower scores, suggesting that integration of such information may introduce noise that generates confusion to the system. The baseline system tends to hallucinate, introducing entities that were not identified and creating relationship combinations that are not predefined. For the BENTMistral and BentMistralSemantic systems, they often fail to identify ternary tag-based relations in many examples. Similar to the baseline, they generate non-predefined combinations, such as "subject_label": "anatomical location," "predicate": "located in," and "object_label": "anatomical location." Additionally, the CostParsing system also produces incorrect relationship combinations and incorrect predicates.

Subtask 6.2.3 posed significant challenges for all systems, with negligible F1 scores. Both ConstParsing and BENTMistralSemantic scored zero in all evaluation metrics, highlighting a complete failure to generalize on this task, and once more reinforcing the idea that the additional semantic information introduced noise into the systems. For this subtask, a major issue was the coverage of the systems. The baseline system only extracted ternary mention-based relations from nine out of forty examples. Additionally, it faced problems such as incorrect combination sets and hallucination of predicates. The BENTMistralSemantic system extracted relationships from thirteen abstracts, but had numerous errors in the combination and order of relations. For instance, it incorrectly identified "gene" as a head entity when it should only be a tail entity. The BENTMistral system identified relations in twenty-one abstracts but struggled with unexpecific entities and incorrect relation sets. The CostParsing system identified relations in seventeen abstracts, but it returned the constituents as text spans instead of the original identified entities. For example, it returned "the gut" instead of just "gut," resulting in no match with the corresponding text span in the ground truth.

Table 2
Performance Metrics

Subtask	System	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
T61	BENTMistral	0.221	0.103	0.086	0.347	0.196	0.251
	MistralBaseline	0.000	0.000	0.000	0.000	0.000	0.000
T622	BENTMistral	0.120	0.066	0.077	0.457	0.066	0.115
	BENTMistralSemantic	0.011	0.010	0.010	0.308	0.016	0.031
	Baseline	0.112	0.048	0.062	0.409	0.037	0.068
	ConstParsing	0.080	0.062	0.065	0.393	0.045	0.081
T623	BENTMistral	0.027	0.005	0.008	0.093	0.005	0.010
	BENTMistralSemantic	0.000	0.000	0.000	0.000	0.000	0.000
	Baseline	0.022	0.001	0.002	0.067	0.001	0.003
	ConstParsing	0.000	0.000	0.000	0.000	0.000	0.000

6. Conclusions and Future Work

Our systems did not reach competitive results in this edition of GutBrainIE challenge. Therefore, we analyzed their main limitations for future improvement. Addressing the multiple challenges proposed by the competition required a combination of a number of techniques, which is not trivial. The main issues across our systems were inefficient performance in NER, which affected downstream subtasks, as well as hallucinations and LLM's inability to consistently follow the predefined rules, such as relation combination sets.

Entity Annotation: limited performance in Named Entity Recognition (Subtask 6.1) was one of the main factors contributing to the poor results of our systems, as it affected the downstream subtasks that relied on accurate entity annotation. The system's ability to accurately identify and extract relationships was seriously compromised by insufficient coverage and precision in entity extraction, as subsequent subtasks relied on accurate entity recognition. Additionally, our results revealed that using zero-shot NER with Mistral was insufficient for extracting relevant entities. Despite their promise, LLMs still exhibit notable limitations, such as hallucinations and inconsistent adherence to prompts [26]. The large language model performed better when supplemented with seed entities from the specialized domain tool, BENT. This suggests that, for this task, a specialized domain-specific tool remains preferable, as it outperforms a general-purpose LLM—even those with strong generalization capabilities. Moreover, BENT is highly dependent on the pretrained models, where its vocabulary is defined, which leads to its main limitation on addressing the full range of entities involved in this challenge.

