Understanding Gut-Brain Interplay in Scientific Literature: A Hybrid Approach from Classification to Generative LLM Reasoning

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

Chaeeun Lee[1,†], Simona E. Doneva[2,†], Maria Juliana Rodriguez-Cubillos[1,†], Elisa Castagnari[1,†], Antoine D. Lain[3,†], Joram M. Posma[3,*] and T. Ian Simpson[1,*]

[1]School of Informatics, University of Edinburgh, 10 Crichton Street, EH8 9AB, Edinburgh, UK
[2]Center for Reproducible Science, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland
[3]Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion, and Reproduction, Faculty of Medicine, Imperial College London, London W12 0NN, United Kingdom

Abstract

In this work we present our approach to Task 6 GutBrainIE in the CLEF2025 BioASQ, in which we develop Natural Language Processing (NLP) systems to extract structured information from biomedical literature related to the gut microbiome and its connection to neurological disease and mental health. The task consists of a multi-class Named Entity Recognition (NER) subtask (6.1) and three Relation Extraction (RE) subtasks(6.2.1 Binary Relation Extraction, 6.2.2 Ternary Tag-Based Relation Extraction and 6.2.3 Ternary Mention-Based Relation Extraction). Our system adopts a two-stage pipeline. First, we addressed the NER as a token classification task with encoder-only BERT-based models. To address the complexity of the multi-class NER including significant class imbalance, we explored a range of training and post-processing strategies, such as span-based ensemble of predictions from models trained on different subsets of labels. For RE, we investigated both an encoder-based classification approach and a generative approach where we fine-tuned a large language model (LLM) on generated reasoning traces. Our systems achieved competitive performance for both NER and RE subtasks, with our best RE system ranking 3rd for mention-level RE on the official leaderboard and our best NER system ranking 4th, demonstrating the effectiveness of combining structured classification with generative reasoning in biomedical information extraction. In addition, we provide qualitative insights into the challenges of multi-class NER for domain-specific corpus and complementary strengths and limitations of encoder-based and generative approaches for RE. Our findings underscore the value of combining structured classification with interpretability-oriented generative reasoning in information extraction pipelines.

Keywords

Biomedical Natural Language Processing, Named Entity Recognition, Relation Extraction, Information Retrieval, Gut Microbiota, Gut-Brain

1. Introduction

The gut–brain axis is a complex biological system enabling bidirectional signalling between the brain and the gut. A growing body of studies have highlighted the potential role of the gut microbiome in neurological and psychiatric conditions [1, 2, 3]. Much of this information is only accessible as unstructured text in scientific journal articles [4] limiting our ability to use these data to deepen our understanding of the gut-brain axis. Natural Language Processing (NLP)-based information extraction (IE) methods offer a promising way to harness this information for research use through reliable IE tools that are capable of identifying and organising relevant biomedical knowledge[5]. The GutBrainIE

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
chaeeun.lee@ed.ac.uk (C. Lee); simona.doneva@uzh.ch (S. E. Doneva); juliana.rodriguez@ed.ac.uk (M. J. Rodriguez-Cubillos); e.castagnari@ed.ac.uk (E. Castagnari); a.lain@imperial.ac.uk (A. D. Lain); jmp111@ic.ac.uk (J. M. Posma); ian.simpson@ed.ac.uk (T. I. Simpson)

task [6] in BioASQ Laboratory [7] provides an annotated dataset for developing such tools, with a focus on extracting structured information from biomedical abstracts related to gut microbiota and their roles in psychiatic and neurological disease.

The GutBrainIE task [6] is composed of four subtasks designed to evaluate biomedical IE tools for gut-brain interplay. The first, Subtask 6.1: Named Entity Recognition (NER), requires systems to identify and classify entity mentions into one of thirteen predefined biomedical categories. The remaining three subtasks focus on Relation Extraction (RE) at varying levels of detail. Subtask 6.2.1: Binary RE asks participants to detect whether a relation exists between two identified entities within a document, without specifying the type of relation. Subtask 6.2.2: Ternary Tag-Based RE extends this by requiring systems to predict not only the presence of a relation but also its type from a predefined set of relation predicates. Finally, Subtask 6.2.3: Ternary Mention-Based RE requires participants to identify the exact entity mentions involved in a relation and classify the relation type between them. These four subtasks were created using a fine-grained annotation schema, which defines 13 distinct entity categories and 25 relation types. This level of granularity supports richer biomedical understanding and enables applications such as knowledge-graph construction [8] and evidence synthesis.

Biomedical IE is challenging because many biomedical categories overlap and entity meanings are often ambiguous. For instance, 'bacteria' typically refers to bacterial taxonomy, whereas 'microbiome' covers the broader microbial community and thus requires contextual interpretation. Likewise, distinguishing among 'chemical', 'dietary supplement', and 'drug' mentions is difficult, since the same compound may belong to different categories depending on use context or regulatory status. Gene mentions also require nuanced understanding of the context. In the literature, a 'gene' can denote the gene itself, its protein product, an enzyme, or even an entire biological pathway, again depending on context. Such contextual variability complicates the design of a fine-grained annotation schema that generalises, especially when combined with relation-extraction tasks that must identify both interaction type and directionality.

This paper presents the work of our team (ICUE) for the GutBrainIE task [6] in BioASQ Laboratory [7]. We discuss the dataset analysis and distribution, our methodology for each subtask, experimental setup, results, limitations and directions for future work.

2. Related Work

With much of biomedical knowledge often only accessible through unstructured scientific literature, biomedical information extraction continues to be a rapidly evolving area of research. Pioneering resources such as the GENIA corpus [9] laid the foundation two decades ago, and the field has since progressed from feature-engineering approaches to transformer-based models (e.g., BioBERT [10], SciBERT [11]) and, most recently, to LLMs capable of few-shot reasoning. Over the years, many datasets and systems have been developed for both NER and RE tasks. In NER, previous research has primarily focused on identifying entities such as diseases, drugs/chemicals, genes/proteins, and species [12, 13, 14, 15, 16, 17, 18, 19]. For RE, the focus has typically been on identifying the presence or absence of a relationship between entities such as genes and diseases or proteins and chemicals [20, 21, 22, 23]. However, while these well-known and widely used datasets cover established biomedical categories, they do not capture the finer distinctions introduced in this challenge, such as separating chemicals from drugs or dietary supplements, or distinguishing bacteria from the broader microbiome. Additionally, entity types such as biomedical technique and statistical technique are rarely annotated in existing corpora.

