# Relationship Extraction using Retrieval Augmented Generation for biomedical Dataset

Anonymous ACL submission

## Abstract

With the increasing number of structured and unstructured data, obtaining reliable information effectively has become crucial. In the biomedical domain, extracting information from the scientific papers is crucial in order to stay up-to-date with accurate information, given the increased pace by which new research studies are published. This work focuses on identifying relationships between entities that are extracted from the abstracts and titles of biomedical research papers. In this work, we developed a Retrieval Augmented Generation (RAG) based system to automatically identify relations between biomedical entities. We evaluate multiple open source Large Language Models (LLMs) and the number of examples (shots) required to improve the LLM's results. We evaluate our methods using precision, recall and F-1 scores and compare our approach to traditional deep learning methods using DeBERTa with a Convolutional Neural Network (CNN). Our results indicate that Qwen models using the RAG approach with 10-shot examples achieved the highest macro F1 score compared to the baseline and other LLMs under the same setting. At 35 shots, Qwen reasoning and Qwen non-reasoning model performed best, exhibiting the fewest hallucinated labels and maintaining high macro F1 scores.

## 1 Introduction

With the rapid development of technology, the Internet has become part of daily life for everyone. This makes it difficult to stay up to date with useful information all the time. Information Extraction (IE) is an important task in Natural Language Processing. It usually focuses on a specific domain or task, looking for information that is relevant to a user's interests (Hobbs and Riloff, 2010). The need for IE from large-scale scientific resources continues to increase over time (Brokman et al., 2025).

Relationship Extraction (RE) is a subpart of Information Extraction (IE) with many use cases across different domains like automatically extracting relations between entities from biomedical text. RE aims to identify semantic relations between entities in text and build structured knowledge bases (Wang et al., 2022). RE from biomedical corpora is important research to biomedical experts. One common application of biomedical RE is extracting Drug-drug interaction from texts. Drug-drug interaction (DDI) is defined as a change in the effects of one drug by the presence of another drug (Baxter, 2010). It is important to extract knowledge about DDIs from pharmaceutical papers comprehensively to understand the information and have new research around it. It can also be useful for people who need summarization and quick understanding of abstracts when they lack deep domain expertise, helping them save time and avoid being overwhelmed by the technical jargon. Reviewing abstracts from recent and previous biomedical papers is a tedious task, and getting insights from those papers is time-consuming.

Earlier for RE, fine tuned models were fundamentally limited by their pre-trained knowledge scope and often hallucinate, as they are not easily updatable with new biomedical findings. In contrast, large language models (LLMs) such as GPT (Mann et al., 2020) have demonstrated strong performance through prompt engineering. There are a number of prompting techniques such as few-shot, chain-of-thought, and generated knowledge prompting that can enhance a LLM's ability to answer questions. Prompting can serve as a useful tool in a variety of general domains. However, in some cases for specialized domains, such as the biomedical domain, their performance falls short, and finetuning these models is often infeasible due to computational restraints or time. This sets the stage for Retrieval-Augmented

Generation (RAG) approaches. These approaches incorporate domain specific knowledge from a vector store into prompts fed into an LLM.

In this project, we explore efficient approaches for RE using state-of-the-art methods (Lewis et al., 2020). For each relationship extracted from the text between different entities, the system computes the embedding similarity between the query entities and those in the training dataset. Then it retrieves the top-$k$ most similar candidates using vector search (Olasunkanmi et al., 2025). The results demonstrate that the RAG approach effectively mitigates hallucination and incorrect reasoning in relation extraction tasks (Chen et al., 2024).

To take the limitations of LLMs into account, Retrieval-Augmented Generation (RAG) integrates external information, dynamically retrievable resources with LLM generation, increasing factual grounding, and enhancing robustness against noisy, vague or counterfactual evidence-key concerns in biomedicine.

The diverse applications of Retrieval Augmented Generation (RAG) opens new perspectives, giving dynamic knowledge access and superior robustness against hallucinations and has recent applications in biomedical information extraction challenges like GutBrainIE. Inspired by the potential of RAG for biomedical relationship extraction, the research goals of this work are to address the following questions:

- How is the performance of Retrieval-Augmented Generation for RE (RAG4RE) pipeline when using different available LLMs compared to traditional neural RE models (such as DeBERTa + CNN ) across the Biomedical dataset?

- How many prompt examples (*shots*) are necessary for a given model to achieve optimal results in relationship extraction for biomedical dataset? At what point does increasing the number of shots result in saturated performance for the best performing LLMs?

