Zero-Shot Cross-Sentential Scientific Relation Extraction via Entity-Guided Summarization

Vani Kanjirangat[1], **Fabio Rinaldi**[1]

[1]SUPSI, IDSIA, Switzerland

{vani.kanjirangat, fabio.rinaldi}@supsi.ch

Abstract

Structured information extraction (IE) from scientific abstracts is increasingly leveraging large language models (LLMs). A crucial step in IE is relation extraction (RE), which becomes challenging when entity relations span sentences. Traditional path-based methods, such as shortest dependency paths, are often unable to handle cross-sentential relations effectively. Although LLMs have been utilized as zero-shot learners for IE tasks, they continue to struggle with capturing long-range dependencies and multi-hop reasoning. In this work, we propose using GPT as a zero-shot entity-guided summarizer to encapsulate cross-sentential context into a single-sentence summary for relation extraction. We perform intrinsic evaluations, comparing our approach against direct zero-shot prompting on biomedical scientific abstracts. On the Chemical-Disease Relation (CDR) dataset, our method achieves a 7-point improvement in overall F-score and 6 points for cross-sentential relations. On the Gene-Disease Association (GDA) dataset, we observe an 8-point gain for inter-sentential relations. These results demonstrate that entity-guided summarization with GPT can enhance zero-shot biomedical RE, supporting more effective structured information extraction from scientific texts.[1].

1 Introduction

In structured information extraction from scientific literature, identifying and extracting entity relations is a key intermediate step, for example, in building knowledge graphs. A typical IE pipeline includes named entity recognition (NER), entity linking/normalization, relation extraction (RE), optional event/fact extraction, and knowledge base population (KBP) (Dagdelen et al., 2024; Jaradeh et al., 2023). With advances in generative language models, zero-shot (ZSL) and few-shot learning (FSL) have become increasingly popular for IE and other NLP tasks (Dagdelen et al., 2024; Hou et al., 2024; Savelka, 2023; Shu et al., 2022; Wu et al., 2025). Parallel research has explored the limitations of zero-shot learning (ZSL) across various domains (Manikandan et al., 2023; Lauscher et al., 2020; Kanjirangat et al., 2024; Al Nazi et al., 2025). GPT-based models (Radford et al., 2019; Liu et al., 2023; Achiam et al., 2023) and open-source models such as Falcon (Almazrouei et al., 2023), Bloom (Le Scao et al., 2023), LLaMA (Touvron et al., 2023), and Mistral (Jahan et al., 2023) have demonstrated strong capabilities in knowledge-intensive tasks, including question answering and summarization. However, their performance in classification tasks can be limited by factors such as domain specificity. For example, they excel in sentiment analysis or intent classification (Wei et al., 2021) but often struggle with clinical or biomedical classification. These limitations are especially pronounced in complex tasks like relation identification and causality detection(Armengol-Estapé et al., 2021; Khondaker et al., 2023; Lai et al., 2023; Yang et al., 2023; Bi et al., 2025; Chen et al., 2025). Considering the above points, this work focuses on the relation extraction task under two key constraints: (i) addressing complex cross-sentential relations and (ii) focusing the task within the biomedical domain. Concerning relation extractions, efforts have been made to leverage LLMs, specifically focusing on improving prompting approaches (Li et al., 2023; Wadhwa et al., 2023; Laskar et al., 2025), which have demonstrated performance upgrades. The potentials and limitations of GPT models in biomedical information extraction have been reported in multiple studies. It has been shown that even though GPT-4 had achieved near state-of-the-art results in few-shot knowledge transfer in open-domain NLP tasks, it underperformed the domain-specific models such as BioBERT (Lee et al., 2020) or SciBERT (Beltagy et al., 2019), which are or-

[1]Experimental codes will be made available

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 22–33

ders of magnitude smaller than them (Chen et al., 2024; Moradi et al., 2021; Ateia and Kruschwitz, 2023; Nori et al., 2023; Waisberg et al., 2023). The limitations and capacity of zero-shot LLMs are less explored (Jahan et al., 2023; Shang et al., 2025) in addressing complex cross-sentential relations, even though such relations are plentiful in scientific literature.

Limited work explores generative RE, for instance, El Khettari et al. (2025) used this concept with instruction-tuned LLMs in the microbiome domain. In contrast, Zhang et al. (2025) utilizes entity-pair relation summarizations for triplet fact judgments, whereas the proposed approach focuses primarily on extracting inter-sentential relations and integrating cross-sentential spans of information in an entity-guided summary.