In terms of methods, the advent of the Transformer architecture [24] spurred the development of biomedical language models based on fine-tuned transformers. BioBERT [10], a domain-specific variant of BERT pre-trained on PubMed and PMC articles, achieved state-of-the-art performance on various biomedical NER benchmarks, including F1 scores of 89.71% on NCBI Disease [12] and 87.15% on BC5CDR Disease [14]. PubMedBERT [25], trained exclusively on PubMed abstracts, offers improved performance in specific biomedical categories, particularly gene (F1 score of 79.10% on JNLPBA [19])

and disease mentions (F1 score of 85.62% on BC5-disease [14]), and outperformed BioBERT for RE with a F1 score of 83.96% on GAD [20] and 77.24% on ChemProt [23].

There has been notable parallel progress in RE methods. Following early successes with transformer-based encoder-only models for sequence classification, generative sequence-to-sequence approaches using encoder-decoder architectures have also shown strong potential. REBEL [26] is an autoregressive sequence-to-sequence model for RE. It frames RE as a generation task, translating raw text into structured relation triplets. Built on a BART-based Transformer architecture, REBEL uses a linearisation approach with special tokens to represent triplets, enabling efficient autoregressive decoding. The model has been evaluated on standard RE benchmarks such as TACRED [27], DocRED [28], and CONLL04 [29], where it achieves competitive or state-of-the-art results, as well as better generalisation in low-resource settings. While REBEL does not have a distinct NER module or requires it as a preliminary step, its end-to-end generation process directly identifies and outputs the entity spans and their types as part of generating the complete relation triplet. SciSpacy [30], a spaCy extension with pre-trained NER models for biomedical texts, has proven effective in lightweight applications but lacks the contextual reasoning capabilities of transformer-based approaches.

LLMs have shown promise in few-shot biomedical classification tasks [31, 32, 33]. However, these models face limitations when applied to NER. A key challenge arises from the divergence between the parameter knowledge of LLMs, learned during pretraining, and the specific annotation guidelines of the biomedical corpus of interest. This misalignment often results in trade-offs between precision and recall, especially in context-dependent cases. Entity types like gene or microbiome are particularly difficult, as their meanings are heavily dependent on the surrounding context. Despite these challenges, LLMs have demonstrated remarkable capabilities in natural language understanding and generation tasks, driven by Transformer-based architectures that have scaled to hundreds of billions of parameters. Pre-trained on vast text corpora using self-supervised objectives, these models are typically fine-tuned on task-specific data using strategies like prompting or supervised training. One notable advancement has been the incorporation of chain-of-thought (CoT) prompting, which enables LLMs to perform complex, multi-step reasoning tasks by introducing explicit reasoning traces into few-shot examples [34]. This has substantially improved both accuracy and interpretability in many tasks.

Recent research also focuses on scaling reasoning at inference time, where models generate explicit reasoning tokens interspersed with normal output, allowing for more interpretable and structured chains of thought [35, 36]. Building on these advances, knowledge distillation has emerged as a promising method for transferring the reasoning capabilities of larger models to smaller ones. In this setup, a teacher model generates both final answers and intermediate rationales, which are then used to train a student model to replicate both outputs and reasoning traces [37]. Incorporating synthetic reasoning data during distillation has been shown to significantly boost performance. Distilled models trained in this way can match the zero-shot performance of larger models on specific reasoning tasks. However, due to their smaller parameter size, these models are still constrained when handling knowledge-intensive problems, and some information loss is inevitable during the distillation process. While these approaches have shown limited effectiveness for NER, they appear more promising for RE tasks, where reasoning plays a greater role than static entity knowledge.

3. Dataset

The dataset used in this study consists of titles and abstracts from PubMed articles, with a thematic focus on the gut microbiota and its connection to Parkinson's disease and mental health. The data is divided into three primary subsets: a training set of 1,567 articles, a development set of 40 articles, and a test set of 40 articles.

Training data is stratified by annotation quality into four levels. The highest quality, Platinum-standard annotations, are expert-curated and externally reviewed by biomedical professionals. Gold-standard annotations are also expert-curated but without external review. Silver-standard annotations were created by trained student annotators under supervision and are further subgrouped based on anno-

tator consistency, where documents in `StudentA` where annotated by annotators with more consistent performance compared to `StudentB`. Bronze-standard annotations were generated automatically using GLiNER [38] for NER and ATLOP [39] for RE.

Each article is annotated with entity mentions and, where applicable, relations between entity pairs. Entities are labeled according to a predefined schema of 13 biomedical categories. The test set is drawn from the gold and platinum standard data and includes only titles and abstracts. It is constructed to provide broad coverage of the entity and relation types relevant to the task.

Table 1 summarises the entity labels, definitions, and their frequency in the training data.

Table 1
Entity labels, concise definitions, and their frequency in the training data.

Entity label	Definition (abridged)	Count	Unique
Disease, Disorder, or Finding (DDF)	Observations, test results, and other pathology-related concepts.	17 267	4 338
Chemical	Substances with constant composition; includes metabolites and neurotransmitters.	5 639	2 387
Microbiome	Entire microbial habitat (organisms, genomes, environment).	4 853	471
Human	Members of *Homo sapiens*.	3 757	1 004
Bacteria	Unicellular prokaryotes of the domain *Bacteria*.	3 175	1 204
Anatomical Location	Named body parts or regions.	2 409	470
Dietary Supplement	Orally taken nutrients (macro / micro).	1 844	629
Biomedical Technique	Methods applying biological or physiological principles in medicine.	1 772	1 195
Animal	Multicellular, non-human organisms capable of movement.	1 674	649
Drug	Substances that alter physiology, used therapeutically or recreationally.	1 251	498
Statistical Technique	Methods to calculate, analyse, or present statistical data.	758	538
Gene	Hereditary units occupying defined chromosomal loci.	692	452
Food	Substances consumed for nutrition.	434	233

Statistics on relation frequencies and entity-type pairings are illustrated in Figures 1 and 2.

4. Methodology

4.1. NER

We describe here the methods and configurations used to develop our systems for the NER task. We utilised encoder-only Transformer models for token classification, mainly focusing on domain-specific pretrained models with various preprocessing, fine-tuning, and postprocessing steps. Each system was trained as a token-level sequence tagger using the IOB2 labeling scheme [40], with model configurations varying in backbone architecture, class subset coverage, and ensemble composition. Data preparation involved token alignment, label assignment, and filtering based on entity presence. Postprocessing recovered entity spans from token-level predictions, resolving subword splits and validating offset mappings. The systems were trained and evaluated using a unified framework built on the HuggingFace Transformers library [41]. To improve clarity and reproducibility, we defined a set of standardised codes for backbone models, class coverage, ensemble strategies, and post-processing steps (Table 2). These codes are then used in Table 3 to describe each of the top 5 system based on test set performance.

