GLiM: Integrating Graph Transformer and LLM for Document-Level Biomedical Relation Extraction with Incomplete Labeling

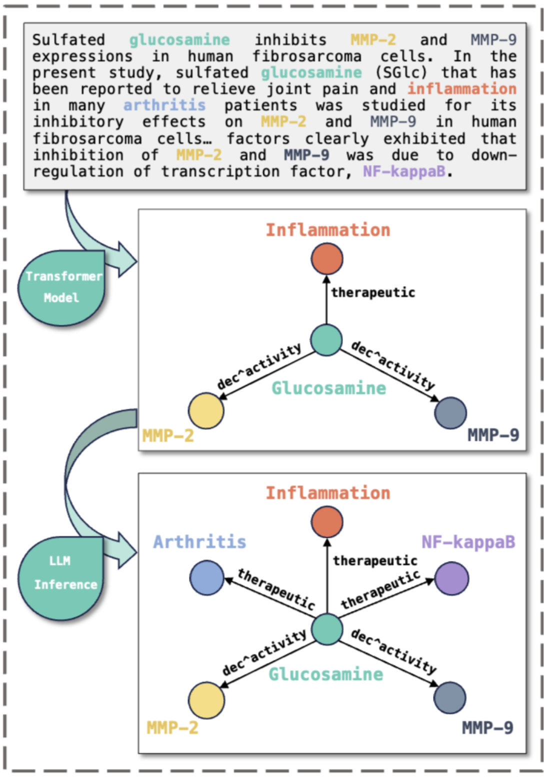
Hao Fang[1], Yuejie Zhang[1*], Rui Feng[1,2*], Yingwen Wang[2], Qing Wang[2], Wen He[2], Xiaobo Zhang[2] Tao Zhang[3*], Shang Gao[4]

[1]College of Computer Science and Artificial Intelligence, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, [2]National Children's Medical Center, Children's Hospital of Fudan University, Shanghai, [3]School of Information Management and Engineering, Shanghai University of Finance and Economics, [4]School of Information Technology, Deakin University

{hfang23, 22211240021}@m.fudan.edu.cn, {yjzhang, fengrui, yingwenwang, hewen}@fudan.edu.cn, zhangxiaobo0307@163.com, taozhang@mail.shufe.edu.cn, shang.gao@deakin.edu.au

Abstract

Document-level relation extraction (DocRE) identifies relations between entities across an entire document. However, as the number and complexity of entities and entity-pair relations grow, the problem space expands quadratically, causing incomplete annotations and frequent false negatives, especially in biomedical datasets due to high construction costs. This leads to low recall in real-world scenarios. To address this, we propose GLiM, a novel framework that reduces the problem space using a graph-enhanced Transformer-based model and leverages large language models (LLMs) for reasoning. GLiM employs a cascaded approach: first, a graph-enhanced Transformer processes entity-pair relations with finer granularity by dynamically adjusting the graph size based on the number of entities; then, LLM inference handles challenging cases. Experiments show that GLiM boosts average recall and F1 scores by +6.34 and +4.41, respectively, outperforming state-of-the-art models on biomedical benchmarks. These results demonstrate the effectiveness of combining graph-enhanced Transformers with LLM inference for biomedical DocRE. Code will be released at https://github.com/HaoFang10/GLiM.



Sulfated glucosamine inhibits MMP-2 and MMP-9 expressions in human fibrosarcoma cells. In the present study, sulfated glucosamine (SGlc) that has been reported to relieve joint pain and inflammation in many arthritis patients was studied for its inhibitory effects on MMP-2 and MMP-9 in human fibrosarcoma cells… factors clearly exhibited that inhibition of MMP-2 and MMP-9 was due to down-regulation of transcription factor, NF-kappaB.

Figure 1: Illustration of biomedical document-level relation extraction using our cascaded framework on ChemDisGene dataset.

1 Introduction

Relation Extraction (RE) plays a crucial role in Information Extraction (IE), aiming to identify relations between entities in a given text. Document-level relation extraction (DocRE) extends this task by identifying relations between all entity pairs across an entire document. DocRE is of paramount importance for performance enhancement of downstream applications such as question answering, knowledge graph construction, and recommendation systems.

Compared to sentence-level relation extraction, which focuses on identifying entity relations within a single sentence, DocRE faces greater challenges in dataset quality. As the number of entities increases, the problem space grows quadratically, complicating the annotation process. Moreover, DocRE needs to handle longer contexts and more complex relation types, leading to inevitable issues such as incomplete labeling and a high prevalence of false negatives. These challenges are especially prominent in biomedical DocRED datasets, such as ChemDisGene. In the automatically annotated training set for ChemDisGene, the average number of relations per document is only about one-third of

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 14131–14146
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

that in the expert-annotated test set. Consequently, models trained on such datasets often suffer from low recall in real-world applications, making this a critical issue to address.