Our core idea is to strategically leverage these generative abilities to enhance zero-shot RE performance, as it remains a valuable strategy for querying LLMs, particularly for non-expert users. In this paper, we formulate two main **research questions:** (i) What are the zero-shot relation extraction capabilities of LLMs (GPT) for cross-sentential RE in the biomedical domain? (ii) How can we simplify and tackle the problem of cross-sentential RE with LLMs' generative capability?

For the current experiment, we used open-sourced GPT-4-0-mini primarily due to its computational efficiency and accessibility, which allowed for extensive experimentation under limited resource constraints, while having core instruction-tuning and generative reasoning capabilities. **RQ1** explores the limitations and potentials of GPT with simple zero-shot prompting in the context of biomedical RE. In **RQ2**, we use LLMs in RE, but not directly as a relation classifier; instead, we explore the *generation capability of LLMs, serving as a summarizer*. In this way, we propose to use GPT's zero-shot capacity to generate an *entity-guided summary that converts cross-sentential relations to intra-sentential relations*. This can also help alleviate the problem of capturing long-range dependencies and complex multi-hop navigation. The current focus is not on maximizing absolute task performance, but instead on better understanding the relative behavior, strengths, and limitations of the approaches under controlled settings.

2 Dataset

We used the BioCreative V *Chemical Disease Relation (CDR)*(Li et al., 2016)[2] and *Gene-Disease Association (GDA)* (Wu et al., 2019) datasets for our experiments. They include abstracts from the scientific biomedical literature. In CDR, we need to identify the binary relations between chemical-induced diseases (CID). The dataset can be considered a good representative of cross-sentential relations, attributed to its complexity and diversity of entity spans, which makes the task challenging. Among the test samples, we extracted 1,800 (negative) and 266 (positive) cross-sentential samples and 748 (positive) and 1,716 (negative) intra-sentential samples. To assess generalizability, we applied the approach to a subset of the GDA dataset (Wu et al., 2019). Since our approach primarily evaluates cross-sentential RE, we specifically selected 1,491 cross-sentential samples (i.e., entity pairs with cross-sentential relations in the given abstract). As cross-sentence and intra-sentence entity pairs can sometimes overlap, following existing works (Christopoulou et al., 2019; Verga et al., 2018; Zhao et al., 2020), we consider cross-sentence subsets to be approximate, rather than strictly disjoint from intra-sentence ones. The details are given in Appendix A.

3 Methods

In this section, we describe the proposed and the baseline approaches used in our controlled experimentation setup.

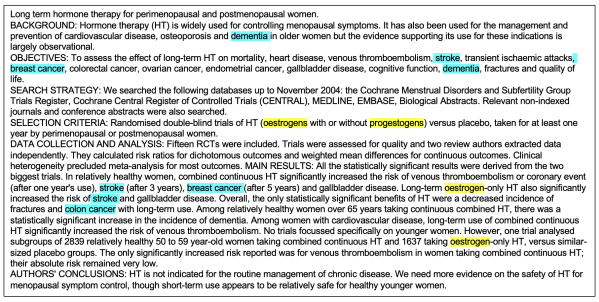
3.1 Direct Zero-shot Learning

As a baseline, we employ a vanilla zero-shot prompting approach to evaluate GPT's capabilities in biomedical RE. We use a simple prompt template that asks GPT to predict whether the entity pair has a relation, given the input text as the context. In this case, the inputs are the abstracts and the corresponding entity pairs, whose relation needs to be classified. For instance, in the CDR dataset, it asks: *"Does the Given chemical entity induce the given disease or not"*.

3.2 Proposed Approach

In the proposed work, we aim to use GPT's zero-shot generative power as an intermediate step to enhance the relation classification pipeline. The

[2] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/



('progestogens', 'stroke')	The use of progestogens was associated with a significant increase in the risk of stroke in women taking hormone therapy.
('progestogens', 'dementia')	Among relatively healthy women over 65 years, the long-term use of combined continuous hormone therapy with progestogens significantly increased the risk of dementia.
(('oestrogen or oestrogens', 'breast cancer')	long-term use of oestrogen or oestrogens was linked to a significant increase in the risk of breast cancer in women
('oestrogen or oestrogens', 'colon cancer')	long-term use of oestrogen or oestrogens was associated with a decreased incidence of colon cancer in women

Figure 1: A CDR abstract with the chemical entities highlighted in yellow and disease entities in blue. The right side shows the (chemical, disease) entity pairs and the corresponding summaries produced by the zero-shot entity-guided summarizer (GPT).

existing models struggle to capture cross-sentential relations for various reasons: The relations that define the entities are not contained within a single sentence. In this case, multi-hop reasoning approaches are needed, which the model may not inherently possess. Secondly, the semantic encodings may not capture sufficient context for identifying such relations due to the presence of long-range dependencies. Thirdly, some sentences or contexts can even act as noise to the model due to the span of entities in multiple sentences. Further, the general path-based approaches used in relation extractions, such as shortest dependency path (SDP) methods, only directly apply to intra-sentential relations.