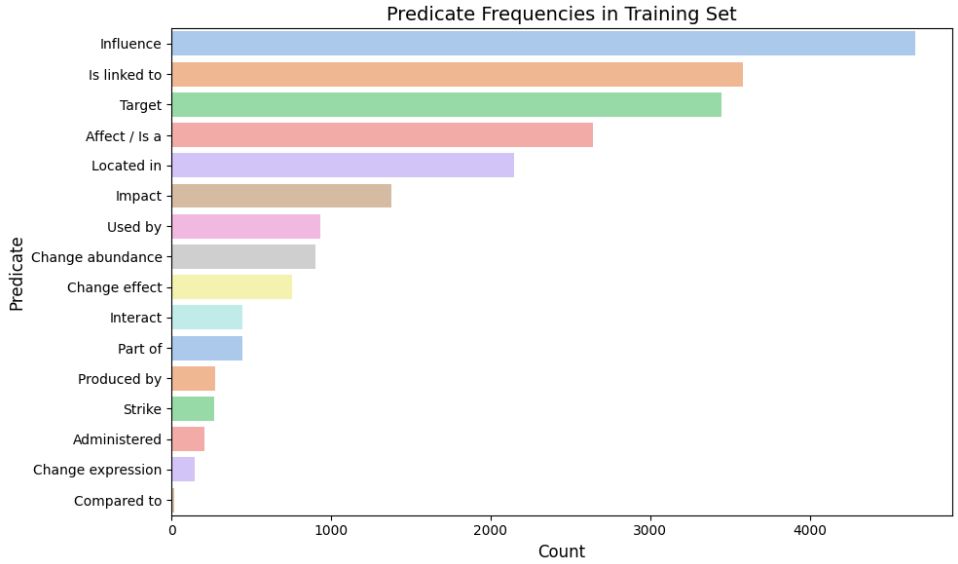

Figure 1: Predicate frequencies in the training set.

Figure 2: Distribution of head vs. tail entity types in annotated relations

4.1.1. Preprocessing

To prepare the data for model training, article texts were first merged with their corresponding annotations using PubMed IDs. Tokenization was then performed using either the `bert-base-cased` or `bert-base-uncased` tokenizer, depending on the model configuration.

During tokenisation, words may be split into subwords. These subwords were re-aligned to form original tokens, and entity labels were propagated across subwords according to the standard IOB2 tagging scheme. Tokens were labeled as `B-`, `I-`, or `O` based on whether they marked the beginning, continuation, or absence of an entity span. Each document was processed independently, grouped by PubMed ID and sentence location. For each token, we checked for overlaps with annotated entity spans

Table 2
Legend of codes used in our NER system submissions, including backbone models, entity class subsets, ensemble strategies, and post-processing steps.

Pre-trained model		**Class-coverage**	
GL	GLiNER (organiser baseline)	ALL	all 13 entity classes
PMBL	`BiomedNLP-BiomedBERT-large-uncased-abstract`	high-count	DDF, microbiome, bacteria, anatomical location, animal, human, chemical, gene, drug
PMBB	`BiomedNLP-BiomedBERT-base-uncased-abstract`	Food	food only
BLB	`BioLinkBERT-large`	DDF	DDF only
		Microbiome	microbiome only
Ensemble method		**Post-processing**	
single	no ensemble (single model)	PT	add PubTator gene annotations
LS	longest-span rule		
US	union of spans rule		
MINK	retain span if $\geq k$ models agree		

Table 3
Description of NER methods and configurations for top five systems based on test set metrics. Each system is described using the standardised codes from Table 2

Run ID	**Backbone**	**Classes**	**Ensemble**	**Post-proc**
ensemble5	PMBL, PMBB, BLB	ALL, DDF, Microbiome, high-count, Food	MINK (K=10)	–
ensemble11	PMBL, PMBB, BLB, GL	ALL, DDF, Microbiome, high-count	MINK (K=10)	PT
ensemble10	PMBL, PMBB, BLB, GL	ALL, DDF, Microbiome, high-count	MINK (K=7)	PT
ensemble4	PMBL, PMBB, BLB	ALL, DDF, Microbiome, high-count, Food	MINK (K=4)	–
ensemble9	PMBL, PMBB, BLB	ALL, DDF, Microbiome, high-count	MINK (K=4)	PT

and assigned corresponding labels. The system supported multi-label settings and handled overlapping annotations by prioritising the longest match.

Label coverage was defined by a configurable label set. By default, models were trained on the full set of 13 entity types, but we also trained models on subsets (e.g., high-frequency labels or task-specific classes like food, DDF, or microbiome). All IOB2-converted data were exported to JSON format. For sequences exceeding 512 tokens, the input was split into chunks to conform to model input size limits of 512 with overlap size of 12.

4.2. Model Architecture

We experimented with four model families for NER. The organiser baseline system was GLiNER,[1] a lightweight span classification model designed to generalise to unseen entity types via transfer learning. It does not require predefining a label set, making it suitable for few-shot or open-domain scenarios.

In our systems, mainly domain-specific BERT variants trained on biomedical corpora were utilised: `BiomedNLP-BiomedBERT-base-uncased-abstract` and `BiomedNLP-BiomedBERT-large-uncased-abstract` [25], as well as `BioLinkBERT-large` [42].

Each model was trained either on all 13 entity types or on a focused subset of entity types. This design enabled experimentation with label subset selection to test performance trade-offs.

[1] https://github.com/kakaobrain/gliner

4.2.1. Inference and Postprocessing

At inference time, models produced IOB2-encoded label predictions. Subword tokens were merged back into full words, and offset alignment was re-computed to extract contiguous entity spans from token-level labels. Predictions were filtered using span-level validation checks to ensure consistency with input text and IOB2 rules. For ensemble systems, we applied span-based majority voting. Specifically in Table 3, MINK denotes a voting strategy, where a span was retained if predicted as an entity by at least K models.

Given the pronounced class imbalance across the 13 entity labels, we experimented with targeted post-processing techniques tailored to individual classes. We observed that the *food* entities in the development dataset were commonly recognized, but misclassified as *dietary supplement*. This may be attributed to the overlapping contextual usage of these concepts, combined with the dominance of annotations for the *dietary supplement* class. To address this, we used WordNet, a lexical database of English, to extract hierarchically structured sets of food- and beverage-related terms [43]. Specifically, we retrieved all hyponyms of the synsets "food.n.02" and "beverage.n.01". The use of those subsets was motivated by their similarity to the guideline definition of food: "a group of solid, semi-solid, and liquid substances which are consumed by humans and animals". This closely matches the WordNet definitions: "any solid substance (as opposed to liquid) that is used as a source of nourishment" ("food.n.02") and "any liquid suitable for drinking" ("beverage.n.01").

Extracted terms were normalized by lowercasing, removing underscores, and applying lemmatization to reduce morphological variance. These normalized term sets were then used to relabel entities initially labeled as "dietary supplement": if a phrase or any of its constituent words matched the food or drink term sets, the label was overwritten as "food". In addition, a small manually curated keyword list, based on the annotation guidelines for food, was included to capture relevant edge cases not covered by WordNet[2]. Example relabelings included: *"dairy products"* $\rightarrow$ *"food"* and *"unpasteurised milk"* $\rightarrow$ *"food"*.