## 2 Related Work

This section reviews prior work on biomedical Relation Extraction and Retrieval Augmented Generation systems that our approach builds upon. RAG utilizes both parametric and non-parametric memory to improve the performance of LLMs over NLP tasks. Parametric memory refers to the knowledge stored internally within the model's learned weights, while non-parametric memory is the information extracted from the external documents(Lewis et al., 2020). In traditional RAG, retrieval is performed directly from the initial user query, without any refinement or classification along the way (Lewis et al., 2020).

Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), replaced the traditional sparse technique (e.g.,BM25) (Robertson et al., 2009) with improved dual-encoder embeddings for both queries and passages. In open domain Question Answering, DPR achieved a top-20 passage recall of 85.4% on Natural Question(NQ), which is compared to BM25 with 64.3% and 21.2% absolute gain. When it is integrated with RAG pipeline, DPR systems reached EM (Exact Match) scores exceeding 42% and F1 improvements of 9-11% across major QA datasets, measuring dense retrieval as high-quality selection.

The queries and external knowledge base are converted into vector space for efficient retrieval. This is known as the embedding step. Then, these vectors get stored into a vector database, allowing for fast similarity search(Lewis et al., 2020) which is known as indexing. The embedding process has evolved significantly from early approaches. While BERT-based embeddings (Devlin et al., 2019) initially dominated, recent work shows that task-specific fine-tuning of embedders substantially improves retrieval quality (Ram et al., 2021). The indexing strategy matters as well: dense passage retrieval (DPR) (Chen, 2024) excels at semantic similarity, but could miss lexically important terms, motivating hybrid approaches that combine dense and sparse representations (Glass et al., 2022a). A critical, but often overlooked aspect is the chunking strategy: how documents are segmented can affect both retrieval precision and context preservation, with optimal chunk sizes varying by domain and query type (Li et al., 2025).

For relationship extraction, frameworks like RAG4RE (Efeoglu and Paschke, 2024) used prompt and data augmentation by allowing neighboring examples and sentences with context similarity. This improved the model's F1 score significantly in comparison to simple queries, the report shows up to 94.6% F1 on TACREV, 93% on Re-TACRED, and robust zero-shot performance while using a fine-tuned LLM i.e Mistral-7B. Augmented queries have helped retrieval recall improve

by 30%, demonstrating that quality of prompt and prompt expansion helps RAG systems extract information effectively, even from vague or multiple-hop questions. RAG approach with Multi-Source and Retrieval Augmentation(Lee et al., 2024) Hybrid retrieval pipelines, both HYBGRAG and Re2G(Glass et al., 2022b), combine different retrieval approaches-dense(FAISS), sparse (BM25), and sometimes graoh-based or multiple query which can be used as source within ensemble combination and reranking modules. Both of these approaches shows how integrating multiple different context sources, allow for critical selection and lead up to better precision i.e. 88% recall i.e. 65.7% and 34% improvements on multi-hop benchmarks in comparison to traditional RAG. These hybrid models allow to retrieve from more diverse context, which is more relevant and have trustworthy evidence.

## 3 Dataset

The dataset consists of 2,127 train samples with 18 relation labels and 20,227 test samples with 15 relation labels. It focuses on biomedical titles and abstracts related to the gut microbiota and its effects on mental health, following Martinelli et al. (2025).The dataset focuses on biomedical titles and abstracts related to the gut microbiota and its effects on mental health (Martinelli et al., 2025).

The dataset has 15 distinct relations in Test partition: strike, used by, is linked to, located in, change effect, target, interact, influence, impact, part of, change expression, is a, compared to, administered, and change abundance; and Train partition has 3 more distinct relation i.e NONE, affect, produced by; which are not in the Test partition as shown in the Table 1. The diversity and imbalance in frequency among these relations as represented in the Table 1 provide defined types common and rare but biologically significant which cause a challenge for relationship extraction. These relations show the entire range of annotated connections in the dataset used and forms a base for all metric computation in this work.