In the proposed approach, we deviate from the general approach of path-based or multi-hop reasoning (combined with or without encoder/decoder variants) by enabling LLMs' generative capabilities to adapt cross-sentential sentences to intra-sentential ways. Specifically, we want to *convert cross-sentential sentences to a single-sentence entity-guided summary*. Given the impressive results of GPT in generation tasks[3], we used *GPT as a zero-shot entity-guided single-sentence summarizer*. For instance, consider the abstract from the

Dataset	Direct ZSL	Proposed ZSL
CDR (Cross)	0.35	**0.41** (+0.07) ↑
GDA (Cross)	0.49	**0.57** (+0.08) ↑

Table 1: Performance comparison of Direct ZSL and Proposed ZSL on cross-sentential biomedical RE (F-scores).

CDR dataset in Figure 1 (enlarged figures are in Appendix B) with the entity pairs under consideration marked. Here *estrogens* and *progestogens* are the chemical entities, and *{dementia, breast cancer, colon cancer, stroke}* are the disease entities [4]. It can be observed that the relations are cross-sentential, and entities can span across multiple sentences. The entity pairs and the corresponding zero-shot summary generated by GPT-4 are shown in Figure 1.

Considering the entity pair *(progestogens, stroke)*, the relation is not apparent, and proper reasoning is required to classify the relation. Firstly, the model should consider the sentence - *"double-blind trials of HT (oestrogens with or without progestogens)"*, which is the only mention of progestogens in the abstract, and should deduce (entity normalization) that *HT* refers to *"Hormone Therapy"*. Further, it should be related to the sentence - *"In relatively healthy women, combined continuous HT significantly increased the risk of venous thromboembolism or coronary event (after one year's use), stroke (after 3 years), breast cancer (after 5 years) and gallbladder disease."* for capturing the actual relation.

The proposed approach initially uses a prompt to generate a zero-shot entity-guided summary for each cross-sentential entity pair (Figure 3). For instance, for the previous example, we generated a summary that directly conveys the cross-sentential relationship (*"The use of progestogens was associated with a significant increase in the risk of stroke in women taking hormone therapy."*). Similarly, a negative relation is indicated for the entity pair *('estrogen or estrogens,' 'colon cancer')*. These generated summaries were used as inputs for the second step, where the actual relation classifica-

[3] https://github.com/openai

[4] Only the entities required for illustration are highlighted. There are more entity relations in this abstract.

ZSL	Type	F-score	Recall	Precision
Direct ZSL	Overall	0.49	0.83	0.34
	Intra	0.55	0.91	0.40
	Inter/cross	0.35	0.66	0.24
Proposed ZSL	Overall	**0.56**	0.80	0.43
	Intra	**0.65**	0.81	0.54
	Inter/cross	**0.41**	0.75	0.28

Table 2: Comparing proposed ZSL with direct ZSL in CDR dataset

Model	Type	F-score	Recall	Precision
BioBERT_Proposed	Overall	**0.57**	0.56	0.58
	Intra	**0.64**	0.62	0.65
	Inter/cross	**0.41**	0.40	0.41
BioBERT_Baseline	Overall	0.25	0.17	0.44
	Intra	0.36	0.29	0.48
	Inter/cross	0.24	0.21	0.28

Table 3: Fine-tuned Encoder-only Model Performance on CDR dataset

tion is performed (Figure 4). Note that the intra-sentences were directly extracted from the abstract by considering sentences that mention both entities.

4 Results & Comparisons

In Table 1, we report the zero-shot results on the cross-sentential RE in the CDR and GDA datasets obtained with the baseline GPT model (Direct ZSL) and compare them with those of the proposed approach (Proposed ZSL). In the baseline approach, the input is the abstract directly, while, for the proposed approach, it is the entity-guided summary generated by GPT for cross-sentential relations. For intra-sentences, we use the sentences where both entity mentions are present. A 7-point F-score improvement can be observed in the CDR dataset, while in GDA, an 8-point increase is reported.