4.3. RE

We approached RE tasks using two main strategies: (1) sequence classification with a BERT-based encoder-only models, and (2) supervised fine-tuning of LLM using generated reasoning traces. In addition, we implemented two baseline methods: one based on rule-based dataset statistics, and another using a generative encoder-decoder model for RE. Below, we describe the baselines, our two primary approaches, and various pre- and post-processing methods that we explored.

We approached all three RE subtasks (6.2.1–6.2.3) using the same underlying methods. Predictions were generated uniformly and task-specific outputs were derived by including the appropriate fields as required by each subtask. Model selection was based on performance on the development set for subtask 6.2.3 ternary mention-based RE.

4.3.1. RE Baseline 1: Rule-based Method

We implemented a simple baseline for RE using relation frequency statistics and co-occurrence patterns observed in the training data. This method relies on identifying commonly annotated *(subject, object)* label pairs and using these to predict relations in unseen documents, with optional filtering based on distance and likelihood.

To support this, we first computed relation statistics from the training corpus. For each annotated relation, we extracted the subject and object labels, the predicate, and the character distance between the subject's end and the object's start index. We tracked the frequency of each *(subject, object)* pair, the number of unique annotators who labeled it, and predicate frequencies per pair (excluding annotations by distant supervision). Additionally, we computed distance-based metrics from the character distances between the subject's end and the object's start index, including mean, median, minimum, maximum, and robust percentiles (5th and 95th) for each pair.

[2]The manual keyword list included: *diets*, *diet*, *product*, *products*, *food*, and *foods*.

We also calculated entity label co-occurrence frequencies across all documents, independent of whether a relation was annotated. These co-occurrence statistics allowed us to define a *relation likelihood* as the ratio of annotated frequency to total co-occurrence frequency for each pair. For example, the pair *<DDF, animal>* co-occurred 5,285 times in the dataset, but the relation "target" was annotated only 547 times, resulting in a relation likelihood of 0.10.

For binary relation prediction, we used a filtering-based approach grounded in the training statistics described above. A *(subject, object)* pair was considered valid if it met all of the following criteria: (1) it was annotated by at least one non-distant annotator, (2) the total number of predicate annotations for the pair met a minimum frequency threshold, and (3) its relation likelihood exceeded a predefined cutoff (default: 0.01). We further refined the candidate entity pairs by comparing their character-level distance against the learned statistics from the training data. A candidate pair was retained only if it satisfied two conditions: (1) the direction and magnitude of the distance had to be consistent with the average distance observed in training (e.g., subjects typically preceding objects), and (2) the distance had to lie within the 5th to 95th percentile range of training distances for that label pair.

To extend binary predictions to full relation triples, we assigned predicates to each filtered *(subject, object)* pair using frequency statistics from the training data. For each pair, we retrieved the distribution of observed predicates from the training set. If `predict_all` was enabled, all known predicates for the pair were predicted, simulating an upper-bound scenario. Otherwise, we selected the single most frequently annotated predicate as the predicted relation.

4.3.2. RE Baseline 2: REBEL

We employed the REBEL framework using the publicly available implementation from https://github.com/Babelscape/rebel. To adapt the model to the domain-specific dataset, we fine-tuned *Babelscape/rebel-large* with the custom entity and relation types from the challenge. This process involved creating new configuration files for data and training parameters, as well as implementing a dataset loader to handle our specific data format. We also modified core REBEL source files to support the new schema, including `pl_modules.py`, `train.py`, and `test.py`, to integrate the custom relation and entity definitions.

```
<triplet> gut microbiota <microbiome> central nervous system
<anatomical_location> located in
<microbiome> depression <ddf> is linked to
<microbiome> depressive disorder <ddf> is linked to

<triplet> neurotransmitters <chemical> gut microbiota <microbiome> impact

<triplet> gut peptides <chemical> gut microbiota <microbiome> produced by

<triplet> gut microbiota <microbiome> mental health <ddf> is linked to

...
```

Figure 3: Example of a linearized triplet generated for REBEL fine-tuning. Full text is abridged here.

4.3.3. Sequence Classification with Encoder-only Transformers

As our main approach to RE, we formulated RE as a binary sequence classification task. Given a sentence containing two tagged entities, the model predicts whether a given relation exists between them. To this end, we fine-tuned a BERT-based Transformer model for sequence classification, which consists of a pre-trained encoder followed by a classification head. For each input sequence, the model leverages the contextual representation of the special [CLS] token, which serves as a summary embedding of the entire input sequence. This representation is passed through a classification head to predict a binary

label indicating the presence or absence of the candidate relation. We fine-tuned the model end-to-end using binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ (Equation 1), where N is the number of training examples in a batch, y_i is the ground-truth binary label for the i-th example, and $\hat{y}_i$ is the predicted probability that the relation exist. We evaluated the model on the development set using standard classification metrics including precision, recall, and F1 score.

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\Big], \tag{1}$$

To construct input instances for the binary classification model we first generated all legal entity pairs from the ground truth NER labels, based on the annotation guidelines (Section 4.1 Relation Labels), which defines valid combinations of head entity type, tail entity type, and relation predicate. Using ground-truth NER annotations, we extracted all entity pairs that matched one of these legal entity type combinations. For each such pair, we created a classification instance by inserting tags around the head and tail entities in the full sentence. Since there were cases where multiple valid relation types could exist between two entity types, we added a prefix sentence to explicitly indicate which relation was being classified (Figure 4). This enables the model to disambiguate between different predicates applicable to the same entity pairs. Each resulting sentence was treated as a binary classification example, with the label indicating whether the specific relation was present or not.

> Is "influence" the correct relation between the subject entity **<chemical> proinflammatory cytokines </chemical>** and the object entity **<DDF> depression </DDF>** in the following text?
> **TEXT:**
>
> …
>
> Moreover **<DDF> depression </DDF>** can be induced by administration of **<chemical> proinflammatory cytokines </chemical>**, including IL-2 or IFN-α.
>
> …

Figure 4: Example RE classication instance for encoder-only models. Full text is abridged here.

