Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Summarization for Generative Relation Extraction in the Microbiome Domain

Oumaima El Khettari[1] Solen Quiniou[1] Samuel Chaffron[1]
(1) Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000
oumaima.el-khettari@ls2n.fr, solen.quiniou@ls2n.fr,
samuel.chaffron@ls2n.fr

RÉSUMÉ

Résumé automatique pour l'extraction générative de relations dans le domaine du microbiome
Nous explorons une approche générative pour l'extraction de relations, adaptée à l'étude des interactions dans le microbiote intestinal, un domaine biomédical complexe et faiblement doté en données annotées. Notre méthode s'appuie sur le résumé automatique par des grands modèles de langage, pour affiner le contexte, avant d'extraire les relations via une génération guidée par instruction. Les premiers résultats sur un corpus dédié montrent que le résumé automatique améliore les performances de l'approche générative en réduisant le bruit et en orientant le modèle. Cependant, les méthodes d'extraction de relations à l'aide de modèles de type BERT restent plus performantes. Ce travail en cours met en évidence le potentiel des approches génératives pour l'étude de domaines spécialisés en contexte de faibles ressources.

ABSTRACT

We explore a generative relation extraction (RE) pipeline tailored to the study of interactions in the intestinal microbiome, a complex and low-resource biomedical domain. Our method leverages summarization with large language models (LLMs) to refine context before extracting relations via instruction-tuned generation. Preliminary results on a dedicated corpus show that summarization improves generative RE performance by reducing noise and guiding the model. However, BERT-based RE approaches still outperform generative models. This ongoing work demonstrates the potential of generative methods to support the study of specialized domains in low-resources setting.

MOTS-CLÉS : Extraction de relations générative, Ajustement par instruction, Domaine à faible ressources, Microbiome.

KEYWORDS: Generative Relation Extraction, Instruction-tuning, Low-Resource Domain, Microbiome.

1 Introduction

The biomedical field is witnessing a rapid expansion in the study of specialized subdomains under the scope of NLP (He *et al.*, 2023), such as the microbiome interactions (Hogan *et al.*, 2024). The study of the gut microbiome is of high importance (Heintz-Buschart & Wilmes, 2018), as it is a key modulator of human health, influencing gastrointestinal diseases like Inflammatory Bowel Disease, metabolic and neurological conditions such as Parkinson's disease and depression (Shreiner *et al.*, 2015; Góralczyk-Bińkowska *et al.*, 2022; Borrego-Ruiz & Borrego, 2025). Its role in immune regulation, nutrient

Actes de l'atelier Traitement du langage médical à l'époque des LLMs 2025 (MLP-LLM), pages 68–82.

metabolism, and the gut-brain axis has made it a focal point for biomedical research. However, the microbiome remains insufficiently characterized, especially at the species interaction level, due to the vast diversity of microbes (Zhang *et al.*, 2025), variation in microbial communities from one individual to another, and the scattered, often implicit, nature of knowledge in scientific publications. The areas involving this subdomain are often underrepresented in large-scale annotated corpora, making them inherently low-resource for tasks such as information extraction. Among these, relation extraction (RE) is essential for structuring biomedical knowledge and supporting downstream applications like knowledge graph construction and hypothesis generation.

Traditional RE methods typically rely on supervised learning with manually annotated data, which is often scarce or unavailable in emerging subdomains. However, the recent rise of LLMs, pre-trained on extensive corpora, has opened new avenues for addressing low-resource scenarios. These models can be adapted to information extraction task to be performed in a generative manner (Zhu *et al.*, 2023; Xu *et al.*, 2024). In a low resource setting, the generative paradigm can offer flexibility in dealing with unseen relation types and complex, context-rich inputs. In this work, we explore how such models can be leveraged to improve RE in a low-resource biomedical domain, particularly through a generative approach using summarization and instruction tuning. The contributions of this work are as follows:

- We conduct a comprehensive comparison of zero-shot, instruction-tuned, and encoder-based approaches for document-level biomedical relation extraction, highlighting their strengths and limitations ;
- We show that combining summarization with instruction tuning significantly improves generative model performance, especially for smaller models in low-resource settings ;
- We provide insights into model behavior, focusing on hallucinations, output consistency, and adherence to predefined labels in generative models.

2 Related Work

RE is a fundamental task in biomedical NLP, that consists in identifying and classifying semantic relations between entities of interest. Since this task is modeled as a classification problem, the dominant approaches have used supervised classification models, such as convolutional neural networks (Liu *et al.*, 2016; Gu *et al.*, 2017), recurrent neural networks (Jettakul *et al.*, 2019; Mandya *et al.*, 2018), and Transformer-based encoders (Lee *et al.*, 2020; Gu *et al.*, 2021; Beltagy *et al.*, 2019).