With the remarkable capabilities demonstrated by large language models (LLMs) across various domains, reframing relation extraction as a text generation task and leveraging LLMs has emerged as a promising approach to overcoming the aforementioned challenges. However, the relatively underwhelming performance of LLMs in RE tasks is largely due to the massive problem space inherent to DocRE. Therefore, our focus is on reducing the problem space while leveraging the intrinsic knowledge of LLMs to recover false negative relations without compromising model performance.

In light of the above, we propose an innovative cascaded framework named **Graph LLM integration Model (GLiM)**. GLiM first employs a Transformer-based model for initial processing, which not only produces preliminary results but also significantly reduces the problem space. This reduction allows subsequent LLM to fully exploit its intrinsic knowledge, thereby improving both recall and F1 scores. Furthermore, to enhance the first-step model's ability to capture finer-grained and complex entity relations, we incorporate adaptively-adjusted graph structures into the Transformer-based model. This enhancement enables GLiM to achieve state-of-the-art (SOTA) performance on the biomedical BioRED and ChemDisGene benchmark datasets, even without utilizing LLMs. When LLMs are used, the framework achieves further significant performance gains. An example of GLiM is shown in Figure 1.

The main contributions of this paper include:

(1) **A novel framework for biomedical DocRE:** GLiM reduces the problem space by decomposing the DocRE task into cascading processing steps. The initial step uses a graph-enhanced Transformer model, followed by LLM inference. This approach substantially enhances performance on incompletely-annotated biomedical DocRE tasks.

(2) **Adaptively-adjusted Graph structure:** We integrate a sequential graph structure into the Transformer-based model and dynamically adjust the graph structure size based on the number of entities, achieving SOTA performance on BioRED and ChemDisGene benchmark datasets even without relying on LLMs.

(3) **Extensive experimental validation:** Experimental results across multiple datasets demonstrate the complementary effectiveness of Transformer-based models and LLMs. GLiM achieves average recall and F1 score improvements of +6.34 and +4.41 on ChemDisGene and BioRED benchmark datasets, respectively. On the general-domain DocRED dataset, GLiM maintains a competitive F1 score while achieving SOTA recall performance.

2 Related Work

2.1 Document-level Relation Extraction

Document-level relation extraction (DocRE) differs from sentence-level relation extraction because it requires handling longer contexts and more complex entity-pair relations. Moreover, a document with N entities results in $N(N-1)$ possible relation predictions, causing the problem space to grow quadratically as the number of entities increases. This leads to significant costs in constructing a complete DocRE dataset. Consequently, DocRE tasks commonly face challenges such as multi-label relations, long-tail relation distributions, and incomplete labeling with false negatives.

To address multi-label and multi-entity challenges, Zhou et al. (2021) proposed an adaptive thresholding and local context pooling model (ATLOP). Zhang et al. (2021) introduced a document U-shaped network (DocuNet), which regards DocRE as a semantic segmentation task. DocuNet predicts relation types between entity pairs via pixel-level mask prediction and uses a balanced softmax approach to handle relation distribution imbalances. Similarly, several methods introduced novel loss functions (Tan et al. (2022a); Wang et al. (2022)) to address these issues.

2.2 Document-level Biomedical RE with Incomplete Labeling

In the biomedical field, the high cost of constructing DocRE datasets intensifies the issue of false negatives caused by incomplete labeling. To address this problem, mainstream methods include knowledge distillation (Tan et al. (2022a); Ma et al. (2023); Gao et al. (2024)), negative sampling (Li et al., 2021) to avoid overfitting on false negatives, and Positive-Unlabeled (PU) learning (Wang et al. (2024); Wang et al. (2022)), which adjusts the loss weight based on the distribution of each relation class. Moreover, Tan et al. (2023) proposed a class-adaptive re-sampling self-training framework that iteratively samples pseudo-labels for training each