## 4 System Architecture

Here in our system, we have incorporated RAG for RE. It's a four stage pipeline for RE on biomedical dataset which consists of pre-processing and instance construction, dense embedding and indexing, retrieval and few-shot construction, and LLM-based relation generation as shown in the Figure 1. This work compares relationship extraction on biomedical texts using both transformer-based neural model (DeBERTa+CNN) and different LLMs in a RAG pipeline. The goal is a systematic, head-to-head comparison between a strong transformer-based neural model and a range of modern retrieval-augmented LLM pipelines for relationship extraction. All experiments are conducted on BioASQ Conference and Labs of the Evaluation Forum (CLEF) 2025 GutbrainIE challenge dataset (Martinelli et al., 2025), containing PubMed abstracts on annotated entities. This comparison allows for better understanding of the strengths and limitations to both the methodologies addressing real-world biomedical relationship extraction tasks.

Table 1: Ground truth class frequencies for Train and Test dataset.

| Label | Train Freq | Test Freq |
|---|---|---|
| **NONE** | 1053 | 0 |
| TARGET | 173 | 1273 |
| CHANGE EXPRESSION | 36 | 494 |
| IMPACT | 84 | 927 |
| INFLUENCE | 199 | 1955 |
| IS LINKED TO | 171 | 1126 |
| LOCATED IN | 141 | 1865 |
| CHANGE ABUNDANCE | 55 | 1568 |
| USED BY | 82 | 225 |
| **AFFECT** | 105 | 0 |
| PART OF | 44 | 1214 |
| COMPARED TO | 5 | 500 |
| INTERACT | 25 | 2165 |
| IS A | 7 | 4788 |
| **PRODUCED BY** | 6 | 0 |
| ADMINISTERED | 8 | 974 |
| STRIKE | 15 | 612 |
| CHANGE EFFECT | 8 | 541 |

### 4.1 Pre-processing and instance construction

In this pre-processing pipeline, annotated abstracts and titles are systematically processed to extract entity pairs, generate context windows, and accurately map character offsets necessary for sentence-level cross-sections. For sentence segmentation spaCy is used. The process begins with reading training and development data from JSON files, containing biomedical text samples with annotated head and tail entities and their labels. The workflow not only works for positive relation extraction but also creates representative negative samples to enhance model robustness. Downsampling NONE relations

to balance the training set. After working on the cross-sentences the data.

### 4.2 Dense Embedding and Indexing

#### 4.2.1 Embedding

All training samples are converted into Document objects with metadata and passed to LlamaIndex's ingestion pipeline. To improve the retrieval by enhancing the contextual understanding each chunk is converted into a dense vector embedding using the HuggingFaceEmbedding class within LlamaIndex, with the all-MiniLM-L6-v2 model. This process encode the semantic meaning of sentences (which are cross-section sentences from abstracts and titles), it also prevents the relationships central to biomedical reasoning. The embeddings are created using the train and test dataset, where each item from dev dataset is a query as shown in Figure 1. These embeddings are the foundation for any retrieval and inference steps, which make sure the similarities in scientific meaning are being detected and utilized correctly.

#### 4.2.2 Vector Database with LlamaIndex & Qdrant

During index creation all the preprocessed documents from the train dataset are passed through the embedding model and which are then mapped to the vector space. The embeddings are stored and indexed in the vector database, in this work we used Qdrant is used for embedding model (Ockerman et al., 2025) via QdrantVectorStrore in LlamaIndex, which is built for the effective semantic search.

### 4.3 Retrieval and few-shot construction

#### 4.3.1 Query

When test sample is processed, its query embedding is computed using the same embedding model as the training data, but it is not stored in the vector index. Instead, this query embedding is used only at inference time to retrieve the most similar training embeddings. LlamaIndex provides the abstraction to seamlessly convert this query to an embedding and leverage Qdrant's search capabilities.

#### 4.3.2 Retrieval

In the retrieval step, both the queries and database entities are represented using embeddings generated by all-MiniLM-L6-v2 model. This approach allows the vectors for new queries and the stored biomedical data being directly compatible, making it possible to accurately compare their meaning within most relevant scientific contexts and retrieve the most relevant sentences every time.

In this work for every query, we retrieve the most similar training nodes based on vector similarity. These nodes are used as few-shot examples for the generative model which provides the context drawn from similar biomedical content as shown in Figure 1. Every retrieved node is fed into the prompt by formatting its passage, entity, labels and relation information. These components are used into a structured prompt which is fed to the LLM.

### 4.4 LLM-based relation generation

This works evaluates several open-source large language models, to generate the prediction of relations. Each model is fed the augmented prompt containing k real examples along with the query-and is instructed to extract and explain the biomedical relationship between the given entities. In this work, to measure the effectiveness of RAG prompting and measure the capabilities of modern open-source models in biomedical relationship extraction, this study evaluates with multiple state-of-the-art LLMs. Here, we have used two reasoning and two non-reasoning models, each have a distinct combination of reasoning strength, instruction tuning and parameter.