More recently, the emergence of LLMs has shifted the paradigm from discriminative to generative approaches in RE (Wadhwa *et al.*, 2023). For example, Asada & Fukuda (2024) investigated RE, using GPT models, from standard biomedical RE datasets, including EU-ADR (Van Mulligen *et al.*, 2012), GAD (Bravo *et al.*, 2015), and ChemProt (Islamaj Doğan *et al.*, 2019). While their use shows promise, they underperform compared to domain-specific models like BioBERT (Lee *et al.*, 2020) and PubMedBERT (Gu *et al.*, 2021). Another study proposed a novel RE method enhanced by LLMs, incorporating techniques such as in-context few-shot learning and chain-of-thought reasoning to improve performance (Zhang *et al.*, 2024a). More specifically, the study of protein-protein interactions as been treated as a binary task (Rehana *et al.*, 2024; Chang *et al.*, 2025). Rehana *et al.* (2024) compared Masked Language Models with different 2 GPT models using different techniques,

and concludes that BERT models perform better on complex datasets, while GPT models can be just as effective on simpler ones. Chang *et al.* (2025) investigated the effect of prompt engineering on the same task and showed that domain-specific prompts significantly enhance the performance of two GPT models. However, a key limitation of these generative approaches is that they are usually evaluated on standard biomedical RE datasets that are either binary or with a limited number of relation types. Brokman *et al.* (2025) applied their method to several biomedical datasets, including BioRED (Luo *et al.*, 2022), a manually annotated corpus with eight relation types and four entity types. This makes BioRED one of the datasets most comparable to our use case in terms of entity and relation diversity. On this dataset, reported F1-scores remained below 10% with GPT-4 and reached 23.2% with OpenAI o1. It is important to note, however, that their evaluation is end-to-end, including both named entity recognition and RE, rather than focusing only on relation classification. Furthermore, BioRED does not focus on a specific low-resource subfield.

LLMs have significantly advanced text summarization, moving beyond extractive methods (Erkan & Radev, 2004; Mihalcea & Tarau, 2004; Nallapati *et al.*, 2017) to enable fluent and context-aware generation (Lewis *et al.*, 2020; Gupta & Gupta, 2019). Modern summarization systems can synthesize information across sentences, rephrase content, and adapt to various domains, including scientific and medical texts (Goyal *et al.*, 2022; Zhang *et al.*, 2024b). Recent directions further explore personalized (Pu *et al.*, 2023), human-in-the-loop (Chen *et al.*, 2023), and real-time summarization, reflecting a shift toward more interactive and user-adaptive systems. These advances highlight the role of LLMs in redefining summarization as a dynamic and domain-sensitive task.

Summarization and RE share similar challenges, as both require identifying and condensing key information from text (Zhang *et al.*, 2023). LLMs perform well in summarization tasks due to their extensive pretraining on diverse and domain-specific corpora (Van Veen *et al.*, 2023; Tang *et al.*, 2023; Shaib *et al.*, 2023), but their effectiveness in structured information extraction tasks like RE remains limited (Naguib *et al.*, 2024; Ma *et al.*, 2023; Zhang *et al.*, 2024c). The combination of summarization and RE, two tasks with clear conceptual overlap in the generative paradigm, remains largely unexplored, especially in the context of closed generative RE (Li *et al.*, 2023).

3 Methodology

We explore two strategies for performing RE using LLMs: a direct approach and a summarization-based approach.

3.1 Direct RE Approach

In the direct RE setting, we adopted a zero-shot framework. The model is prompted with an instruction, the input passage, and a list of candidate relation types. Then, it is expected to identify the correct relation between the target entities, without additional training. We designed a prompt to elicit the relation type between two target entities within a given text passage. The prompt consists of three main components:

- Task instruction: Instructs the model to generate the class of the relation between entity1 and entity2 based on the input paragraph.

- Constraints: Specifies two requirements (i) the model must output only the relation label, and (ii) no justification should be provided. Classes: Increase, Decrease, Stop, Start, Improve, Worsen, Presence, Negative_correlation, Affects, Causes, Complicates, Experiences, Interacts_with, Location_of, Marker/Mechanism, Prevents, Reveals, Treats, Physically_related_to, Part_of, Possible, Associated_with, None.
- Input: A passage excerpt from the source article containing the target entities.

3.2 Summarization-Based RE Approach

The summarization-based approach is a two-step process:

1. A summarization step: an LLM is instructed to generate a concise summary that captures the relation between a pair of entities, based on the surrounding context. This step reduces input noise and mitigates hallucination by focusing on the core semantic relation;
2. A fine-tuning step: a model is instruction-tuned on relation classification task, and is given the generated summaries as input.

This approach relies on a lighter input than the previous one, as the relation types are learned during the fine-tuning step and are not included in the prompt.