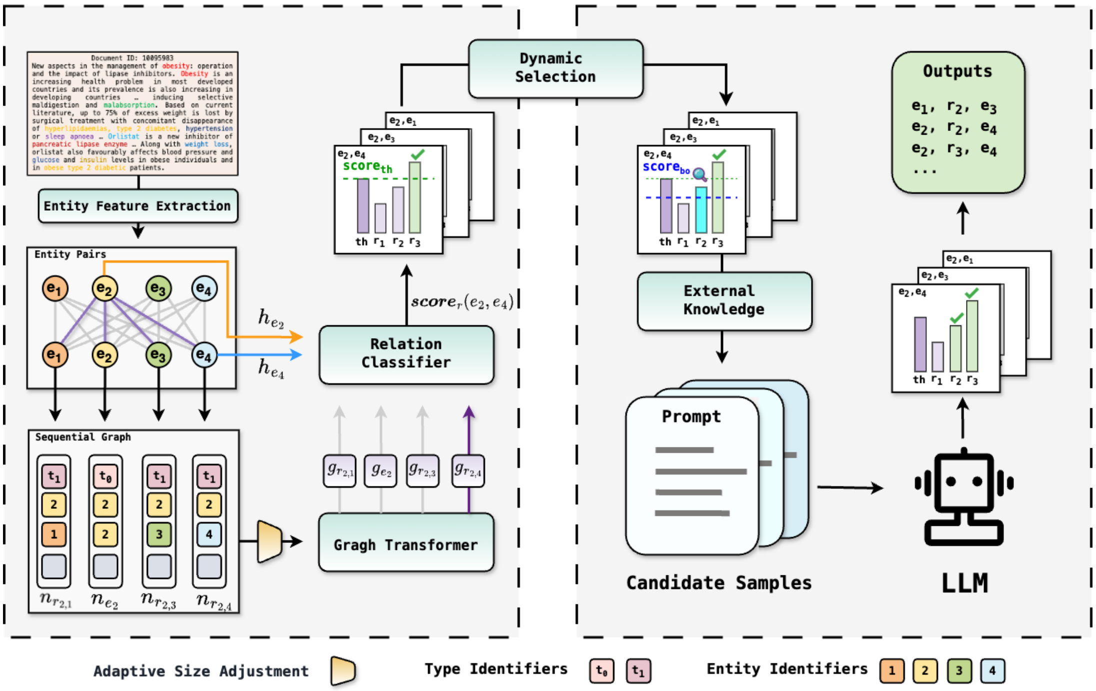

Document ID: 10095983
New aspects in the management of obesity: operation and the impact of lipase inhibitors. Obesity is an increasing health problem in most developed countries and its prevalence is also increasing in developing countries … inducing selective maldigestion and malabsorption. Based on current literature, up to 75% of excess weight is lost by surgical treatment with concomitant disappearance of hyperlipidaemias, type 2 diabetes, hypertension or sleep apnoea … Orlistat is a new inhibitor of pancreatic lipase enzyme … Along with weight loss, orlistat also favourably affects blood pressure and glucose and insulin levels in obese individuals and in obese type 2 diabetic patients.

Figure 2: Overall framework of GLiM.

class.

Unlike these methods, our approach fully leverages the intrinsic knowledge capabilities of LLMs to mitigate distribution shifts between training and test sets while addressing false negatives. This strategy ultimately improves both the recall and F1 scores.

2.3 Relation Extraction with LLMs

Large language models (LLMs) have demonstrated significant potential across various domains. Several recently proposed relation extraction (RE) methods advocate leveraging LLMs' string-based encoding to perform RE in an interpretable manner.

Paolini et al. (2021) proposed reframing structured prediction tasks, including RE, as sequence-to-sequence problems. Wan et al. (2023) further advanced this approach by prompting GPT-3 for RE. Expanding this research direction, Wadhwa et al. (2023) explored few-shot learning within LLMs for sentence-level RE. Meanwhile, Xue et al. (2024) introduced AutoRE, a model that decomposes DocRE into three sub-tasks: extracting relations, entity heads, and entity tails. AutoRE fine-tunes LLMs with Quantized Low-Rank Adapter (QLoRA) to enhance performance. Similarly, Wei et al. (2024) proposed ChatIE, which decomposes the complex RE process into multiple rounds of question-answering, with outputs combined into a final structured format.

However, despite these advancements, LLM performance in RE tasks still lags behind that of BERT-based models. Han et al. (2023) found that ChatGPT struggles to comprehend subject-object relations in RE tasks. Likewise, Li et al. (2023) pointed out that in standard information extraction tasks, ChatGPT typically underperforms compared to BERT-based models. In DocRE, LLMs' performance remains unsatisfactory, largely due to the vast potential problem space.

To overcome this limitation, our approach utilizes a pre-trained model for the first step of processing, significantly reducing the problem space. This enables LLM inference in the second step to fully exploit its reasoning capabilities, identifying false negative relations and further enhancing model performance.

3 Methodology

3.1 Problem Definition

In the DocRE task, we use a model to handle document D that consists of M sentences, N entities, and R candidate relation types. Given a known set of entities $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$, the task is to determine whether each entity pair (e_i, e_j) in the document D has a relation r from the candidate relation set $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$. Each entity pair may be associated with multiple relations, and entities may appear multiple times in D under different mentions. The final problem space is $N \times (N-1) \times R$.