To prepare input sequences, we experimented with three different strategies. In the first setting, we included only the sentences that explicitly contained both the head and tail entities, ensuring the input was focused on the mention span where the relation might be expressed. To account for cases where the entities were mentioned in separate sentences but still shared a contextual link, we also tried a broader context window by selecting the sentences containing the head and tail entities along with all sentences between them. Lastly, we also experimented with using full-text including both the title and full abstract as input and tagging head and tail entities where they appear. Due to the nature of the task setup where binary classification instances were created for all legal subject-object entity pairs, the resulting dataset was highly imbalanced, with far fewer positive instances (i.e., ground-truth annotated relations) compared to the large number of negative pairs. To address this, we generated multiple balanced versions of the training dataset by randomly sampling different subsets of negative instances, while keeping the full set of positive instances fixed across all splits. This ensured that each model variant saw the complete set of annotated relations while being exposed to diverse, representative samples of negatives. We trained separate models on each of these balanced training splits, evaluated them individually on a shared development set, and ultimately ensembled their predictions to improve robustness and mitigate the effects of label imbalance and sampling variance.

4.4. LLM-based RE via Supervised Fine-Tuning

Beyond encoder-based binary classifiers, we explored the use of LLMs for RE via supervised fine-tuning (SFT) with reasoning traces. Inspired by recent work on interpretable reasoning with LLMs [35, 44], we

framed relation classification as a two-option multiple-choice question answering (QA) task, where the model must not only classify a relation but also justify it through an intermediate reasoning trace.

To build the training corpus, we used the more capable `DeepSeek-R1` [45] as a teacher model to generate reasoning traces via API for binary-choice QA instances. Each prompt contained: (1) an instruction clarifying the relation to be classified, (2) the full document text with the subject and object entities highlighted using custom tags (in the same format as our encoder models), and (3) two candidate options, one being affirmative of the given relation between tagged entities and the other indicating an absence of such relation between the entities. We present an example input provided to both the teacher and student LLMs in Table 4.

The reasoning model outputs a chain-of-thought (CoT) explanation followed by a final answer label, either "A" or "B". From the generated dataset, we curated a fine-tuning corpus by keeping only those traces whose final answer matched the gold label. This filtered set was then used to fine-tune a smaller LLM with token-level cross-entropy loss $\mathcal{L}_{\text{CE}}$ (Equation 2), where N is the batch size, T_i is the length of the i-th sequence, $x_{i,t}$ is the ground-truth token at position t, and p_θ is the model's predicted probability for the given token. Each reasoning trace and its final answer were concatenated and treated as a single target sequence.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T_i}\log p_\theta\big(x_{i,t} \mid x_{i,<t}\big), \tag{2}$$

We used `deepseek-ai/DeepSeek-R1-Distill-Qwen-7B` as our student model. Fine-tuning was performed using HuggingFace `TRL` with DeepSpeed ZeRO-3 for efficient multi-GPU training. We evaluated the resulting model in two settings:

(1) **Standalone prediction** – the model directly outputs reasoning and a relation decision.

(2) **Post-processing verifier** – used as a verifier after the BERT-based classification.

The latter approach was motivated by the observation that the BERT-based models achieved high recall but comparatively lower precision and F1 scores; the LLM was used to verify or filter predicted positive relations, thus acting effectively as a re-ranker.

5. Experimental Setup

All experiments were conducted using NVIDIA H100 GPUs. The encoder-based NER and RE classifiers were trained on a single H100 GPU with 40GB of memory. For both NER and RE tasks, we fine-tuned pretrained BERT models with a batch size of 16 and a learning rate of 2×10^{-5}. Models were trained for up to 15 epochs and the best checkpoint was saved based on the development set overall F1 score. No hyperparameter-tuning was done.

For supervised fine-tuning (SFT) of LLMs using reasoning traces, we utilised four H100 GPUs with 80GB of memory each. We used a batch size of 2 per device, 1 epoch, a cosine learning rate schedule (min ratio 0.1), learning rate of 5×10^{-5}, and `bfloat16` precision. Flash Attention 2 and gradient checkpointing were enabled for memory efficiency.

6. Results

We submitted 25 system runs for each four subtasks. Given the number of configurations explored, we present here the top 5 systems for the NER task and top 3 systems for each RE subtask, based on their official test set performance. These represent the most competitive and informative combinations of model architecture, training strategy, and post-processing, providing insight into which method and configurations were most effective across subtasks.

Table 4
Example input to teacher and student LLM for supervised fine-tuning. The reasoning trace and final answer presented is from the teacher LLM.

PMID 36244970: *Depression and fatigue in active IBD from a microbiome perspective-a Bayesian approach to faecal metagenomics*

LLM Input

Depression and fatigue in active IBD from a microbiome perspective-a Bayesian approach to faecal metagenomics. Extraintestinal symptoms are common in inflammatory bowel diseases (IBD) and include depression and fatigue.
...
Based on taxonomic and functional metagenomic profiles of **<microbiome>faecal gut microbiota</microbiome>**, we used Bayesian statistics to investigate the associative networks and triangle motifs between bacterial genera, functional modules and symptom severity of self-reported fatigue and depression. Associations with moderate to strong evidence were found for 3 genera (Odoribacter, Anaerotruncus and **<bacteria>Alistipes</bacteria>**) and 3 functional modules (pectin, glycosaminoglycan and central carbohydrate metabolism) with regard to depression and for 4 genera
...

A) <bacteria>Alistipes</bacteria> is part of <microbiome>faecal gut microbiota</microbiome>.
B) <bacteria>Alistipes</bacteria> is not part of <microbiome>faecal gut microbiota</microbiome>.

Teacher LLM Reasoning

Okay, let's try to figure out whether Alistipes is part of the faecal gut microbiota based on the provided study.
...
clearly lists Alistipes as one of the genera identified in the faecal samples. Since the study is about the microbiome in fecal samples of IBD patients, and Alistipes is listed among the genera found there, that supports option A.
...
Final Answer. A

6.1. Subtask 6.1 NER

All top five systems were constructed by ensembling predictions from multiple NER models. Our best-performing system (Run ID `ensemble5`) was an ensemble of predictions from 11 different single-model runs. The ensembling process was span-based, meaning we aggregated predictions at the level of entity spans rather than tokens. In this system, we applied the `MINK` ensemble strategy, where a predicted span was retained as an entity mention only if at least K individual models (in this case, K=10) predicted the exact same start and end offsets for that span as well as the same entity type. This voting-based filtering helped reduce spurious predictions while preserving spans that were consistently identified across multiple models and improved the consistency of the final annotations especially in cases involving overlapping or ambiguous spans.

The ensemble included models trained on different subsets of entity classes (e.g., ALL, DDF, Microbiome) as well as those from different backbone models. Configurations for each top 5 submissions can be found on Table 3 and Table 5 shows development and test set metrics. The results indicate that span-based ensembling contributes meaningfully to NER performance, particularly when aggregating predictions from diverse models. All top five submissions employed ensemble methods.

Another observation was made that training models on subsets of entity classes, rather than the full label set, can still be effective when such models are integrated within an ensemble. Several top systems included models specialised in high-frequency or semantically similar labels (e.g., DDF, Microbiome, or

Food), which, while not necessarily strong on their own, contributed to performance increase when ensembled with broader models. These results suggest that selective training on entity type subsets, when paired with robust ensembling, can be a practical strategy in complex multi-label NER tasks.