All the models are integrated into the RAG pipeline using LlamaIndex and Ollama infrastructure. Every model received the same structured, context-enriched prompts and produced predictions in standardized JSON format. By systematically using the different number of retrieved examples, this work measured each model's accuracy quality and robustness against ground-truth biomedical relationships from the GutBrainIE dataset.

## 5 Models

The Baseline Model (DeBERTa+CNN) focuses on creating context-aware token embeddings and enhances them with feature extraction using one-layer of convolution.

### 5.1 Baseline Model

In this work, the baseline was conducted using DeBERTa+CNN pipeline which is designed to extract and classify relationships within biomedical texts. Every document is embedded using DeBERTa-v3-base transformer, which captures

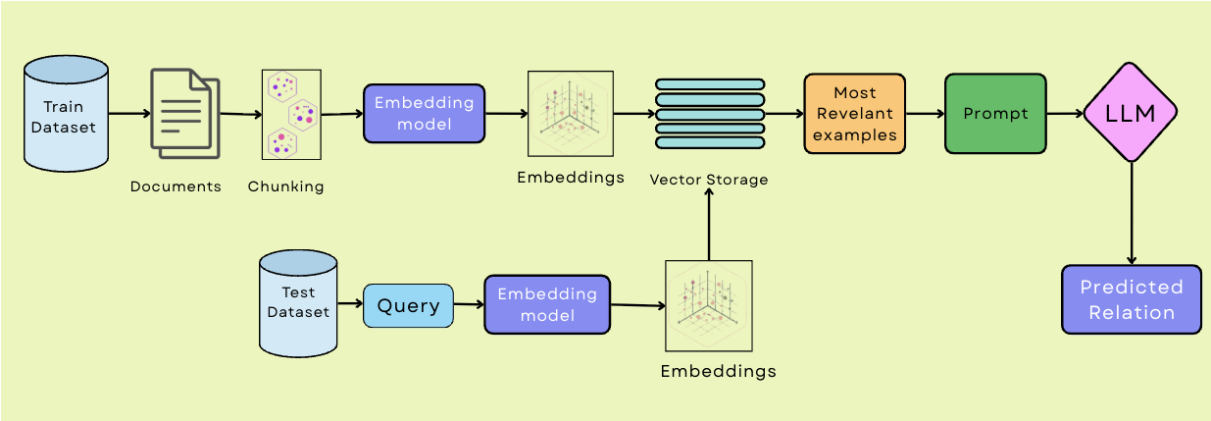


Figure 1: Methodology: RAG for RE on biomedical dataset.

context-dependent meanings of tokens throughout the input text. To further capture semantic information. CNN layer is used directly after embedding. This approach begins with tokenizing input text and feeding it to the DeBERTa-v3-base transformer, which then generates context-aware token embeddings that help the model to have entity recognition capabilities. This transformer-based embeddings are then passed to convolutional layer, consisting of 256 filters and kernel size of three, which allowed for the extraction of local features that are informative for entity interactions. This resulting vector is then passed to a classification layer, giving an 18-dimensional output corresponding to the set of possible relation types defined by the challenge.

### 5.2 Llama-3.1-8B-Instruct - Non-Reasoning Model

In this work, we used instruction-tuned model from the Llama family is designed for conversational and task-oriented NLP. The purpose of using this model was to allow the analysis being fair how compact models maintain-perform over the relations to larger systems when provided with the real domain-specific, high quality retrieval-augmented prompts. While Llama models (Touvron et al., 2023) are strong generalists, it tends to work slightly behind reasoning-focused architectures in complex biomedical tasks.

### 5.3 Gemma3:12B - Reasoning Model

This open-source LLM produced by google (Team et al., 2024), was used in this work to add and broaden the comparison scope and assess performance across diverse reasoning-intensive and biomedical scenarios. Gemma delivers consistent results, especially strong in educational reasoning and standard question-answering tasks.

### 5.4 Qwen3:14B - Non-Reasoning Model

Qwen is a comprehensive language model series that encompasses distinct models with varying parameter counts (Bai et al., 2023). This version of Qwen model does not explicitly include the reasoning of the prediction it made.

### 5.5 Qwen3:14B - Reasoning Model

This version of Qwen model have complex reasoning and advanced comprehension, which makes it particularly effective for interpreting biomedical relationships.The reasoning model responded with the relation and a reasoning of the relation provided.