Unlike sentence-level RE tasks, document-level RE involves more entities and candidate relations. Additionally, the problem space grows quadratically as N^2, increasing the complexity of entity interactions and causing incomplete annotations and frequent false negatives.

To address the challenges of DocRE, we propose GLiM (Graph LLM Integration Model), a novel cascaded framework illustrated in Figure 2. GLiM first integrates a Transformer-based model with an adaptively-adjusted graph structure for fine-grained entity-pair relation processing. To handle complex cases, GLiM dynamically selects high-probability candidate relations using the Transformer's outputs and refines them through LLM inference, leveraging both internal and external knowledge. The final results are obtained by merging the outputs from both steps, significantly improving relation extraction performance.

3.2 Model Based on Transformers and Graph

To improve the model's ability to capture entity-pair relations at a finer granularity, we introduce a **graph model** within the backbone of a **Transformer-based pretrained model**. Inspired by the TokenGT model of Kim et al. (2022), we adopt a Transformer-based Tokenized Graph model. Unlike traditional message-passing Graph Neural Networks (GNNs), Transformer-based graph models have demonstrated greater expressiveness. Therefore, our proposed model is entirely Transformer-based, ensuring unified information fusion throughout the architecture.

Entity Feature Extraction

When constructing graph nodes for the graph model, the first step is to obtain the feature representation of the entities. Based on the ATLOP model, a classic Transformer-based approach for entity representation, given a document $D = [x_t]_{t=1}^{l}$, where l represents the total number of tokens, special symbols "*" are inserted at the beginning and end of each entity mention to mark its position. The document is then fed into a pretrained language model (PLM), and the embedding of the initial "*"is used as the mention embedding:

$$[h_1, h_2, ..., h_l] = PLM([x_1, x_2, ..., x_l]). \quad (1)$$

For an entity e_i with multiple mentions, the logsumexp pooling (Jia et al., 2019) is applied to obtain the entity embedding h_{e_i}.

$$h_{e_i} = \log \sum_{k=1}^{N_{e_i}} \exp(h_{m_k^i}), \quad (2)$$

where $h_{m_k^i}$ represents the embedding of the k-th mention of entity e_i, and N_{e_i} denotes the number of mentions of entity e_i.

Graph Representation Learning

After obtaining the feature representation of the entities, we construct the graph nodes. Following TokenGT, we treat both entities and edges as graph nodes, and introduce Entity Identifiers and Type Identifiers in nodes to fully represent the graph structure, thereby enhancing the graph learner.

$$\begin{aligned} n_{e_i} &= h_{e_i} + [p_{e_i}; p_{e_i}] + type_0 \\ n_{r_{i,j}} &= h_{e_i} + h_{e_j} + [p_{e_i}; p_{e_j}] + type_1 \end{aligned}, \quad (3)$$

where n_{e_i} and $n_{r_{i,j}}$ represent the embeddings of entity e_i node and edge $r_{i,j}$ node, respectively. p_{e_i} and p_{e_j} are orthogonal vectors, referred to as Entity identifiers $\in \mathcal{R}^{d/2}$, and are initialized using a random Gaussian matrix. $type_0$ and $type_1$ are trainable feature vectors called Type Identifiers, each in $\mathcal{R}^d$, initialized to 0 for entity nodes, and 1 for edge nodes, respectively. $h_{e_i} \in \mathcal{R}^d$ represents the embedding of entity e_i.

Since the number of entities varies across different documents, the sequence length input to the Tokenized Graph Transformer also varies. We employ a dynamic thresholding method to pad the input length to the smallest possible upper bound, choosing from $[128, 256, 512, 1024, 1296, 2048, 4096]$. Using a consistent input length allows for better learning of node representation information, which is then processed through a Graph Transformer (GT) consisting of an alternating stack of multi-head self-attention (MSA) layers and feed-forward MLP layers to obtain the final graph node representations $\boldsymbol{g}$.

$$\boldsymbol{g} = GT(\boldsymbol{n}), \quad (4)$$

where $\boldsymbol{n}$ represents the sequence of input graph nodes $[n_{e_1}, \ldots, n_{e_N}, n_{r_{1,2}}, \ldots]$, and $\boldsymbol{g}$ represents corresponding output sequence $[g_{e_1}, \ldots, g_{e_N}, g_{r_{1,2}}, \ldots]$.