During the summarization step, the model is prompted to generate a concise summary that captures the relation between the two target entities, based on the surrounding context. In the instruction-tuning step, the model is prompted as if it were a biology expert. It is told that the provided paragraph describes a relation between two entities and is instructed to identify the expressed relation. The input paragraph is provided without any predefined class list, as the model is expected to learn the relation types during training. The prompt is designed as follows:

- System Role: You are a biology expert. You role is to answer the following instruction.
- Task instruction: This is a paragraph describing the relation between Entity1 and Entity2. Find the expressed relation in the text.
- Input: A passage excerpt from the source article containing the target entities.

4 Experimental setup

4.1 MicrobioRel Corpus

We conducted our experiments on MicrobioRel[1], a domain-specific corpus for multi-class RE in the microbiome-related biomedical literature. The dataset consists of manually annotated excerpts from scientific articles, covering 22 relation types across six entity categories: species, diseases, chemicals, mutations, genes, and cell lines. With only 1,994 annotated relations in total (see Table 2,

[1] https://github.com/Stan8/MicrobioRel-dataset

in Appendix A), MicrobioRel places this study in a low-resource setting, targeting a biomedical subdomain that remains largely unexplored in NLP research.

All named entities in MicrobioRel were initially pre-annotated using PubTator (Wei *et al.*, 2019). As shown in (El Khettari *et al.*, 2023), PubTator2 demonstrates strong performance in recognizing species mentions, which are one of the most important and challenging entity types in our study, alongside diseases. This step is followed by manual correction by domain experts under specific guidelines, to ensure the correctness of entities participating in relations. For relation annotation, annotators were presented with paragraphs containing at least two entities. First, annotators were instructed to identify all relevant relations by selecting pairs of entities that could plausibly be linked. Once a relation was established, it was then categorized according to a predefined schema. A "None" class was automatically introduced to support negative examples that could not be annotated manually due to the complexity of identifying all such instances exhaustively. This was done by generating all possible pairs of entities within a paragraph that were not annotated as being related. To avoid overwhelming annotators with negative examples, the number of "None" instances was limited proportionally to the length of each paragraph.An example from MicrobioRel is given in Appendix B.

4.2 Models

To prevent biasing the model by presenting only a subset of relation classes in demonstrations, we use a zero-shot framework for direct RE, using Llama2 13B (Touvron *et al.*, 2023), Llama3 8B, Llama 3.2-3B-Instruct (Dubey *et al.*, 2024), Mistral 7B-Instruct (Jiang *et al.*, 2023), and BioMistral 7B (Labrak *et al.*, 2024). For the initial summarization task, we employ Llama 3.1 in a few-shot setting. The demonstration examples provided follow a simple structure: *Entity1 Relation Entity2*, expressed in a single sentence, to avoid adding constraints on the output structure. Then, instruction-tuning is performed using parameter-efficient fine-tuning techniques, specifically the Unsloth framework[2] and LoRA, applied to the Llama-3 3B Instruct, Mistral 7B, and Llama3 8B models, with configurations adapted to each model's architecture. We choose to instruction-tune smaller models due to computational constraints and the limited size of the MicrobioRel dataset, which makes smaller models more suitable for effective adaptation. +[] Furthermore, we compare these models to biomedical BERT-based models, BioBERT (Lee *et al.*, 2020), SciBERT (Beltagy *et al.*, 2019), PubMedBERT (Gu *et al.*, 2021) and BioLinkBERT (Yasunaga *et al.*, 2022), fine-tuned for 30 epochs, using a fixed batch size of 4 and a learning rate set to 1×10^{-5}.

4.3 Evaluation metrics

To assess the quality of the generated summaries, we employ BERTScore (Zhang *et al.*, 2019) and cosine similarity. These metrics respectively capture fine-grained semantic alignment and overall representational similarity between the generated summaries and the original texts. We also report the F1-BERTScore and cosine similarity scores prior to RE, in order to evaluate how well the summaries capture relational information. Evaluation of the RE task is carried out using weighted precision, recall, and F1-score. Under an exact match criterion, a prediction is considered correct only if it matches the gold relation label. Although, human evaluation could provide qualitative insights, we

[2] https://github.com/unslothai/unsloth

focus on strict, reproducible classification performance.

$$\text{Weighted Metric} = \sum_{i=1}^{n} \left(\frac{\text{support}_i}{\text{total support}} \times \text{metric}_i \right) \tag{1}$$

In the weighted setting, the precision, recall, and F1-score are computed as the weighted averages of the corresponding per-class scores, where the weights are based on the support of each class (see Equation 1). Note that weighted F1 is derived from individual per-class F1 scores, not from weighted precision and recall. Metrics are computed after post-processing the LLM outputs to remove text beyond the predicted relation class. Only exact matches with the gold labels are considered correct.