Table 5
Metrics for the top five NER submissions based on test set performance. We also include the organiser baseline system and the overall best system with the highest micro-F1 score on the official leaderboard. We highlight in bold the overall best metrics and underline the best metrics among our top five submissions based on test set micro-F1. Since the official leaderboard ranks systems only by micro-F1, we cannot determine the overall best systems for the other metrics.

Task ID	Run ID	Dev micro-F1	Test Set				Notes
			micro-F1	micro-P	micro-R	macro-F1	
6.1	ensemble5	0.8401	<u>0.8331</u>	<u>0.8369</u>	0.8249	<u>0.7546</u>	-
6.1	ensemble11	0.8424	0.8264	0.8274	0.8254	0.7470	–
6.1	ensemble10	0.8102	0.8204	0.7966	0.8456	0.7350	-
6.1	ensemble4	0.7944	0.8100	0.7629	<u>0.8634</u>	0.7170	-
6.1	ensemble9	0.8360	0.8056	0.7582	0.8593	0.7108	-
6.1	Organiser Baseline	-	0.7927	0.7639	0.8238	0.7047	-
6.1	leaderboard-best	-	**0.8408**	0.8384	0.8432	0.7613	-

6.2. RE Subtasks 6.2.1 - 6.2.3

We report here results on the RE subtasks from both the rule-based and generative baseline methods, BERT-based binary classification models, and post-processing with LLM-based reasoning.

6.2.1. Baseline 1: Rule-based Method

The rule-based system, which relies on corpus statistics, performs strongly in terms of recall, achieving 0.90 in both binary and ternary tagging tasks and 0.68 in the more challenging mention-level setting (Table 6). This suggests that the system is effective at identifying a wide range of relation instances by leveraging frequently co-occurring patterns observed in the training data. However, this broad matching strategy comes at the expense of precision, which remains below 0.50 across all tasks, indicating that many extracted relations are incorrect. The performance especially deteriorates sharply in the more fine-grained mention-level task, where the F1 score falls to just 0.18. These results highlight a key limitation of rule-based, corpus-driven methods: while they can achieve high coverage, they often lack the specificity needed for accurate RE.

6.2.2. Baseline 2: REBEL

The REBEL model demonstrates consistent performance across the three RE tasks, with F1 scores of 0.59 for binary tagging, 0.57 for ternary tagging, and 0.35 for the more challenging mention-level extraction.

In the binary relation-extraction setting, REBEL produced several recurring false positives. The pair (microbiome → anatomical location) was falsly predicted most often. Two other pairs—(bacteria → DDF) and (microbiome → DDF)—each was predicted wrongly six times. In the tag-based ternary setting, the most frequent false positive was the complete relation (microbiome located in anatomical location), predicted seven times. Two further ternary relations were over-predicted six times each: (microbiome is linked to DDF) and (microbiome located in animal).

Turning to false negatives, the relation (DDF is a DDF) was missed fourteen times, making it the single most common omission. The model also failed to recover (DDF affects DDF) ten times and overlooked two other DDF-centric relations—(DDF strikes anatomical location) and (DDF targets human)—on six occasions each. Overall, these patterns indicate that REBEL tends to over-generate microbiome-related

links while struggling to capture intra-DDF interactions and DDF relations to human or anatomical entities.

Table 6
Development set metrics for RE baseline methods, underline values show the highest metric reported.

Task	6.2.1 Binary			6.2.2 Tag-based			6.2.3 Mention-based		
	micro-P	micro-R	micro-F1	micro-P	micro-R	micro-F1	micro-P	micro-R	micro-F1
Rule-based	0.47	0.90	0.62	0.44	0.90	0.59	0.10	0.68	0.18
REBEL	0.64	0.56	0.59	0.63	0.53	0.57	0.41	0.31	0.35

6.2.3. BERT-based sequence classification

The BERT-based binary classifiers trained on balanced datasets achieved consistently high recall, often well above organiser baseline and leaderboard best systems (Table 7), accurately identifying most of the true positive relations. However, due to the large number of negative pairs precision was relatively lower. Across multiple training splits (with fixed positives and randomly sampled negatives), the performance was stable, and all models performed comparably on the development set. To improve robustness and mitigate sampling noise, we ensembled predictions from these independently trained models, which led to a modest but consistent increase in F1 score. We experimented with `BiomedNLP-BiomedBERT-large-uncased-abstract`, `BiomedNLP-BiomedBERT-base-uncased-abstract`, and `BioLinkBERT-large`, and all top three systems for each subtasks were based on `BioLinkBERT-large` as the backbone model. We report development and test set metrics for each systems in Table 7.

Table 7
Metrics for the top three RE submissions based on test set performance. We also include organiser baseline system and the overall best system with the highest micro-F1 score on the official leaderboard. We bold overall best metrics and underline best metrics among our top three submissions based on test set micro-F1. Since the official leaderboard ranks systems only by micro-F1, we cannot determine the overall best systems for the other metrics.

Task ID	Run ID	Dev micro-F1	Test Set				Notes
			micro-F1	micro-P	micro-R	macro-F1	
6.2.1	run17	0.5232	0.5476	0.3894	0.9221	0.4751	LLM verifier
6.2.1	run18	0.5232	0.5476	0.3894	0.9221	0.4751	-
6.2.1	run15	0.5232	0.5333	0.3731	0.9351	0.4667	-
6.2.1	Organiser Baseline	–	0.5947	0.7584	0.4892	0.3864	ATLOP
6.2.1	leaderboard-best	–	**0.6864**	0.6304	0.7532	0.5386	-
6.2.2	run22	0.7477	0.6093	0.4974	0.7860	0.4880	-
6.2.2	run15	0.6874	0.6052	0.5241	0.7160	0.4639	-
6.2.2	run17	0.7417	0.6021	0.5262	0.7037	0.4562	LLM verifier
6.2.2	Organiser Baseline	–	0.5751	0.7533	0.4650	0.3745	ATLOP
6.2.2	leaderboard-best	–	**0.6866**	0.6280	0.7572	0.5184	-
6.2.3	run23	0.5936	0.3651	0.2858	0.5054	0.2825	LLM verifier
6.2.3	run15	0.5836	0.3593	0.2886	0.4759	0.2821	-
6.2.3	run17	0.5048	0.3560	0.2967	0.4450	0.2878	LLM verifier
6.2.3	Organiser Baseline	–	0.3288	0.4986	0.2453	0.2123	ATLOP
6.2.3	leaderboard-best	–	**0.4635**	0.4215	0.5147	0.3497	-

6.2.4. LLM Supervised fine-tuning

In Table 7, systems marked with the note "LLM verifier" refer to configurations where a SFT-trained student LLM was used to verify predictions made by the base `BioLinkBERT-large` classifier. This two-stage setup was motivated by the observation that the classifier achieved high recall on the development set, and the LLM was used to improve precision by filtering false positives. For the mention-based RE task, our best-performing system submission was based on the LLM-based verifier, which unexpectedly achieved the highest recall among all submissions despite our initial assumption that it might trade recall for precision.