Feature Fusion and Classification

After obtaining the enhanced edge node features $g_{r_{i,j}}$ from the GT, we fuse these features with the local context embeddings $\boldsymbol{c}^{(i,j)}$ obtained from ATLOP, along with the original entity embeddings. This results in the final representations of each head-tail entity pair, $\boldsymbol{z}_i^{(i,j,r)}$ and $\boldsymbol{z}_j^{(i,j,r)}$, which are then input into the classifier to predict every relation r score as $\boldsymbol{score}_r(e_i, e_j)$:

$$\begin{aligned} \boldsymbol{z}_i^{(i,j,r)} &= \tanh\left(\boldsymbol{W}_i \boldsymbol{h}_{e_i} + \boldsymbol{W}_{c_1}\boldsymbol{c}^{(i,j)} + \boldsymbol{W}_{g_1}\boldsymbol{g}_{r_{i,j}}\right) \\ \boldsymbol{z}_j^{(i,j,r)} &= \tanh\left(\boldsymbol{W}_j \boldsymbol{h}_{e_j} + \boldsymbol{W}_{c_2}\boldsymbol{c}^{(i,j)} + \boldsymbol{W}_{g_2}\boldsymbol{g}_{r_{i,j}}\right)' \end{aligned} \tag{5}$$

where $\boldsymbol{z}_i^{(i,j,r)}$ is entity e_i representation after feature fusion. $\boldsymbol{W}_i$, $\boldsymbol{W}_{c_1}$, $\boldsymbol{W}_{g_1}$, $\boldsymbol{W}_j$, $\boldsymbol{W}_{c_2}$, and $\boldsymbol{W}_{g_2}$ are model parameters ($\in \mathcal{R}^{d\times d}$).

The final score of relation r between entities e_i and e_j is computed as:

$$\boldsymbol{score}_r(e_i, e_j) = \sigma(\boldsymbol{z}_i^{(i,j,r)\top}\boldsymbol{W}_r\boldsymbol{z}_j^{(i,j,r)} + b_r), \tag{6}$$

$\sigma(\cdot)$ is the sigmoid activation function. $\boldsymbol{W}_r \in \mathcal{R}^{d\times d}$ and $b_r \in \mathcal{R}^d$ are model parameters.

Loss Function

To address the issue of false negatives in the dataset's relation triplets, we employ positive-unlabeled learning under the class prior shift of training set (S-PU) to mitigate this problem, as described by Wang et al. (2022).

The following equation formalizes the loss function for S-PU learning, which is designed to weigh the contributions of each relation class differently based on their prior probabilities.

$$\pi_{u,i} = \frac{\pi_i - \pi_{\text{labeled},i}}{1 - \pi_{\text{labeled},i}}, \tag{7}$$

$$\begin{aligned} L_{\text{S-PU}}(f) = \sum_{i=1}^{R}\left(\frac{\pi_i}{n_{\text{P}_i}}\sum_{j=1}^{n_{\text{P}_i}}\ell(f_i(\boldsymbol{x}_j^{\text{P}_i}), +1)\right. \\ + \max\left(0, \left[\frac{1}{n_{\text{U}_i}}\frac{1-\pi_i}{1-\pi_{u,i}}\sum_{j=1}^{n_{\text{U}_i}}\ell(f_i(\boldsymbol{x}_j^{\text{U}_i}), -1)\right.\right. \\ \left.\left.\left. - \frac{1}{n_{\text{P}_i}}\frac{\pi_{u,i} - \pi_{u,i}\pi_i}{1-\pi_{u,i}}\sum_{j=1}^{n_{\text{P}_i}}\ell(f_i(\boldsymbol{x}_j^{\text{P}_i}), -1)\right]\right)\right), \end{aligned} \tag{8}$$

where $\pi_i = p(y_i = +1)$ represents the prior probability of a positive relation for class i, and $\pi_{u,i} = p(y_i = 1 \mid s_i = -1)$ represents the probability of an unlabeled but positive relation for class i. $\pi_{\text{labeled},i} = p(s_i = +1)$ represents the probability of a positive relation for class i calculated from training set. To simplify computation, we introduce a hyperparameter e to approximate π_i as $\pi_i = e \cdot \pi_{\text{labeled},i}$. Here, $\ell(\cdot)$ is a convex loss function, and $f_i(\cdot)$ is the score function.

3.3 Relation Completion Based on LLM

In response to the remaining unextracted relations in the Transformer-based model mentioned earlier, this paper utilizes a LLM for relation completion. Using the relation scores generated by the Transformer model for each entity pair, high-probability candidate relation samples are dynamically selected. This significantly narrows down the search space for relation extraction, allowing the LLM to extract relations using both its internal and external knowledge. Finally, the extracted relations are merged with the previous step's results to obtain the final relation extraction outcome.