To analyze the overall improvements, we conducted similar experiments using intra-sentential samples from the CDR dataset. From Table 2, it can be observed that the proposed ZSL approach presents a 7-point improvement compared to baseline or direct ZSL, in terms of overall F-scores. In terms of recall and precision, it can be observed that GPT generally prioritizes recall, which is understandable given its general-purpose nature. In terms of intra-sentential RE, a 10-point improvement is noted. These improvements indicate the scope of utilizing the inherent generative capacity of these LLMs for the downstream tasks, specifically for zero-shot.

Furthermore, we also compare the performance of the proposed entity-guided summaries when used as inputs to fine-tuned encoder-only models. In this case, we fine-tune a BioBERT model using the generated summaries (BioBERT_Proposed) and compare it with the one fine-tuned directly on the abstracts (BioBERT_Baseline). This is the same as the input to the Direct ZSL and Proposed ZSL approaches, which are reported in Table 2. We fine-tune BioBERT with sentence pair classification - where the <*text,entity_pair*> is the input. As discussed, for the baseline, this text is the abstract, and for the proposed, it will be the entity-guided summary. From Table 3, it can be observed that the model appears to capture more accurate information when using the proposed summaries as input. With the cross-sentential RE, BioBERT_Proposed presents significantly better results, with an improvement of almost 17 points over the baseline counterpart. Based on the experimental results reported in Table 3, it is evident that the proposed entity-guided summaries already improve the performance of the simple BioBERT models. More details of experimental settings are in Appendix D. These intrinsic evaluations under controlled settings show that the proposed approach helps the model capture relations more accurately, guiding the LLM to make better predictions. Our analysis suggests that summarizing cross-sentential information into a single sentence enables simpler, more effective representations, which in turn support more accurate scientific information extraction.

5 Conclusions

In the proposed work, we aim to evaluate the zero-shot capabilities of GPT in biomedical relation extraction, with a focus on cross-sentential relations. We utilized the chemical-induced disease and gene-disease association datasets, which comprise complex inter-sentential spans of entity relations, as a representative dataset. We observed that GPT, in its zero-shot capacity, has considerable scope for improvement in capturing these relations. A novel approach is proposed to utilize the generative capabilities of GPT as an intermediate step in the relation extraction pipeline by using it as a zero-shot entity-guided summarizer. This is used to encapsulate information on cross-sentential relations and convert these relations into intra-sentential ones. We observed a good performance improvement compared to baseline zero-shot performances. We believe that the proposed direction has considerable potential for exploration, where, instead of using GPT directly as a downstream classifier, it would be more reasonable to exploit its inherent generative ability by mapping it to intermediate steps in a logical manner.

6 Related Works

In the field of structured IE from scientific literature, recently LLMs are used widely (Dagdelen et al., 2024; Li et al., 2024; Garcia et al., 2024). The approaches range from simple feature-based extractions to transformer-based to current Large Language Model (LLM) based approaches. A unified schema representation was proposed in Li et al. (2023) to encourage LLMs to follow schemas, learn easily, and extract structured knowledge accurately.

In the existing literature, a wide range of approaches and studies consider the problem of biomedical information extractions (Sciannameo et al., 2024; Fornasiere et al., 2024; Reichenpfader et al., 2024). In the task of relation extractions, Zhang et al. (2018) proposed a hybrid model that uses Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) with Shortest Dependency Features (SDP). An SDP-based feature extraction for candidate cross-sentential sample extractions, coupled with BioBERT models, was presented by Kanjirangat and Rinaldi (2021). The use of biomedical ontologies to enhance neural network knowledge is another direction (Sousa et al., 2020; Sänger and Leser, 2025; Liu et al., 2025). Another promising direction was to explore graph-based models, graph LSTMs (Peng et al., 2017), graph kernels (Panyam et al., 2018), graph CNN with multi-head attentions (Zhao et al., 2021), and multi-view GNNs (Al-Sabri et al., 2022). BERT and its variants have been widely used for biomedical RE tasks (Thillaisundaram and Togia, 2019; Bhasuran, 2022; Su and Vijay-Shanker, 2020, 2022). However, the complex task of cross-sentential RE necessitated more sophisticated approaches. For instance, Wei and Li (2022) proposed a sequence-aware graph model with adaptive margin loss, while Zhu et al. (2024) leveraged dependency and constituency information using Tree-LSTM, GNN, and BERT models.