5 Results and discussions

5.1 Relation Extraction Performance

Table 1 presents the performance of the proposed RE approaches, comparing the direct prompting method, the summarization-based pipeline, and BERT-based models across the 23 relation classes.

RE Approach	Models	P	R	F1
Direct Approach	Llama2 13B	**39.2**	**15.2**	**14.3**
	Llama3 8B	11.3	11.2	8.34
	Llama 3.2-3B-Instruct	4.9	7.47	3.78
	Mistral 7B-Instruct	26.9	10.4	4.95
	BioMistral 7B	9.41	9.87	7.07
Summarization-Based	Llama3 8B	34.6	23.5	24.51
	Llama 3.2-3B-Instruct	62.3	**60.2**	**59.7**
	Mistral 7B-Instruct	**62.9**	46.9	44.6
Regular Fine-tuning	PubMedBERT	**71.9**	**71.7**	**71.3**
	BioLinkBERT	71.2	71.2	70.8
	SciBERT	69.7	69.3	69.0
	BioBERT	68.1	68.5	67.8

Table 1: Weighted Precision, Recall, and F1-scores (in %) of RE models on the test set of MicrobioRel. Regular fine-tuning refers to supervised training on the task, while other rows correspond to generative models used with or without instruction tuning.

Results highlight the comparative strengths and weaknesses of different approaches to RE in the microbiome domain. Traditional biomedical models such as PubMedBERT, BioBERT, SciBERT, and BioLinkBERT continue to outperform generative models, with PubMedBERT achieving the highest F1 score of 71.3% . This confirms the effectiveness of domain-specific pre-training and task-aligned fine-tuning for biomedical RE.

Direct Approach: Zero-Shot Limitations

Zero-shot RE yielded in poor results for both versions of Llama. Llama2, in particular, exhibited

frequent hallucinations, defined here as generating relation types outside the predefined label set. These hallucinations are not limited to under represented relation types, nor do they consistently affect a specific subset of relations. Instead, they depend more on the context in which the relation occurs. They can be classified into two categories: (i) relation types that are semantically similar to valid classes (e.g., *Reduce* instead of *Decrease*), and (ii) context-specific phrases that imply a link but do not correspond to any annotated relation. These issues highlight the lack of strict adherence to instructions and class constraints. Moreover, Llama2 predictions were biased toward more general relation types, resulting in higher recall on dominant classes but limited ability to distinguish fine-grained, domain-specific relations. In contrast, Llama3 produced more consistent outputs that were closer to the target label set. This reduced hallucination rate eased post-processing and demonstrated improved alignment with the task. Notably, Llama3 made more balanced use of the available relation classes but sometimes at the expense of accuracy on broader categories. Overall, the performance of Llama models improved with the increase in model size.

In this setting, Mistral 7B-Instruct showed strong bias toward the *associated_with* class (284 instances predicted), resulting in low label diversity and poor F1 of 4.95%. BioMistral 7B produced more varied outputs, favoring *increase* (240 instances), but it also struggled with precision and class balance since it generated only six relation types out of 23. Overall, both models underperformed, highlighting the difficulty of zero-shot generative RE in specialized biomedical domains.

Summarization-Based Approach: Instruction-Tuning on Smaller Models

Incorporating summarization improved the performance of generative models. By reducing contextual noise and narrowing the focus to task-relevant information, summarization made it easier for models to extract relations. Among generative approaches, Llama 3.2-3B-Instruct achieved the highest F1 score of 59.7%, representing a substantial improvement over its zero-shot performance of 3.78%. With this method, the model produced more aligned, concise, and consistent outputs; reduced hallucinations; and demonstrated a greater ability to map its responses to predefined relation classes.

In contrast, the fine-tuned Mistral 7B-Instruct model continued to exhibit several issues encountered in zero-shot setting. Although fine-tuning provided a moderate improvement in classification accuracy, hallucinations and variability in output formatting persisted. For instance, the model generated semantically similar but structurally inconsistent relation types such as *experience*, *experiences*, and *experiencer*. It also introduced relations not present in the training data, suggesting a tendency to overgeneralize beyond the expected label set. Instruction-tuning Llama3 8B followed the same trend. The inconsistencies in the hallucinations and the output format are better than the ones of the zero-shot approach, but are still persistent regardless of the instruction tuning.

The performance gap between instruction-tuned models can be partly explained by their number of trainable parameters. Llama 3.2-3B (24M) and Llama3 8B (50M) adapted well to the small MicrobioRel dataset, while Mistral 7B (95M) showed signs of overfitting and hallucination. Larger models typically require more data to generalize, and limited supervision can lead to unstable outputs, as supported by prior work (Gekhman *et al.*, 2024).