Dynamic Selection of Candidate Samples

Inspired by ATLOP, the Transformer model outputs scores for R relation types along with a threshold (th) score, which indicates the presence or absence of a relation. This results in a total of $R + 1$ scores. Relations with scores above this threshold are directly extracted by the Transformer model. In the second step, the focus shifts to candidate relation samples with scores below the threshold, identifying previously unextracted relation triplets.

To dynamically construct candidate samples for each entity pair, a scaling factor f is calculated using the development set, combined with a hyperparameter k. We then use the threshold score $\boldsymbol{score}_{th}$ for each entity pair, obtained from the Transformer-based model, to compute $\boldsymbol{score}_{bottom}$, which determines the lower bound of candidate scores. Samples within this range are then included in the LLM prompt.

$$\boldsymbol{f} = \frac{\text{average}(\boldsymbol{score}_{nf}^{dev})}{\text{average}(\boldsymbol{score}_{th}^{dev})}, \tag{9}$$

$$\boldsymbol{score}_{bottom} = \boldsymbol{score}_{th} * \boldsymbol{f} * \boldsymbol{k}, \tag{10}$$

where $\boldsymbol{score}_{nf}^{dev}$, $\boldsymbol{score}_{th}^{dev}$ represent the set of scores for undiscovered positive samples and threshold th, respectively, in the development set by the Transformer-based model.

Incorporating External Knowledge

To improve the LLM's performance in relation extraction, additional detailed relation definitions and example samples for each relation are added using external knowledge bases such as Comparative Toxicogenomics Database (*CTD*) (Davis et al., 2021) and *wiki*. For the DocRED relation descriptions, a revised version of AutoRE is referenced. An example is shown below:

> “Chemical-Gene: activity - affects”: “A chemical (head entity) that modifies the activity of a gene (tail entity), potentially altering the gene’s functional outcome without specifying the direction of the change (e.g., chemical Z alters the expression of gene A, leading to unpredictable changes in cellular responses).”

Details about the prompt and relation definitions can be found in *Appendix D* and *Appendix E*.

Inference and Merging of Results

After applying Chain-of-Thought (CoT) reasoning for relation extraction inference and performing regularization, the final determination of whether a candidate sample is valid is obtained. This result is then merged with the first step outcomes. Since the score threshold selection intervals are non-overlapping, the recall scores from both parts can be directly combined, significantly improving the final recall and F1 scores.

4 Experimental Settings

4.1 Datasets

ChemDisGene: ChemDisGene (Zhang et al., 2022) is a biomedical multi-label document-level RE dataset. In its training set, entity mentions are obtained from PubTator Central (Wei et al., 2019), while relations are sourced from the *CTD*. The test set consists of 523 abstract documents manually annotated by a team of biologists. As a result, the average number of relations per document in the test set significantly exceeds the average number in the training set, suggesting the presence of incomplete annotations and underreporting in the training data.

BioRED: BioRED (Luo et al., 2022) is a biomedical multi-label document-level RE dataset, designed to predict multiple associations between gene, chemical, disease, and variants. The dataset is manually annotated by experts based on 600 abstracts. Due to the high cost of manual annotation, BioRED is smaller than ChemDisGene but contains more complete annotations.

DocRED: DocRED (Yao et al., 2019) is a large-scale, widely used benchmark for general-domain relation extraction. However, it is known to contain many missing annotations. Re-DocRED (Tan et al., 2022b) is a revised version of DocRED with more complete annotations. In our experiments, we use the incompletely annotated training set from DocRED and the revised development and test sets from Re-DocRED to evaluate the model’s effectiveness.

4.2 Baselines

We compare GLiM with the following four types of baselines: (1) **Standard baseline:** SOTA model ATLOP (Zhou et al., 2021), trained under a fully supervised setting; (2) **PU learning-based methods:** Methods that adjust the loss weights assigned to relational classes in relation extraction, including SSR-PU (Wang et al., 2022) and P^{3}M (Wang et al., 2024); (3) **Sub-symbolic self-training methods:** Approaches that leverage self-training techniques, such as CAST (Tan et al., 2023); 4) **Memory augmented methods:** Memory-based approaches, such as TTM-RE (Gao et al., 2024).

4.3 Implementation Details

For relation representation learning, we follow the SOTA TTM-RE method and apply PubmedBert-Large (Gu et al., 2021) to ChemDisGene and BioRED, and RoBERTa-Large (Liu et al., 2019) to DocRED. All models are implemented using Huggingface’s Transformers (Wolf et al., 2020), with AdamW (Loshchilov and Hutter, 2019) used as the optimizer. More information can be found in *Appendix A*.