Generative models are now being explored in biomedical RE, where their performance has been reported to vary based on the complexity of the dataset and task at hand Zhang et al. (2024); Asada and Fukuda (2024). Some of the findings reported good performances, but were limited to intra-sentential relations. A few studies (e.g., (El Khettari et al., 2025)) have explored generative approaches to relation extraction (RE) using instruction-tuned large language models (LLMs). In contrast, (Zhang et al., 2025) focuses on leveraging entity-pair relation summarization for triplet fact evaluation. In our proposed approach, we primarily address inter-sentential relation extraction, emphasizing the integration of cross-sentential contextual spans within an entity-guided summarization framework.

In the proposed work, we focus on exploring the zero-shot capability of GPT in cross-sentential RE. Moving a step further, we propose an approach to possibly utilize the generative capability of GPT in the RE pipeline, which is the inherent potential of generative models. This deviates from the general trend of using these generative models directly for classifications, a use case that does not fully align with their intrinsic generative nature.

7 Limitations

The proposed approach could propagate errors from the summarization module, as we introduce it as an intermediate path in the relation extraction pipeline. An explicit evaluation of the zero-shot summarization component is challenging, which limits the understanding of the summarizer's performance. Currently, the experiments are done only on the CDR and GDA biomedical datasets. These could be considered as representative datasets for

complex cross-sentential relations; however, a proper generalization of the proposed approach has to be verified by extending the experiments to other datasets with cross-sentential relations, Chemical Reaction (CHR) dataset (Peng et al., 2017), or general-purpose datasets, such as DocRed (Yao et al., 2019), Codred (Yao et al., 2021), CrossRE (Bassignana and Plank, 2022), etc. Furthermore, GPT responses can be limited by multiple factors, including sensitivity to prompts, context, post-processing, controversies, ambiguities, efficiency, and costs (Kocoń et al., 2023). In general, the low performance of GPT models can be attributed to several factors, including the lack of domain-specific training, entity disambiguation issues in biomedical data, and the need for multi-hop reasoning to address inter-sentential relations. While refining prompts can mitigate some issues, prompt sensitivity remains a challenge. Soft prompting techniques offer a potential solution to improve robustness, though naive zero-shot prompting still holds value for user-centric applications across various domains. We also have the scope of experimenting with different LLMs (open-sourced). Finally, considering the state-of-the-art approaches, we still have considerable scope for improvement, even though our approach focuses on zero-shot capability.

Acknowledgments

The work of Vani Kanjirangat has been supported by HASLER STIFTUNG through the project "RING-P- Relation Extractions at Inter-sentential N-ary levels using Graph Prompting", project No. 24002.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Zabir Al Nazi, Md Rajib Hossain, and Faisal Al Mamun. 2025. Evaluation of open and closed-source llms for low-resource language with zero-shot, few-shot, and chain-of-thought prompting. *Natural Language Processing Journal*, 10:100124.

Raeed Al-Sabri, Jianliang Gao, Jiamin Chen, Babatounde Moctard Oloulade, and Tengfei Lyu. 2022. Multi-view graph neural architecture search for biomedical entity and relation extraction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1221–1233.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2021. On the multilingual capabilities of very large-scale english language models. *arXiv preprint arXiv:2108.13349*.

Masaki Asada and Ken Fukuda. 2024. Enhancing relation extraction from biomedical texts by large language models. In *International Conference on Human-Computer Interaction*, pages 3–14. Springer.

Samy Ateia and Udo Kruschwitz. 2023. Is chatgpt a biomedical expert?–exploring the zero-shot performance of current gpt models in biomedical tasks. *arXiv preprint arXiv:2306.16108*.

Elisa Bassignana and Barbara Plank. 2022. Crossre: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Balu Bhasuran. 2022. Biobert and similar approaches for relation extraction. In *Biomedical Text Mining*, pages 221–235. Springer.

Jing Bi, Ziqi Wang, Haitao Yuan, Xiankun Shi, Ziyue Wang, Jia Zhang, MengChu Zhou, and Rajkumar Buyya. 2025. Large ai models and their applications: Classification, limitations, and potential solutions. *Software: Practice and Experience*, 55(6):1003–1017.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, and 1 others. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1):3280.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2024. Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4925–4936.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418.

Oumaima El Khettari, Solen Quiniou, and Samuel Chaffron. 2025. Summarization for generative relation extraction in the microbiome domain. In *Actes de l'atelier Traitement du langage médical à l'époque des LLMs 2025 (MLP-LLM)*, pages 68–82.