Regular Fine-tuned Models

BERT-based biomedical models such as PubMedBERT, BioBERT, SciBERT, and BioLinkBERT outperform all other tested models, with PubMedBERT achieving the highest F1 score. This highlights the effectiveness of domain-specific pretraining and regular fine-tuning for RE in biomedical texts.

Despite the rise of instruction-tuned and generative models, specialized encoders remain highly effective, especially when task formulation is well aligned with traditional classification.

Overall, our findings show that while zero-shot RE with generative models struggles in a specialized multi-class biomedical RE, combining instruction tuning with summarization leads to marked improvements, particularly for smaller, more adaptable models. Instruction tuning helps enforce task constraints, and summarization reduces contextual noise, making this strategy especially promising in low-resource settings. Although fine-tuned encoder models like PubMedBERT still achieve the best overall performance, generative models can be presented as an alternative when annotation resources are limited. In future work, it would be valuable to explore the threshold at which datasets become too small for traditional fine-tuning to remain effective, opening the door for lightweight, tuned generative models to be better fitted for the task.

5.2 Summary Quality Evaluation

To assess the quality of the generated summaries, we evaluated their semantic similarity to the original passages using two complementary metrics: cosine similarity and F1-BERTScore.

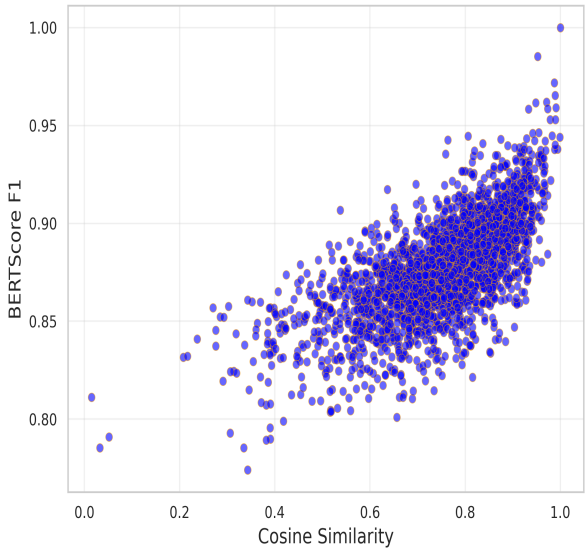

Figure 1: Relationship between Cosine Similarity and F1-BERTScore, for the evaluated textual pairs.

As shown in Figure 1, there is a moderate to strong positive Pearson correlation between these metrics, indicating that both capture semantic alignment reliably. Interestingly, F1-BERTScore remains relatively high (0.80–0.90) even when cosine similarity is low (< 0.5), highlighting its robustness in detecting semantic overlap beyond surface-level similarity. When cosine similarity is high (approaching 1.0), both metrics converge, suggesting strong alignment between summary and original text.

Additionally, correlation analysis revealed that longer original passages tend to slightly reduce similarity scores, which is likely due to increased textual noise, while longer summaries are weakly associated with higher similarity. Overall, this evaluation supports that the generated summaries preserve the core semantic content needed for downstream RE.

5.3 Error Analysis

Given that our summarization-based method is a two-step pipeline, it is susceptible to error propagation. To better understand the origin of performance limitations, we conduct an error analysis aimed at distinguishing errors coming from the summarization step from those introduced during the relation classification phase. We manually inspect a subset of 30 examples from the test set where the predictions diverge from the gold annotations. These cases are categorized into three distinct error types: errors from the summaries, where the summary does not contain the relevant interaction (Class 1), counting 6 elements, errors originating from the RE model (Class 2), counting 13 elements, and predictions that, although different from the annotation, are valid and semantically plausible interpretations of the input (Class 3), counting 11 elements.

Class 1 Errors The fewest errors were attributed to the summarization step (Class 1), suggesting that the summaries generally preserved the key interactions necessary for RE. This indicates that the summarization phase was effective in retaining relevant information. An illustrative example of this type of errors is presented below:

> **Original text:** "Nevertheless, the role of FXR in diabetes is controversial. Although study has reported that upregulating FXR in diabetic mice can significantly improve hypercholesterolemia, other study found that inhibition of intestinal FXR can reduce hepatic gluconeogenesis, promote the secretion of GLP-1, and reduce body weight. Therefore, whether upregulating or inhibiting FXR signaling is an innovative approach in the control of blood glucose in patients with T2D needs further research." **Generated Summary:** "Unfortunately, there is no relation mentioned between FXR (Farnesoid X Receptor) and blood glucose in the given text. The text discusses various topics such as depression, hypertension, gut microbiota, and diabetes, but it does not mention a direct relationship between FXR and blood glucose. "
>
> **Extracted Relation:** None **Annotated Relation:** Affects

In this example, the relevant entities are *FXR* and *Blood Glucose*. While the original passage discusses the potential influence of FXR signaling on blood sugar regulation in type 2 diabetes (T2D), the summary fails to catch this relation. This can be explained by the speculative nature of the original sentence, which could lead the summary to focus on general context rather than nuanced interactions. The extracted relation, therefore, remains correct relative to the generated summary.