For LLM inference, we conduct experiments using Llama3-8B (Grattafiori et al., 2024), Qwen2.5-7B (Yang et al., 2024), gpt-4o-mini, and DeepSeek-V3 (Liu et al., 2024).

4.4 Evaluation Metric

For ChemDisGene and BioRED, we use micro F1 (F1), precision, and recall as evaluation metrics. For DocRED, we use micro F1 (F1), micro ignored F1 (Ign F1), precision, and recall. Ign F1 measures the F1 score while excluding relations shared between the training and test sets. All results are obtained on the test set.

5 Results

5.1 Main Results

Table 1 presents the experimental results on ChemDisGene. The results indicate that the main factor limiting the F1 score is the lower recall. Methods such as P^{3}M-ATLOP show significant improvements in F1 scores compared to ATLOP, primarily due to their enhanced recall. Meanwhile,

Model	ChemDisGene			BioRED			DocRED			
	F1	Precision	Recall	F1	Precision	Recall	IgnF1	F1	Precision	Recall
ATLOP*	42.73	76.17	29.70	42.89	45.96	40.20	49.16	49.32	92.62	33.61
SSR-PU*	48.56	54.27	43.93	46.81	49.44	44.44	59.48	61.05	65.71	57.01
P^{3}M-ATLOP*	53.62	60.20	48.34	49.80	50.52	49.09	61.98	63.06	69.22	57.95
CAST-ATLOP*	42.66	78.41	29.30	44.80	45.26	44.34	64.25	65.32	72.83	59.22
TTM-RE*	53.59	53.83	53.34	43.65	42.88	44.44	-	-	-	-
GLiM (w/o LLM)	54.03	64.67	46.39	50.62	54.21	47.47	62.02	63.17	67.62	59.26
GLiM (Llama3-8B)	56.95	57.01	56.90	51.93	49.81	54.24	60.56	61.98	60.59	63.44
GLiM (Qwen2.5-7B)	58.12	58.57	57.67	**53.24**	50.36	56.46	62.15	63.42	64.60	62.28
GLiM (gpt-4o-mini)	58.21	59.19	57.27	52.11	48.24	**56.67**	61.96	63.30	62.86	**63.76**
GLiM (DeepSeek-V3)	**59.00**	59.56	**58.44**	53.18	50.18	56.57	62.76	64.00	65.04	63.00

Table 1: Experimental results across test datasets: ChemDisGene, BioRED, and DocRED. Models marked with * have some results from Wang et al. (2024), Gao et al. (2024), and Tan et al. (2023).

the proposed GLiM achieves the highest F1 score across all evaluations.

Without using LLM for relation completion inference, GLiM improves precision while maintaining a strong recall score, already outperforming the current SOTA models. This suggests that the adaptively-adjusted Graph structure effectively processes and extracts correct relations with finer granularity.

After incorporating LLM for relation completion inference, recall improves significantly, leading to an overall boost in F1. The F1 score surpasses that of the best-performing SOTA model, P^{3}M-ATLOP, by 5.38 points (59.00 vs. 53.62), while recall increases by 5.1 points compared to TTM-RE (58.44 vs. 53.34). These results demonstrate that GLiM effectively leverages LLM to extract additional correct relations, thereby mitigating the issue of false negatives in the dataset.

For the BioRED dataset results in Table 1, although its training set is much smaller than that of ChemDisGene, our model still achieves the highest scores on the low-data dataset. The first step GLiM model outperforms all SOTA approaches, and after incorporating LLM, recall improves significantly, exceeding the best SOTA recall (P^{3}M-ATLOP) by 7.58 points (56.67 vs. 49.09), and the F1 score by 3.44 points (53.24 vs. 49.80). These results suggest that even with smaller training datasets, GLiM still effectively extracts correct relations, demonstrating its ability to handle multi-label biomedical relation extraction tasks.

In the general domain, our model remains competitive with SOTA approaches. Specifically, GLiM model achieves a recall score higher than the best SOTA model (CAST-ATLOP) on DocRED. After LLM inference, the recall score surpasses that of CAST-ATLOP by 4.54 points (63.76 vs 59.22). This suggests that, even in a broader set of general-domain relations, GLiM can still identify more hidden correct relations while maintaining precision.

Overall, the experimental results across both biomedical and general domains demonstrate that integrating Graph Transformer with LLM inference improves recall and F1 scores, thereby enhancing overall performance. Moreover, GLiM achieves a better balance between recall improvement and precision maintenance, leading to a higher F1 score.