Raffaello Fornasiere, Nicolò Brunello, Vincenzo Scotti, Mark James Carman, and 1 others. 2024. Medical information extraction with large language models. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 1–10. Association for Computational Linguistics.

Gabriel Lino Garcia, Joao Renato Ribeiro Manesco, Pedro Henrique Paiola, Lucas Miranda, Maria Paola de Salvo, and Joao Paulo Papa. 2024. A review on scientific knowledge extraction using large language models in biomedical sciences. *arXiv preprint arXiv:2412.03531*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of chatgpt on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. *arXiv preprint arXiv:2306.04504*.

Mohamad Yaser Jaradeh, Kuldeep Singh, Markus Stocker, Andreas Both, and Sören Auer. 2023. Information extraction pipelines for knowledge graphs. *Knowledge and Information Systems*, 65(5):1989–2016.

Vani Kanjirangat, Alessandro Antonucci, and Marco Zaalon. 2024. On the limitations of zero-shot classification of causal relations by llms (work in progress). *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Vani Kanjirangat and Fabio Rinaldi. 2021. Enhancing biomedical relation extraction with transformer models using shortest dependency path features and triplet information. *Journal of biomedical informatics*, 122:103893.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: a comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, and 1 others. 2023. Chatgpt: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Md Tahmid Rahman Laskar, Israt Jahan, Elham Dolatabadi, Chun Peng, Enamul Hoque, and Jimmy Huang. 2025. Improving automatic evaluation of large language models (llms) in biomedical relation extraction via llms-as-the-judge. *arXiv preprint arXiv:2506.00777*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *arXiv preprint arXiv:2005.00633*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. *arXiv preprint arXiv:2310.05028*.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xiang Li, Zhilei Hu, and 1 others. 2024. Knowcoder: Coding structured knowledge into llms for universal information extraction. *arXiv preprint arXiv:2403.07969*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

Xiaoyong Liu, Xin Qin, Chunlin Xu, and Huihui Li. 2025. A knowledge-enhanced model with syntactic-aware attentive graph convolutional network for biomedical entity and relation extraction. *International Journal of Machine Learning and Cybernetics*, 16(1):583–598.

Hariharan Manikandan, Yiding Jiang, and J Zico Kolter. 2023. Language models are weak learners. *Advances in Neural Information Processing Systems*, 36:50907–50931.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Nagesh C Panyam, Karin Verspoor, Trevor Cohn, and Kotagiri Ramamohanarao. 2018. Exploiting graph kernels for high performance biomedical relation extraction. *Journal of biomedical semantics*, 9(1):7.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 7(1):222.

Mario Sänger and Ulf Leser. 2025. Knowledge-augmented pre-trained language models for biomedical relation extraction. *arXiv preprint arXiv:2505.00814*.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 447–451.

Veronica Sciannameo, Daniele Jahier Pagliari, Sara Urru, Piercesare Grimaldi, Honoria Ocagli, Sara Ahsani-Nasab, Rosanna Irene Comoretto, Dario Gregori, and Paola Berchialla. 2024. Information extraction from medical case reports using openai instructgpt. *Computer methods and programs in biomedicine*, 255:108326.

Yufei Shang, Yanrong Guo, Shijie Hao, and Richang Hong. 2025. Biomedical relation extraction via adaptive document-relation cross-mapping and concept unique identifier. *arXiv preprint arXiv:2501.05155*.

Lei Shu, Hu Xu, Bing Liu, and Jiahua Chen. 2022. Zero-shot aspect-based sentiment analysis. *arXiv preprint arXiv:2202.01924*.

Diana Sousa, Andre Lamurias, and Francisco M Couto. 2020. Using neural networks for relation extraction from biomedical literature. In *Artificial Neural Networks*, pages 289–305. Springer.

Peng Su and K Vijay-Shanker. 2020. Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism. In *2020 ieee international conference on bioinformatics and biomedicine (bibm)*, pages 2522–2529. IEEE.

Peng Su and K Vijay-Shanker. 2022. Investigation of improving the pre-training and fine-tuning of bert model for biomedical relation extraction. *BMC bioinformatics*, 23(1):120.

Ashok Thillaisundaram and Theodosia Togia. 2019. Biomedical relation extraction with pre-trained language representations and minimal task-specific architecture. *arXiv preprint arXiv:1909.12411*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Ying Wei and Qi Li. 2022. Sagdre: Sequence-aware graph-based document-level relation extraction with adaptive margin loss. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2000–2008.