Class 2 Errors Class 2 errors are mainly common when the expected relation is linked to a quantity or a property. The relation *Increase* is particularly present in the examples belonging to this class, because the relation does not refer directly to the entity itself, but to an attribute, resulting in the prediction of a different relation.

> **Original text:** "Patients suffering from PD have different bacterial flora compared with healthy controls. Microbiota have an impact on disease progression, which has been established in the fecal-transplantation experiments. Administration of the microbiota of patients with Parkinson's disease to mice showed the development of neuroinflammatory processes and motor deficits." **Generated Summary:** "**Relation:** The bacterial flora

in patients suffering from PD is different from that of healthy controls.**Effect:** The altered microbiota can impact the progression of PD, specifically leading to neuroinflammatory processes and **motor deficits**."

Extracted Relation: Associated with **Annotated Relation:** Increase

In the example above, the relation to capture is between *Parkison's disease* and *motor deficits*. The reflected relation in the generated summary focuses on the entire context of the text, which includes bacterial flora of PD patients. This mismatch arises because the model predicts a broad association rather than the specific directional change implied by the annotated relation. The increase in motor deficits is indirectly expressed and mediated by microbiota, making it harder for the model to link it causally to Parkinson's disease.

Class 3 Errors For Class 3 errors, the model's predictions are in fact reasonable, despite being different from the annotated relations. These examples highlight the divergence between human annotation and the model's interpretation, revealing that both can arrive at different yet reasonable understandings of the same passage. Many of these examples involve the automatically assigned None class, suggesting that some relations captured by the model may have been missed during annotation.

Original text: "Probiotics Mixture Increases Butyrate, and Subsequently Rescues the Nigral Dopaminergic Neurons from MPTP and Rotenone-Induced Neurotoxicity." **Generated Summary:** "The relation between MPTP (1-methyl-4-phenyl-1,2,5,6-tetrahydropyrine) and Nigral Dopaminergic Neurons is that:MPTP-induced neurotoxicity can damage nigral dopaminergic neurons, but a probiotics mixture can increase butyrate levels, which in turn rescues these neurons from MPTP-induced neurotoxicity."

Extracted Relation: Worsen **Annotated Relation:** Affects

In summary, most errors originate from the RE step rather than the summarization one, which generally preserves the semantic information of interest. The analysis highlights the challenge of balancing contextual understanding with precise semantics, especially in subtle or ambiguous cases. It also points out the limitations of automatically assigning "no relation" labels, and the need for refinement in handling nuanced interactions.

6 Conclusion

This on-going work investigated the use of LLMs for biomedical RE, introducing summarization as an effective intermediate step. The approach helped reduce textual noise and better align model outputs with structured task requirements, even in low-resource settings. While promising, the study also highlighted persistent challenges such as hallucinations, inconsistent outputs, and difficulty capturing subtle context-dependent relations. These findings point to the need for further refinement in generative RE pipelines, particularly in balancing semantic precision with contextual understanding, specifically in low resource specialized context.

7 Limitations

This study focuses exclusively on the MicrobioRel dataset, as our primary interest lies in the analysis of interactions within the gut microbiome. At present, MicrobioRel is one of the few available manually annotated resources having a multi-class setting (with more than ten relation types) and multiple entity types. However, we acknowledge the importance of expanding our evaluation to additional datasets in future work, even if they differ in scope as they could provide valuable perspectives and robustness checks for our approach. Moreover, we do not compare our method directly with other generative RE approaches, as most of them either address simpler binary tasks or evaluate performance in an end-to-end way by including named entity recognition, making direct comparisons difficult. Adapting our pipeline to these settings would be a promising direction for future extensions.

Another limitation of our study is that we did not generate multiple summaries per passage or apply any selection strategy. Instead, we relied on a single generated summary, selected based on its semantic similarity to the original text. Future work could explore generating multiple summaries and selecting the most relevant one using task-specific quality criteria.

Finally, while our summary-based method shows encouraging results, traditional BERT-based RE models continue to outperform it in terms of raw extraction accuracy. This suggests that further research is needed to enhance the effectiveness of generative models, particularly if they are to offer a computational or practical advantage over current methods.

References

ASADA M. & FUKUDA K. (2024). Enhancing relation extraction from biomedical texts by large language models. In *Proc. of the HCI International Conference*, p. 3–14.

BELTAGY I., LO K. & COHAN A. (2019). SciBERT: A pretrained language model for scientific text. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *Proc. of EMNLP-IJCNLP*, p. 3615–3620.