Syntax and Semantic Features While incorporating semantic information can help system performance by providing richer contextual cues, it may also introduce noise or complexity. This was observed in the BENT-Mistral-Semantic and ConstParsing systems, where the incorporation of these features seems to degrade performance, likely due to overwhelming information injection, leading to hallucinations. Striking the right balance between leveraging semantic enrichment and managing the possibility of noise remains a significant challenge.

The work of Jin et al.[27], also encountered challenges when trying to provide additional semantic information to an LLM. In their study, they used Abstract Meaning Representations (AMR) to capture distilled semantic information about the text, including entities, events, and the relationships between them. They provided the AMR graph representation along with the text as input to the LLM. The main finding was that when the AMR information was included, the model's performance actually degraded compared to using just the text alone. Although the AMR information did help improve the model's performance in some specific cases. This suggests that using structured semantic representations could have potential for LLM understanding in the future, even though the current approaches are still not adapted for this.

One potential solution is the use of better integration strategies, such as prompt tuning, which allows the model to receive semantic information in a controlled and interpretable manner. By carefully designing prompts, it is possible to guide the model's attention toward relevant semantic features while minimizing confusion.

6.1. Future Work

Future work for the Named Entity Recognition task should focus specifically on rectifying the gap among entity types and improving recall. One potential direction is to fine-tune existing models using high-quality datasets, such as those provided specifically for this challenge. Additionally, combining these datasets with other publicly available corpora could further enhance model performance and generalization. Another path worth exploring is the hybrid approaches that combine already well-established BERT-based NER models with large language models. By integrating the strong contextual embeddings and proven effectiveness of BERT with the broad knowledge and generative capabilities of LLMs, it may be possible to achieve improved accuracy and robustness in entity recognition.

Future work on addressing the challenges of incorporating semantic information could explore methods such as adaptive prompting and semantic filtering, aimed at optimizing how additional context

is interpreted during inference. These approaches could help systems selectively attend to relevant semantic cues while reducing the risk of noise or overfitting. In parallel, prompt engineering offers a practical strategy for guiding models to effectively leverage semantic knowledge. This could involve providing well-designed examples that illustrate the intended features or integrating informative semantic tags that highlight key elements. Such techniques can improve the interpretability of prompts and ensure the model processes semantic input in a controlled and meaningful way.

Acknowledgments

The authors would like to express their gratitude to Dr. Maria Fernandes for her invaluable assistance and for her thoughtful review of the manuscript. This work was supported by FCT (Fundacão para a Ciência e a Tecnologia) through funding of the PhD Scholarships with ref. UI/BD/153730/2022 and DOI identifier https://doi.org/10.54499/UI/BD/153730/2022 attributed to SIRC, and the LASIGE Research Unit, ref. UID/00408/2025.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 and LanguageTool in order to: Grammar, spelling check and paraphrase. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. Gurow, D. C. Joshi, J. Gwasikoti, N. Joshi, Gut microbial control of neurotransmitters and their relation to neurological disorders: A comprehensive review, Hormone and Metabolic Research (2025).

[2] L. Mitrea, S.-A. Nemeș, K. Szabó, B.-E. Teleky, D. C. Vodnar, Guts imbalance imbalances the brain: A review of gut microbiota association with neurological and psychiatric disorders, Frontiers in Medicine 9 (2022). URL: https://api.semanticscholar.org/CorpusID:247845361.

[3] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics 33 (2017) i37–i48. doi:10.1093/bioinformatics/btx228.

[4] J. Liu, H. Ren, M. Wu, J. Wang, H. jin Kim, Multiple relations extraction among multiple entities in unstructured text, Soft Computing 22 (2018) 4295–4305. doi:10.1007/s00500-017-2852-8.

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

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

[7] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, E. Chen, Large language models for generative information extraction: A survey, Frontiers Comput. Sci. 18 (2023) 186357. URL: https://api.semanticscholar.org/CorpusID:266690657.