Ling-I Wu, Yuxin Su, and Guoqiang Li. 2025. Zero-shot construction of chinese medical knowledge graph

with gpt-3.5-turbo and gpt-4. *ACM Transactions on Management Information Systems*, 16(2):1–17.

Ye Wu, Ruibang Luo, Henry CM Leung, Hing-Fung Ting, and Tak-Wah Lam. 2019. Renet: A deep learning approach for extracting gene-disease associations from literature. In *International Conference on Research in Computational Molecular Biology*, pages 272–284. Springer.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2023. Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-aware language modeling. *arXiv preprint arXiv:2306.11489*.

Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. 2021. Codred: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4452–4472.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Fu Zhang, Hongsen Yu, Jingwei Cheng, and Huangming Xu. 2025. Entity pair-guided relation summarization and retrieval in LLMs for document-level relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4022–4037, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeffrey Zhang, Maxwell Wibert, Huixue Zhou, Xueqing Peng, Qingyu Chen, Vipina K Keloth, Yan Hu, Rui Zhang, Hua Xu, and Kalpana Raja. 2024. A study of biomedical relation extraction using gpt models. *AMIA Summits on Translational Science Proceedings*, 2024:391.

Yijia Zhang, Hongfei Lin, Zhihao Yang, Jian Wang, Shaowu Zhang, Yuanyuan Sun, and Liang Yang. 2018. A hybrid model based on neural networks for biomedical relation extraction. *Journal of biomedical informatics*, 81:83–92.

D. Zhao, J. Wang, Y. Zhang, X. Wang, H. Lin, and Z. Yang. 2020. Incorporating representation learning and multihead attention to improve biomedical cross-sentence n-ary relation extraction. *BMC Bioinformatics*, 21(1):312.

Di Zhao, Jian Wang, Hongfei Lin, Xin Wang, Zhihao Yang, and Yijia Zhang. 2021. Biomedical cross-sentence relation extraction via multihead attention and graph convolutional networks. *Applied Soft Computing*, 104:107230.

Xudong Zhu, Zhao Kang, and Bei Hui. 2024. Fcds: Fusing constituency and dependency syntax into document-level relation extraction. *arXiv preprint arXiv:2403.01886*.

A Dataset Details

The CDR dataset annotation identifies entities that hold a relation (class 1/positive), and all remaining entity pairs fall into the negative category/class 0. The CID relations can be either intra-sentential or cross-sentential. There are no mention-level annotations in the CDR dataset. Hence, we can use the entire abstract as the context or deduce methodologies to extract the context that can convey possible relations (based on the presence of entities).

The Gene–Disease Associations (GDA) dataset is a large-scale biomedical corpus constructed from MEDLINE abstracts using distant supervision. In line with Christopoulou et al. (2019), we partition the data into 23,353 documents for training and 5,839 documents for development. The task is formulated as a binary classification problem, where the goal is to determine whether a given gene and disease entity pair is associated or not. A notable characteristic of the dataset is that many associations span across multiple sentences, which makes it particularly suitable for assessing methods that aim to capture long-range dependencies and inter-sentential relations.

B Methods

The enlarged examples for CDR abstracts and the entity guided summaries are shown in Figures 2a and 2b.

C Prompt Templates

The prompt templates for vanilla and the proposed approaches are given in Figures 3 and 4.

D Experiments

We used GPT4-o-mini [5] for our experiments (Approximately 150 USD was spent). The experiments were conducted on an HPC cluster with 1 GPU (NVIDIA A100 80GB PCI). For BERT-based experiments, we used BioBERT v1.1 (+ PubMed 1M), which refers to the BioBERT model trained on PubMed for 1M steps as the pre-trained model. The experiments were done using PyTorch HuggingFace implementations [6] by fine-tuning the model on the respective datasets. The model is fine-tuned for 10 epochs, using the Adam optimizer and a learning rate of 2e-5 on the training data.