## 6 Evaluation metrics

To have a fair analysis of relationship extraction performance on biomedical dataset, this work measures Precision and Recall to measure macro and micro F1 score. As in our work there is class imbalance we are measuring both macro F1 score. This capability not only strengthens model interpretability but enhances sensitivity toward minority relations while maintaining overall accuracy.

## 7 Results

In this work, the results indicates that RAG for RE using LLM have performed better DeBERTa + CNN baseline. Here, we measure the performance of both approaches for relationship extraction on the biomedical dataset for a fair comparison.In Table **??**, the macro-averaged metrics is used for evaluation as this method shows how well a model

handles minority classes, which is particularly important in biomedical relationship extraction, when there is class imbalance, which often carry important biomedical meaning.

### 7.1 Comparision between Baseline and other LLMs

Here, in this work DeBERTa + CNN measured Macro F1 score of $0.4404$ and Micro F1 of $0.5625$, giving a foundation for RE to measure againstRAG-based LLM models (Table **??**). Since there are few different LLMs, we compared evaluation score across four of them and the baseline model. Empirically, two of the models i.e Qwen3:14B (Reasoning Model) and Qwen3:14B (Non Reasoning Model), had higher evaluation scores with 10 relevant examples (shots) compared to baseline and other LLMs with RAG. The best RAG systems are Qwen3:14B (non-reasoning) with macro F1 $0.766$ and micro F1 $0.753$, and Qwen3:14B (reasoning) with macro F1 $0.758$ and micro F1 $0.742$, both substantially outperforming the baseline. Using $n = 10$ shots, Llama-3.1-8B-Instruct reaches macro F1 $0.211$ and micro F1 $0.370$, while Gemma3:12B reaches macro F1 $0.3179$ and micro F1 $0.658$, which remain below the Qwen models but illustrate the gains from RAG over the purely supervised baseline in some metrics.

**Llama-3.1-8B-Instruct (Non-Reasoning Model)**: This model shows the weakest performance with macro f1 of $0.211$ and micro F1 of $0.370$; which is even less than the baseline performance, macro F1 of $0.4404$, micro F1 of $0.5625$; showing challenges for the biomedical dataset. As shown in Table 3, this model tend to over-predict some relations with relatively low ground-truth counts, such as *strike* (612 instances vs. 2171 predictions by Llama-3.1-8B-Instruct), and under-predict others, such as *compared to* (500 instances vs. only 64 predictions by Llama-3.1-8B-Instruct). At the same time, prediction behavior is more stable for frequent relations like *is linked to*, *is a*, and *located in*, whose predicted counts stay closer to their ground-truth frequencies across models.

**Gemma3:12B (Reasoning Model)**: This model underperformed the baseline approach than Qwen3 reasoning and non-reasoning models, with macro F1 of $0.317$ and micro F1 of $0.658$, showing limitation in reasoning or comprehending the biomedical dataset for this task, and several classes where it predicts far fewer relations than ground truth (e.g. "change abundance" and "strike) as it can be seen in Table 3.

**Qwen3:14B (Non Reasoning Model)**: For this model, explicitly reasoning is disabled, this version performed comparably well; it achieved macro F1 of $0.763$ and micro F1 of $0.754$. These results shows that even without explicit reasoning outputs, the predicted labels are closer to ground truth(GT) in several categories and performed better than the other models. For example "is a" (3648 vs. GT 4788) and "influence" (3124 vs. GT 1955) as it can be seen in Table 3.

**Qwen3:14B (Reasoning Model)**: This model has consistently outperformed others across evaluation metrics, with macro F1 of $0.758$ among all the other models and highest micro F1 of $0.742$ as it can be seen in Table 2. The prediction across different classes throughout different number of shots shown in Table 3 have significantly improved the F1 score per class as well, as seen in Table **??**. The alignment between the macro and micro results shows that, this model has strong generalization across both frequent relations and non-frequent ones, likely due to improved context understanding supported by retrieval based reasoning. For relations such as "located in" (2302 vs. GT 1865) and "is a" (3862 vs. GT 4788), the model predicts close to ground truth, reflecting improved context comprehension and less over-prediction compared to non-reasoning. The Qwen3 reasoning model's architecture benefits from retrieval augmentation as the number of relevant retrieved example increases it performs better, but it gets saturated after $n = 60$ as it is shown in Figure 2.