BORREGO-RUIZ A. & BORREGO J. J. (2025). Human gut microbiome, diet, and mental disorders. *International Microbiology*, **28**(1), 1–15.

BRAVO À., PIÑERO J., QUERALT-ROSINACH N., RAUTSCHKA M. & FURLONG L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, **16**, 1–17.

BROKMAN A., AI X., JIANG Y., GUPTA S. & KAVULURU R. (2025). A benchmark for end-to-end zero-shot biomedical relation extraction with llms: Experiments with openai models. *arXiv preprint arXiv:2504.04083*.

CHANG Y.-C., HUANG M.-S., HUANG Y.-H. & LIN Y.-H. (2025). The influence of prompt engineering on large language models for protein–protein interaction identification in biomedical literature. *Scientific Reports*, **15**(1), 15493.

CHEN J., DODDA M. & YANG D. (2023). Human-in-the-loop abstractive dialogue summarization. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proc. of ACL*, p. 9176–9190.

DUBEY A., JAUHRI A., PANDEY A., KADIAN A., AL-DAHLE A., LETMAN A., MATHUR A., SCHELTEN A., YANG A., FAN A. *et al.* (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

EL KHETTARI O., QUINIOU S. & CHAFFRON S. (2023). Building a corpus for biomedical relation extraction of species mentions. In D. DEMNER-FUSHMAN, S. ANANIADOU & K. COHEN, Éds., *Proc. of BioNLP*.

ERKAN G. & RADEV D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, **22**, 457–479.

GEKHMAN Z., YONA G., AHARONI R., EYAL M., FEDER A., REICHART R. & HERZIG J. (2024). Does fine-tuning LLMs on new knowledge encourage hallucinations? In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proc. of EMNLP*, p. 7765–7784.

GÓRALCZYK-BIŃKOWSKA A., SZMAJDA-KRYGIER D. & KOZŁOWSKA E. (2022). The microbiota–gut–brain axis in psychiatric disorders. *International journal of molecular sciences*, **23**(19), 11245.

GOYAL T., LI J. J. & DURRETT G. (2022). News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

GU J., SUN F., QIAN L. & ZHOU G. (2017). Chemical-induced disease relation extraction via convolutional neural network. *Database*, **2017**, bax024.

GU Y., TINN R., CHENG H., LUCAS M., USUYAMA N., LIU X., NAUMANN T., GAO J. & POON H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, **3**(1), 1–23.

GUPTA S. & GUPTA S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, **121**, 49–65.

HE Z., WANG Y., YAN A., LIU Y., CHANG E., GENTILI A., MCAULEY J. & HSU C.-N. (2023). MedEval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proc. of EMNLP*.

HEINTZ-BUSCHART A. & WILMES P. (2018). Human gut microbiome: function matters. *Trends in microbiology*, **26**(7), 563–574.

HOGAN W., BARTKO A., SHANG J. & HSU C.-N. (2024). Midred: An annotated corpus for microbiome knowledge base construction. In *Proc. of BioNLP*, p. 398–408.

ISLAMAJ DOĞAN R., KIM S., CHATR-ARYAMONTRI A., WEI C.-H., COMEAU D. C., ANTUNES R., MATOS S., CHEN Q., ELANGOVAN A., PANYAM N. C. *et al.* (2019). Overview of the biocreative vi precision medicine track: mining protein interactions and mutations for precision medicine. *Database*, **2019**, bay147.

JETTAKUL A., WICHADAKUL D. & VATEEKUL P. (2019). Relation extraction between bacteria and biotopes from biomedical texts with attention mechanisms and domain-specific contextual representations. *BMC bioinformatics*, **20**, 1–17.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b.

LABRAK Y., BAZOGE A., MORIN E., GOURRAUD P.-A., ROUVIER M. & DUFOUR R. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proc. of ACL*, p. 5848–5864.

LEE J., YOON W., KIM S., KIM D., KIM S., SO C. H. & KANG J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, **36**(4), 1234–1240.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proc. of ACL*, p. 7871–7880.

LI G., WANG P. & KE W. (2023). Revisiting large language models as zero-shot relation extractors. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proc. of EMNLP*.

LIU S., TANG B., CHEN Q. & WANG X. (2016). Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, **2016**(1), 6918381.

LUO L., LAI P.-T., WEI C.-H., ARIGHI C. N. & LU Z. (2022). Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, **23**(5), bbac282.

MA Y., CAO Y., HONG Y. & SUN A. (2023). Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proc. of EMNLP*.

MANDYA A., BOLLEGALA D., COENEN F. & ATKINSON K. (2018). Combining long short term memory and convolutional neural network for cross-sentence n-ary relation extraction. *arXiv preprint arXiv:1811.00845*.

MIHALCEA R. & TARAU P. (2004). Textrank: Bringing order into text. In *Proc. of EMNLP*, p. 404–411.