On the development set, systems with an LLM verifier exhibited improved precision but a drop in recall. We hypothesize that this trade-off is due in part to the model's limited ability to capture global document-level annotation patterns. Specifically, the LLM was trained on individual (subject, predicate, object) triples along with full-text in isolation, without access to the surrounding annotation distribution. In cases where the same textual mention of an entity pair occurred in different positions across the document, the model often predicted the negative option, since there were many more negative instances with that entity pair text span in the training set.

To compensate for these limitations, we applied the fine-tuned LLM as a post-hoc verifier on the outputs of the BERT-based classifier. In this setting, the LLM was used to re-evaluate positive predictions from the previous step with encoder-only models, with the goal of reducing false positives. This postprocessing approach led to a modest increase in development and test set metrics, as shown in Table 7.

7. Conclusion

In this work, we explored multiple approaches to biomedical information extraction, addressing both NER and RE tasks within a unified framework. For NER, we framed the task as a multilabel token classification problem and experimented with a diverse set of strategies, including training on individual subsets of labels and combining predictions via span-level ensembling. This enabled more balanced handling of under-represented classes and improved overall entity coverage. Our analysis revealed that while larger backbone models helped with complex entity types, ensemble strategies were especially effective in reducing false positives and improving consistency across classes.

For RE, we combined classical encoder-based classification with LLM's reasoning capabilities. We formulated RE as a binary classification task using BERT-based sequence classifiers, incorporating explicit entity markers in the input. To improve robustness, we addressed class imbalance by generating multiple training splits, each containing all positive instances and a different subset of sampled negative examples. Independent models were trained on each split, and their predictions were combined through ensembling to enhance recall and reduce variance, although the overall improvement was modest. To harness the reasoning capabilities of LLMs, we generated natural language reasoning traces using a more capable teacher LLM. From these generations, we selected only those that concluded with the correct label to construct a supervised fine-tuning dataset. A smaller student LLM was then fine-tuned using token-level cross-entropy loss. Given that our classification-based system achieved high recall but lower precision, we used the fine-tuned LLM as a second-step verifier to better filter out false positives and improve overall precision.

Together, our findings highlight the strength of encoder-based classification for NER, and the benefit of combining classical classification system with LLM-based reasoning for RE. This hybrid approach was particularly beneficial in cases where consistency with annotation patterns across the whole corpus was important, a setting where document-level LLM reasoning alone often fell short, but classical classification models were effective when used as a first step.

Funding

C.L. was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. M.J.R.C. was supported by EASTBIO - East of Scotland Biosciences consortium, UKRI doctoral training program. E.C. was supported by the United Kingdom Research and Innovation (grant EP/Y030869/1), UKRI AI Centre for Doctoral Training in Biomedical Innovation at the University of Edinburgh. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. J.M.P. and A.D.L. are supported by the CoDiet project. The CoDiet project is funded by the European Union under Horizon Europe grant number 101084642 and supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 101084642].

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check, paraphrase and reword, and improve writing style. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. F. Cryan, T. G. Dinan, Mind-altering microorganisms: the impact of the gut microbiota on brain and behavior, Nature Reviews Neuroscience 13 (2012) 701–712. doi:10.1038/nrn3346.

[2] J. A. Foster, K.-A. McVey Neufeld, Gut-brain axis: how the microbiome influences anxiety and depression, Trends in Neurosciences 36 (2013) 305–312. doi:10.1016/j.tins.2013.01.005.

[3] P. Tiwari, R. Dwivedi, M. Bansal, M. Tripathi, R. Dada, Role of gut microbiota in neurological disorders and its therapeutic significance, J Clin Med 12 (2023) 1650. doi:10.3390/jcm12041650, pMID: 36836185; PMCID: PMC9965848.

[4] Y. Zang, X. Lai, C. Li, D. Ding, Y. Wang, Y. Zhu, The role of gut microbiota in various neurological and psychiatric disorders—an evidence mapping based on quantified evidence, Mediators of Inflammation 2023 (2023). URL: https://onlinelibrary.wiley.com/doi/10.1155/2023/5127157.

[5] J. S. Loh, W. Q. Mak, L. Tan, C. X. Ng, H. H. Chan, S. H. Yeow, J. B. Foo, Y. S. Ong, C. W. How, K. Y. Khaw, Microbiota–gut–brain axis and its therapeutic applications in neurodegenerative diseases, Signal Transduction and Targeted Therapy 9 (2024). URL: https://www.nature.com/articles/s41392-024-01743-1.

[6] 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.

[7] 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, F. N. Maria Di Nunzio, Giorgio, 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: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[8] D. Choi, A. D. Lain, J. M. Posma, M. Kozdoba, B. Perets, S. Mannor, From medical literature to predictive features: An evidence-based knowledge graph approach, in: Proceedings of the LMRL Workshop at the International Conference on Learning Representations (ICLR), 2025. URL: https://openreview.net/forum?id=qCSNi1BRPc.

[9] J.-D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, Genia corpus - a semantically annotated corpus for bio-textmining, Bioinformatics 19 Suppl 1 (2003) i180–2. URL: https://academic.oup.com/bioinformatics/article/19/suppl_1/i180/227927.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234 – 1240. URL: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[11] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: Conference on Empirical Methods in Natural Language Processing, 2019. URL: https://aclanthology.org/D19-1371/.

[12] R. I. Dogan, R. Leaman, Z. Lu, Ncbi disease corpus: A resource for disease name recognition and concept normalization, Journal of biomedical informatics 47 (2014) 1–10. URL: https://www.sciencedirect.com/science/article/pii/S1532046413001974?via%3Dihub.

[13] Ö. Uzuner, B. R. South, S. Shen, S. L. Duvall, 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text, Journal of the American Medical Informatics Association : JAMIA 18 5 (2011) 552–6. URL: https://academic.oup.com/jamia/article/18/5/552/830538.

[14] 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: The Journal of Biological Databases and Curation 2016 (2016). URL: https://academic.oup.com/database/article/doi/10.1093/database/baw068/2630414.