### 7.2 Qwen Reasoning vs Non-Reasoning behavior

The observed differences between reasoning and non-reasoning variants, between Qwen-Reasoning and Qwen-Non Reasoning models is very minimal. Due to the class imbalance in the dataset, we use macro F1 to give weight to minority classes, which give weight to minority of classes as shown in Table-3, class labels and their count of ground truth.

Here we have conducted ablations across both the Qwen3:14B reasoning and non-reasoning variants in the RAG pipeline over 0–60 shots. Both exhibit dynamic class prediction behavior, with predicted label frequencies evolving. At $n = 10$, both

Table 2: Comparison of baseline and LLM models on Macro and Micro metrics.

| **Model** | **Macro P** | **Macro R** | **Macro F1** | **Micro P** | **Micro R** | **Micro F1** |
|---|---|---|---|---|---|---|
| Baseline | 0.5181 | 0.4330 | 0.4404 | 0.6585 | 0.4909 | 0.5625 |
| Llama-3.1-8B-Instruct | 0.290 | 0.232 | 0.211 | 0.370 | 0.370 | 0.370 |
| Gemma3:12B | 0.354 | 0.323 | 0.3179 | 0.658 | 0.658 | 0.658 |
| Qwen3:14B non-reasoning | 0.579 | 0.554 | **0.766** | 0.753 | 0.753 | **0.753** |
| Qwen3:14B reasoning | 0.599 | 0.581 | **0.758** | 0.742 | 0.742 | **0.742** |

Table 3: Per-class predicted frequencies for all models compared to ground truth.

| Class Label | Ground Truth | Gemma3:12B | Llama-3.1-8B-Instruct | Qwen3:14B Non-Reasoning | Qwen3:14B Reasoning |
|---|---|---|---|---|---|
| is linked to | 1126 | 3348 | 5754 | 2568 | 2816 |
| compared to | 500 | 479 | 64 | 626 | 670 |
| located in | 1865 | 1930 | 1845 | 2207 | 2302 |
| change abundance | 1568 | 281 | 303 | 746 | 628 |
| is a | 4788 | 3690 | 1229 | 3648 | 3682 |
| target | 1273 | 958 | 1386 | 1091 | 1052 |
| strike | 612 | 2171 | 37 | 356 | 365 |
| interact | 2165 | 1521 | 255 | 1123 | 1267 |
| impact | 927 | 1469 | 678 | 1608 | 1052 |
| influence | 1955 | 2771 | 4313 | 3124 | 3005 |
| administered | 974 | 1352 | 1342 | 777 | 766 |
| change expression | 494 | 334 | 414 | 391 | 376 |
| used by | 225 | 240 | 227 | 117 | 157 |
| part of | 1214 | 1169 | 1295 | 1145 | 1164 |
| change effect | 541 | 425 | 748 | 585 | 568 |

Table 4: Macro F1 and hallucinated labels: Qwen3:14B Reasoning vs Non-Reasoning.

| #Shots | Reasoning Macro F1 | #Halluc. | Non-Reasoning Macro F1 | #Halluc. |
|---|---|---|---|---|
| 0 | 0.084 | 24 | 0.106 | 39 |
| 5 | 0.607 | 5 | 0.607 | 9 |
| 10 | 0.758 | 4 | 0.767 | 5 |
| 15 | 0.816 | 1 | 0.799 | 2 |
| 20 | 0.849 | 0 | 0.852 | 1 |
| 25 | 0.870 | 0 | 0.878 | 1 |
| 30 | 0.876 | 1 | 0.885 | 1 |
| **35** | **0.885** | **1** | **0.892** | **0** |
| 40 | 0.887 | 4 | 0.896 | 4 |
| 45 | 0.893 | 2 | 0.905 | 3 |
| 50 | 0.893 | 5 | 0.899 | 2 |
| 55 | 0.891 | 12 | 0.903 | 10 |
| 60 | 0.891 | 17 | 0.902 | 16 |

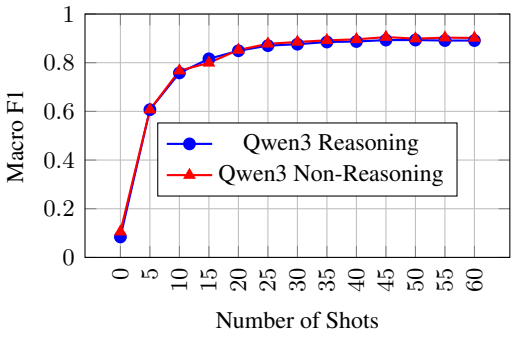


Figure 2: Macro F1 for Qwen3:14B reasoning and non-reasoning across shot counts.

surpass the baseline by ∼0.3 Macro F1, steadily reaching to saturation beyond $n = 60$ (Figure 2).