[5] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

[6] https://github.com/huggingface/transformers

Long term hormone therapy for perimenopausal and postmenopausal women.
BACKGROUND: Hormone therapy (HT) is widely used for controlling menopausal symptoms. It has also been used for the management and prevention of cardiovascular disease, osteoporosis and dementia in older women but the evidence supporting its use for these indications is largely observational.
OBJECTIVES: To assess the effect of long-term HT on mortality, heart disease, venous thromboembolism, stroke, transient ischaemic attacks, breast cancer, colorectal cancer, ovarian cancer, endometrial cancer, gallbladder disease, cognitive function, dementia, fractures and quality of life.
SEARCH STRATEGY: We searched the following databases up to November 2004: the Cochrane Menstrual Disorders and Subfertility Group Trials Register, Cochrane Central Register of Controlled Trials (CENTRAL), MEDLINE, EMBASE, Biological Abstracts. Relevant non-indexed journals and conference abstracts were also searched.
SELECTION CRITERIA: Randomised double-blind trials of HT (oestrogens with or without progestogens) versus placebo, taken for at least one year by perimenopausal or postmenopausal women.
DATA COLLECTION AND ANALYSIS: Fifteen RCTs were included. Trials were assessed for quality and two review authors extracted data independently. They calculated risk ratios for dichotomous outcomes and weighted mean differences for continuous outcomes. Clinical heterogeneity precluded meta-analysis for most outcomes. MAIN RESULTS: All the statistically significant results were derived from the two biggest trials. In relatively healthy women, combined continuous HT significantly increased the risk of venous thromboembolism or coronary event (after one year's use), stroke (after 3 years), breast cancer (after 5 years) and gallbladder disease. Long-term oestrogen-only HT also significantly increased the risk of stroke and gallbladder disease. Overall, the only statistically significant benefits of HT were a decreased incidence of fractures and colon cancer with long-term use. Among relatively healthy women over 65 years taking continuous combined HT, there was a statistically significant increase in the incidence of dementia. Among women with cardiovascular disease, long-term use of combined continuous HT significantly increased the risk of venous thromboembolism. No trials focussed specifically on younger women. However, one trial analysed subgroups of 2839 relatively healthy 50 to 59 year-old women taking combined continuous HT and 1637 taking oestrogen-only HT, versus similar-sized placebo groups. The only significantly increased risk reported was for venous thromboembolism in women taking combined continuous HT; their absolute risk remained very low.
AUTHORS' CONCLUSIONS: HT is not indicated for the routine management of chronic disease. We need more evidence on the safety of HT for menopausal symptom control, though short-term use appears to be relatively safe for healthy younger women.

(a) A CDR abstract with chemical entities highlighted in yellow and disease entities in blue.

('progestogens', 'stroke')	The use of progestogens was associated with a significant increase in the risk of stroke in women taking hormone therapy.
('progestogens', 'dementia')	Among relatively healthy women over 65 years, the long-term use of combined continuous hormone therapy with progestogens significantly increased the risk of dementia.
(('oestrogen or oestrogens', 'breast cancer')	long-term use of oestrogen or oestrogens was linked to a significant increase in the risk of breast cancer in women
('oestrogen or oestrogens', 'colon cancer')	long-term use of oestrogen or oestrogens was associated with a decreased incidence of colon cancer in women

(b) (chemical, disease) entity pairs and the corresponding summaries produced by the proposed zero-shot entity-guided summarizer (GPT).

Figure 2: Illustration of the proposed approach: (a) a CDR abstract with highlighted entities; (b) entity pairs and generated summaries.

```
system_msg="You are a helpful summarizer who performs an entity-guided summarization
based on given entity pairs."

"Based on the given text and entity pair, perform an entity-guided single-sentence
summarization of the text.
Give focus on the terms or keywords that can distinguish whether
the given entities can have a relation or not?.
The output should be a single sentence with the entity mentions in it."""

instructions_msg="You are a helpful summarization assistant. You will be provided
with the text and a (chemical,disease) entity pair.

 Text:<Text>{text}</Text>
 Entity_pair:<Text>{ent_pair}</Text>

Provide the final summary within the tags <summary> </summary>."

 '''
```

Figure 3: A zero-shot prompt-template for an Entity-Guided Summarizer(The prompts will vary slightly based on the experimental datasets. This prompt is tailored for the CDR dataset).

```
system_msg = "You are a helpful medical assistant who tells whether a given chemical

induce a given disease or not."

instructions_msg= You will be provided with the text and a list of chemical and
disease entities.

            Text: <Text>{text}</Text>
            Chemical_list:{chem}
            Disease_list:{dis}

For each pair of (chemical, disease), predict whether the chemical induce the disease

or not?.
You should predict 1 if the chemical induce the disease and 0 if not.
Your response should be only based on the given text.

Provide all your final answers within the tags <Answer> </Answer> with entity pairs
expressed as a tuple with its corresponding prediction."
```

Figure 4: A zero-shot prompt-template for a Relation Classification (The prompts will vary slightly based on the experimental datasets. This prompt is tailored for the CDR dataset).