NAGUIB M., TANNIER X. & NEVEOL A. (2024). Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. In *Proc. of EMNLP*, p. 6829–6852.

NALLAPATI R., ZHAI F. & ZHOU B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proc. of AAAI*, p. 3075–3081.

PU X., GAO M. & WAN X. (2023). Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

REHANA H., ÇAM N. B., BASMACI M., ZHENG J., JEMIYO C., HE Y., ÖZGÜR A. & HUR J. (2024). Evaluating gpt and bert models for protein–protein interaction identification in biomedical text. *Bioinformatics Advances*, **4**(1), vbae133.

SHAIB C., LI M., JOSEPH S., MARSHALL I., LI J. J. & WALLACE B. (2023). Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proc. of ACL*, p. 1387–1407.

SHREINER A. B., KAO J. Y. & YOUNG V. B. (2015). The gut microbiome in health and in disease. *Current opinion in gastroenterology*, **31**(1), 69–75.

TANG L., SUN Z., IDNAY B., NESTOR J. G., SOROUSH A., ELIAS P. A., XU Z., DING Y., DURRETT G., ROUSSEAU J. F. *et al.* (2023). Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, **6**(1), 158.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

VAN MULLIGEN E. M., FOURRIER-REGLAT A., GURWITZ D., MOLOKHIA M., NIETO A., TRIFIRO G., KORS J. A. & FURLONG L. I. (2012). The eu-adr corpus: annotated drugs, diseases, targets, and their relationships. *Journal of biomedical informatics*, **45**(5), 879–884.

VAN VEEN D., VAN UDEN C., BLANKEMEIER L., DELBROUCK J.-B., AALI A., BLUETHGEN C., PAREEK A., POLACIN M., REIS E. P., SEEHOFNEROVA A. *et al.* (2023). Clinical text summarization: adapting large language models can outperform human experts. *Research Square*.

WADHWA S., AMIR S. & WALLACE B. (2023). Revisiting relation extraction in the era of large language models. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proc. of ACL*.

WEI C.-H., ALLOT A., LEAMAN R. & LU Z. (2019). Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, **47**(W1), W587–W593.

XU D., CHEN W., PENG W., ZHANG C., XU T., ZHAO X., WU X., ZHENG Y., WANG Y. & CHEN E. (2024). Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, **18**(6), 186357.

YASUNAGA M., LESKOVEC J. & LIANG P. (2022). LinkBERT: Pretraining language models with document links. In S. MURESAN, P. NAKOV & A. VILLAVICENCIO, Éds., *Proc. of ACL*, p. 8003–8016.

ZHANG J., WIBERT M., ZHOU H., PENG X., CHEN Q., KELOTH V. K., HU Y., ZHANG R., XU H. & RAJA K. (2024a). A study of biomedical relation extraction using gpt models. *AMIA Summits on Translational Science Proceedings*, **2024**, 391.

ZHANG J., ZHANG D., XU Y., ZHANG J., LIU R., GAO Y., SHI Y., CAI P., ZHONG Z., HE B. *et al.* (2025). Large-scale biosynthetic analysis of human microbiomes reveals diverse protective ribosomal peptides. *Nature Communications*, **16**(1), 3054.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

ZHANG T., LADHAK F., DURMUS E., LIANG P., MCKEOWN K. & HASHIMOTO T. B. (2024b). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, **12**, 39–57.

ZHANG W., LU W., WANG J., WANG Y., CHEN L., JIANG H., LIU J. & RUAN T. (2024c). Unexpected phenomenon: LLMs' spurious associations in information extraction. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proc. of ACL*.

ZHANG Z., ELFARDY H., DREYER M., SMALL K., JI H. & BANSAL M. (2023). Enhancing multi-document summarization with cross-document graph-based information extraction. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proc. of EACL*.

ZHU Y., YUAN H., WANG S., LIU J., LIU W., DENG C., CHEN H., LIU Z., DOU Z. & WEN J.-R. (2023). Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.

A MicrobioRel: Relations Statistics

Relation type	Count	Relation type	Count
Associated_with	210	Possible	56
Part_of	204	Worsen	51
Affects	202	Marker-Mechanism	41
Increase	185	Prevents	39
Location_of	146	Physically_related_to	31
Causes	144	Negative_correlation	23
Experiences	132	Treats	17
Decrease	121	Complicates	10
Start	118	Presence	9
Improve	102	Reveals	7
Interacts_with	86	Stop	6

Table 2: Relation label counts in the MicrobioRel corpus

B MicrobioRel Example

Text: "During PD development, microbiota changes can be accompanied by reduced concentrations of branched-chain amino acids (BCAAs) and aromatic amino-acids in comparison with the healthy control group."

Entity 1: PD
Entity 2: branched-chain amino acids
Annotated Relation: Decrease