At low shots (0 to 10), predictions align closely with ground truth but introduce hallucinations (24 vs 39 in zero-shot). Qwen reasoning variants achieve zero hallucinations at 20 and 25 shots(Table 4), yet hallucinated labels got predicted again beyond 30. At 35 shots under Qwen non reasoning there was zero hallucinated label predicted

At high shots (35 to 60), expanded context drives predictions beyond annotated relations by both the Qwen reasoning and non-reasoning model, capturing subtle biomedical interactions absent from the test set. For example, at $n = 40$, both predict low-frequency spurious classes ("administered to" (2), "is caused by" (1), "is" (1), "located_in" (2)). These minimally impact Macro F1 due to their rarity (1 to 4 instances) against 20,227 samples, allowing F1 gains despite increased hallucinations (Figures 3, 2).

Non-reasoning slightly leads post-30 shots while matching hallucination patterns, suggesting few-shot RAG regularizes both similarly.

## 8 Conclusion

This work suggests that retrieval-augmented LLMs can go beyond a strong DeBERTa+CNN baseline for biomedical relationship extraction, especially on rare but important relations, when they are supported with good domain examples. The Qwen3:14B models—most notably the reasoning

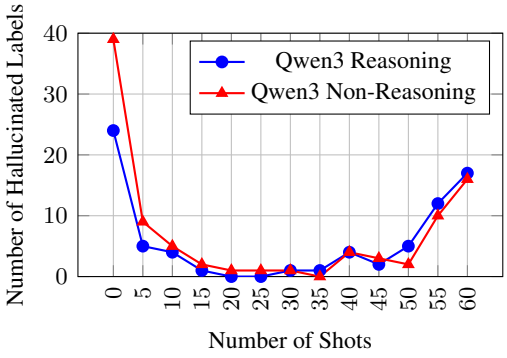


Figure 3: Number of hallucinated labels for Qwen3:14B reasoning and non-reasoning across shot counts.

variant—struck the best balance between macro and micro F1, while also offering clear rationales that make their decisions easier to check and trust. Experiments over different numbers of retrieved shots showed quick improvements up to about 10 examples and a plateau around 60, with a small trade-off: higher performance came alongside a few extra, plausible but unannotated relations.

## 9 Limitations

There are several limitations that motivate future directions. First, the evaluation was conducted on a dataset where every test instance contained at least one annotated relation, so the models were never required to explicitly decide that "no valid relation" exists between entity pairs. This setting does not reflect many real-world biomedical scenarios, where unrelated entities frequently co-occur and models must reliably abstain from predicting a relation. They can hallucinate relation labels, especially under distribution shift or with poorly chosen shots. Generated outputs should be treated as research signals, not as verified biomedical facts.

Finally, the current work focused on relation extraction and did not extend the approach to upstream tasks such as Named Entity Recognition or to richer retrieval paradigms like GraphRAG-based, entity-centric document retrieval. Without these extensions, the pipeline does not yet capture graph-structured biomedical context, which may be necessary to fully support complex, multi-hop relations in noisy real-world datasets.

**Data documentation.** We provide a data sheet in the supplementary material that documents the motivation, composition, collection and preprocessing, intended uses and known limitations of the biomedical relation extraction dataset used in this work.

**Model documentation.** We also provide model cards for our RAG4RE pipeline and the DeBERTa+CNN baseline in the supplementary material, describing their configurations, intended use, training data, evaluation, and limitations.

## A Data Sheet for Biomedical RE Dataset

### A.1 Motivation

This dataset is used to study biomedical relation extraction on the gut–brain axis, focusing on relations between biomedical entities in PubMed abstracts and titles.

### A.2 Composition

The dataset consists of annotated abstracts and titles with head and tail entities and relation labels (including a NONE label). We downsample NONE relations to reduce class imbalance and improve training stability.

### A.3 Collection and Preprocessing

We use the training and development splits provided by the GutBrainIE challenge organizers. Text is segmented into sentences with spaCy, entities and relations are mapped to sentence-level contexts, and cross-sentence are constructed when head and tail entities span multiple sentences.