[15] M. Krallinger, O. Rabal, F. Leitner, M. Vázquez, D. Salgado, Z. Lu, R. Leaman, Y. Lu, D.-H. Ji, D. M. Lowe, R. A. Sayle, R. T. Batista-Navarro, R. Rak, T. Huber, T. Rocktäschel, S. Matos, D. Campos, B. Tang, H. Xu, T. Munkhdalai, K. H. Ryu, S. V. Ramanan, P. S. Nathan, S. Žitnik, M. Bajec, L. Weber, M. Irmer, S. A. Akhondi, J. A. Kors, S. Xu, X. An, U. K. Sikdar, A. Ekbal, M. Yoshioka, T. M. Dieb, M. Choi, K. M. Verspoor, M. Khabsa, C. L. Giles, H. Liu, K. E. Ravikumar, A. Lamurias, F. M. Couto, H.-J. Dai, R. T.-H. Tsai, C. Ata, T. Can, A. Usie, R. Alves, I. Segura-Bedmar, P. Martínez, J. Oyarzábal, A. Valencia, The chemdner corpus of chemicals and drugs and its annotation principles, Journal of Cheminformatics 7 (2015) S2 – S2. URL: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-7-S1-S2.

[16] L. L. Smith, L. K. Tanabe, R. Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. Baumgartner, L. E. Hunter, B. Carpenter, R. T.-H. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. W. Adriaans, C. Blaschke, R. Torres, M. L. Neves, P. Nakov, A. Divoli, M. Maña-López, J. Mata, W. J. Wilbur, Overview of biocreative ii gene mention recognition, Genome Biology 9 (2008) S2 – S2. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-s2-s2.

[17] M. Gerner, G. Nenadic, C. M. Bergman, Linnaeus: A species name identification system for biomedical literature, BMC Bioinformatics 11 (2010) 85 – 85. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-85.

[18] E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, L. J. Jensen, The species and organisms resources for fast and accurate identification of taxonomic names in text, PLoS ONE 8 (2013). URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0065390.

[19] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, N. Collier, Introduction to the bio-entity recognition task at jnlpba, in: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications, Citeseer, 2004, pp. 70–75.

[20] K. G. Becker, K. C. Barnes, T. J. Bright, S. A. Wang, The genetic association database, Nature Genetics 36 (2004) 431–432. URL: https://www.nature.com/articles/ng0504-431.

[21] E. M. van Mulligen, A. Fourrier-Réglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifirò, J. A. Kors, L. I. Furlong, The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships, Journal of biomedical informatics 45 5 (2012) 879–84. URL: https://www.sciencedirect.com/science/article/pii/S1532046412000573.

[22] M. Krallinger, O. Rabal, A. Miranda-Escalada, A. Valencia, Drugprot corpus: Biocreative vii track 1 - text mining drug and chemical-protein interactions, 2021. URL: https://academic.oup.com/

database/article/doi/10.1093/database/baad080/7453369.

[23] R. O. L. A. Krallinger, M., Overview of the biocreative vi chemical-protein interaction track, Proceedings of the BioCreative VI Workshop, 141-146 (2017). URL: https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-5/.

[24] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Neural Information Processing Systems, 2017. URL: https://arxiv.org/abs/1706.03762.

[25] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2020. arXiv:arXiv:2007.15779.

[26] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204/. doi:10.18653/v1/2021.findings-emnlp.204.

[27] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. D. Manning, Position-aware attention and supervised data improve slot filling, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 35–45. URL: https://aclanthology.org/D17-1004/. doi:10.18653/v1/D17-1004.

[28] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, DocRED: A large-scale document-level relation extraction dataset, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 764–777. URL: https://aclanthology.org/P19-1074/. doi:10.18653/v1/P19-1074.

[29] D. Roth, W.-t. Yih, A linear programming formulation for global inference in natural language tasks, in: Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004, Association for Computational Linguistics, Boston, Massachusetts, USA, 2004, pp. 1–8. URL: https://aclanthology.org/W04-2401/.

[30] M. Neumann, D. King, I. Beltagy, B. W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, ArXiv abs/1902.07669 (2019). URL: https://aclanthology.org/W19-5034/.

[31] Q. Chen, H. Sun, H. Liu, Y. Jiang, T. Ran, X. Jin, X. Xiao, Z. Lin, H. Chen, Z. Niu, An extensive benchmark study on biomedical text generation and mining with chatgpt, Bioinformatics 39 (2023). URL: https://academic.oup.com/bioinformatics/article/39/9/btad557/7264174.

[32] I. Jahan, M. T. R. Laskar, C. Peng, J. X. Huang, A comprehensive evaluation of large language models on benchmark biomedical text processing tasks, Computers in biology and medicine 171 (2023) 108189. URL: https://www.sciencedirect.com/science/article/pii/S0010482524002737.

[33] Q. Chen, J. Du, Y. Hu, V. K. Keloth, X. Peng, K. Raja, Q. Xie, A. Gilson, M. B. Singer, R. A. Adelman, R. Zhang, Z. Lu, H. Xu, A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations, Nature Communications (2025). URL: https://www.nature.com/articles/s41467-025-56989-2.

[34] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, ArXiv abs/2201.11903 (2022). URL: https://arxiv.org/abs/2201.11903.

[35] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, ArXiv abs/2210.03629 (2022). URL: https://arxiv.org/abs/2210.03629.

[36] C. Snell, J. Lee, K. Xu, A. Kumar, Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL: https://arxiv.org/abs/2408.03314. arXiv:2408.03314.

[37] X. Chen, Z. Sun, W. Guo, M. Zhang, Y. Chen, Y. Sun, H. Su, Y. Pan, D. Klakow, W. Li, X. Shen, Unveiling the key factors for distilling chain-of-thought reasoning, 2025. URL: https://arxiv.org/abs/2502.18001. arXiv:2502.18001.

[38] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé, E. Bernard, Nuner: Entity recognition encoder

pre-training via llm-annotated data, 2024. arXiv:2402.15343.

[39] W. Zhou, K. Huang, T. Ma, J. Huang, Document-level relation extraction with adaptive thresholding and localized context pooling, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[40] L. A. Ramshaw, M. P. Marcus, Text chunking using transformation-based learning, ArXiv cmp-lg/9505040 (1995). URL: https://arxiv.org/abs/cmp-lg/9505040.

[41] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[42] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: Annual Meeting of the Association for Computational Linguistics, 2022. URL: https://arxiv.org/abs/2203.15827.

[43] G. A. Miller, Wordnet: a lexical database for english, Commun. ACM 38 (1995) 39–41. URL: https://doi.org/10.1145/219717.219748. doi:10.1145/219717.219748.

[44] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, ArXiv abs/2302.04761 (2023). URL: https://arxiv.org/abs/2302.04761.

[45] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J.-M. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B.-L. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D.-L. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S.-K. Wu, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W.-X. Yu, W. Zhang, W. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X.-C. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y.-J. Zou, Y. He, Y. Xiong, Y.-W. Luo, Y. mei You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. guo Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z.-A. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, Z. Zhang, Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, ArXiv abs/2501.12948 (2025). URL: https://arxiv.org/abs/2501.12948.