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

[9] M. M. Hosen, A lora-based approach to fine-tuning llms for educational guidance in resource-constrained settings, ArXiv abs/2504.15610 (2025). URL: https://api.semanticscholar.org/CorpusID:277993896.

[10] R. Lima, B. Espinasse, F. Freitas, The impact of semantic linguistic features in relation extraction: A logical relational learning approach, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019, pp. 648–654.

[11] M. Jiang, J. Diesner, A constituency parsing tree based method for relation extraction from abstracts of scholarly publications, in: Proceedings of the thirteenth workshop on graph-based methods for natural language processing (textgraphs-13), 2019, pp. 186–191.

[12] X. Zhu, Z. Kang, B. Hui, FCDS: Fusing constituency and dependency syntax into document-level relation extraction, ArXiv abs/2403.01886 (2024). URL: https://api.semanticscholar.org/CorpusID:268247791.

[13] L. Wang, F. Wu, X. Liu, J. Cao, M. Ma, Z. Qu, Relationship extraction between entities with long distance dependencies and noise based on semantic and syntactic features, Scientific Reports 15 (2025) 1–13.

[14] Y. Yang, Z. Zheng, Y. Xu, H. Wei, W. Yan, BioGSF: a graph-driven semantic feature integration framework for biomedical relation extraction, Briefings in Bioinformatics 26 (2025) bbaf025.

[15] D. Jurafsky, J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd ed., 2025. URL: https://web.stanford.edu/~jurafsky/slp3/, online manuscript released January 12, 2025.

[16] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers, ACM Comput. Surv. 56 (2024). URL: https://doi.org/10.1145/3674501. doi:10.1145/3674501.

[17] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, BioGPT: generative pre-trained transformer for biomedical text generation and mining, Briefings in bioinformatics 23 (2022) bbac409.

[18] X. Zhang, Q. Meng, J. Bos, Retrieval-augmented semantic parsing: Using large language models to improve generalization, 2024. URL: https://arxiv.org/abs/2412.10207. arXiv:2412.10207.

[19] J. Zhang, T. Wang, H. Wu, Z. Huang, Y. Wu, D. Chen, L. Song, Y. Zhang, G. Rao, K. Yu, SR-LLM: Rethinking the structured representation in large language model, 2025. URL: https://arxiv.org/abs/2502.14352. arXiv:2502.14352.

[20] L. Zhao, L. Kang, Q. Guo, Zero-shot document-level biomedical relation extraction via scenario-based prompt design in two-stage with llm, 2025. URL: https://api.semanticscholar.org/CorpusID:278310363.

[21] P. Ruas, F. M. Couto, Nilinker: Attention-based approach to nil entity linking, Journal of Biomedical Informatics 132 (2022) 104137. URL: https://www.sciencedirect.com/science/article/pii/S1532046422001526. doi:https://doi.org/10.1016/j.jbi.2022.104137.

[22] 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, 2020. arXiv:arXiv:2007.15779.

[23] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020). doi:10.5281/zenodo.1212303.

[24] N. Kitaev, D. Klein, Constituency parsing with a self-attentive encoder, arXiv preprint arXiv:1805.01052 (2018).

[25] M. Marcus, B. Santorini, M. A. Marcinkiewicz, Building a large annotated corpus of English: The Penn Treebank, Computational linguistics 19 (1993) 313–330.

[26] J. Zhang, M. Wibert, H. Zhou, X. Peng, Q. Chen, V. K. Keloth, Y. Hu, R. Zhang, H. Xu, K. Raja, A study of biomedical relation extraction using GPT models, AMIA Summits on Translational Science Proceedings 2024 (2024) 391.

[27] Z. Jin, Y. Chen, F. Gonzalez, J. Liu, J. Zhang, J. Michael, B. Schölkopf, M. T. Diab, Analyzing the role of semantic representations in the era of large language models, in: North American Chapter of the Association for Computational Linguistics, 2024. URL: https://api.semanticscholar.org/CorpusID:

269502049.