### A.4 Intended Uses and Risks

This dataset is intended for research on biomedical relation extraction and RAG systems. Potential risks include label noise from distant annotations and domain bias toward gut–brain literature.

### A.5 Distribution

Researchers can obtain the original data from the GutBrainIE challenge or associated CLEF resources. Our processed splits and scripts can be released under the same conditions as original dataset.

## B Model Card for RAG for RE Pipeline

### B.1 Model Details

In our work, main system is Retrieval-Augmented Generation (RAG) pipeline that uses all-MiniLM-L6-v2 embeddings with LlamaIndex, a Qdrant vector database, and Qwen-based large language models as generators. We also train DeBERTa-based encoder with a CNN classifier as a supervised baseline.

### B.2 Intended Use and Risk

The models are intended for research on biomedical relation extraction and evaluation of RAG architectures. They are not intended for any high-stakes medical decision-making.

### B.3 Training Data and Setup

The baseline model is trained on the annotated training split with downsampled NONE relations, using standard supervised learning. The RAG pipeline indexes training corpus in Qdrant via LlamaIndex and retrieves top-$k$ similar examples for each query. We study different few-shot configurations (e.g., 0, 5, 10, 25 shots) in prompts for Qwen models.

### B.4 Evaluation

We evaluate using macro F1, precision, and recall on the held-out development set. Qwen models with 10-shot RAG prompting achieve the highest macro F1 compared to the DeBERTa+CNN baseline under the same setting; at 25 shots, the non-reasoning Qwen model performs best with fewer hallucinated labels.

### B.5 Limitations and Risks

The models are trained on a specific biomedical subdomain and may not generalize to other areas.

## References


Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. arXiv preprint arXiv:2309.16609.

Karen Baxter. 2010. Stockley's Drug interactions pocket companion 2010. Pharmaceutical Press 2010.

Aviv Brokman, Xuguang Ai, Yuhang Jiang, Shashank Gupta, and Ramakanth Kavuluru. 2025. A benchmark for end-to-end zero-shot biomedical relation extraction with llms: Experiments with openai models. Preprint, arXiv:2504.04083.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 17754–17762.

Zisong Chen. 2024. The question-and-answer system based on dpr system and llava.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint, arXiv:1810.04805.

Sefika Efeoglu and Adrian Paschke. 2024. Retrieval-augmented generation-based relation extraction. arXiv preprint arXiv:2404.13397.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022a. Re2G: Retrieve, rerank, generate. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. 2022b. Re2g: Retrieve, rerank, generate. arXiv preprint arXiv:2207.06300.

Jerry R Hobbs and Ellen Riloff. 2010. Information extraction. Handbook of natural language processing, 15:16.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In EMNLP (1), pages 6769–6781.

Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2024. Hybgrag: Hybrid retrieval-augmented generation on textual and relational knowledge bases. arXiv preprint arXiv:2412.16311.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459–9474.

Siran Li, Linus Stenzel, Carsten Eickhoff, and Seyed Ali Bahrainian. 2025. Enhancing retrieval-augmented generation: a study of best practices. arXiv preprint arXiv:2501.07391.

Ben Mann, Nick Ryder, Melanie Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, and 1 others. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 1(3):3.

Marco Martinelli, Gianmaria Silvello, Vanessa Bonato, Giorgio Maria Di Nunzio, Nicola Ferro, Ornella Irrera, Stefano Marchesin, Laura Menotti, Federica Vezzani, and 1 others. 2025. Overview of gutbrainie@ clef 2025: gut-brain interplay information extraction. In CLEF.

Seth Ockerman, Amal Gueroudji, Song Young Oh, Robert Underwood, Nicholas Chia, Kyle Chard, Robert Ross, and Shivaram Venkataraman. 2025. Exploring distributed vector databases performance on hpc platforms: A study with qdrant. arXiv preprint arXiv:2509.12384.

Olawumi Olasunkanmi, Mathew Satursky, Hong Yi, Chris Bizon, Harlin Lee, and Stanley Ahalt. 2025. Relate: Relation extraction in biomedical abstracts with llms and ontology constraints. arXiv preprint arXiv:2509.19057.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2021. Learning to retrieve passages without supervision. arXiv preprint arXiv:2112.07708.

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval, 3(4):333–389.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Hailin Wang, Ke Qin, Rufai Yusuf Zakari, Guoming Lu, and Jin Yin. 2022. Deep neural network-based relation extraction: an overview. Neural Computing and Applications, 34(6):4781–4801.