5.2 Ablation Study

We conduct an ablation study to evaluate the effectiveness of the key components of GLiM: LLM inference (LLM), Graph Transformer (GT), and S-PU loss (S-PU). Four versions of GLiM are tested: (1) GT + S-PU, without LLM; (2) GT, without S-PU and LLM; (3) S-PU, without GT and LLM; (4) Base, without all three components. When S-PU loss is removed, adaptive thresholding loss from ATLOP is used as a replacement. The results in Table 2 reveal three key observations.

First, LLM inference is crucial for improving the model's recall. Removing LLM inference leads to a decrease in recall and F1 scores by 8.99 and 2.62 points, respectively.

Second, the S-PU loss, which incorporates Positive Unlabeled learning with prior shift, improves the model's recall performance.

Finally, the Graph Transformer enhances the model's ability to extract correct relations at a finer granularity, further improving overall performance. These findings indicate that each component con-

Model	F1	Precision	Recall
GLiM (Qwen2.5-7B)	53.24	50.36	56.46
GT + S-PU, w/o LLM	50.62	54.21	47.47
GT, w/o (S-PU, LLM)	49.97	53.65	46.77
S-PU, w/o (GT, LLM)	45.71	46.19	45.25
Base, w/o (GT, S-PU, LLM)	45.23	50.56	40.91

Table 2: Ablative experiments on BioRED. LLM, S-PU, GT represent LLM inference, S-PU loss, and Graph Transformer component, respectively.

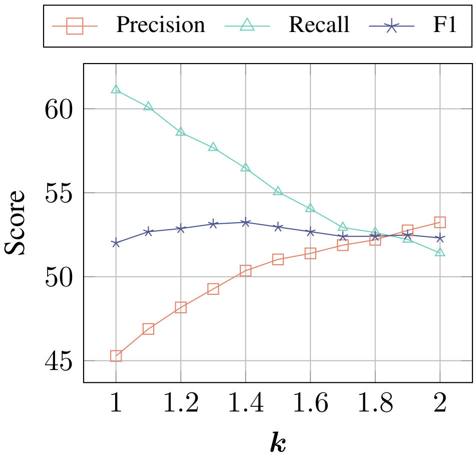


Figure 3: Effect of different $\boldsymbol{k}$ values.

tributes to improving the model's overall effectiveness.

5.3 Effect of $\boldsymbol{k}$

We further analyze the impact of the dynamic selection coefficient k on the final relation extraction results. The experiment is conducted on BioRED using Qwen2.5-7B, and the results are presented in Figure 3.

It is observed that when k is small, the selection range is broader, resulting in a higher recall after LLM inference for relation completion, though at the cost of slightly reduced precision. As k increases, the selection sample range narrows, leading to fewer newly added correct samples. This causes the recall score to drop, while precision improves accordingly. Therefore, as k increases, a balance between recall and precision emerges, allowing the F1 score to peak. Specifically, when k is set to 1.4, the F1 score reaches its highest value.

Effect of k on other datasets can be found in *Appendix B*.

5.4 LLM Studies

Figure 4 shows an F1 performance comparison of different LLMs across multiple datasets. The performance of DeepSeek-V3, gpt-4o-mini, Qwen2.5-7B and Llama3-8B, as well as a base model without LLM (base), is evaluated on the ChemDisGene, BioRED, and DocRED datasets.

DeepSeek-V3 significantly outperforms the other LLMs, achieving the highest scores across multiple datasets and demonstrating strong capability in handling complex documents and relation extraction tasks. Given its overall performance, DeepSeek-V3 is the preferred model for this task.

It is worth noting that the performance of Llama3-8B declines compared to the base model, primarily due to a decrease in precision caused by LLM hallucination, which leads to a lower F1 score. More information about hallucination can be found in *Appendix C*.

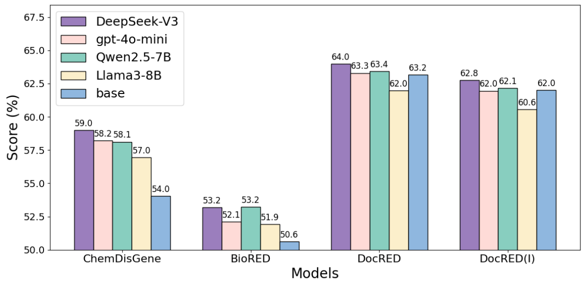


Figure 4: F1 Comparison of different LLMs. DocRED (I) reports IgnF1.

5.5 Efficiency Analysis

Our method adopts a cascaded architecture, where the Stage 1 Transformer-based pre-trained model significantly reduces the candidate relation search space before the Stage 2 LLM Inference.

As shown in Table 3, the reduction rates for the ChemDisGene, BioRED, and DocRED are 1.01%, 3.03%, and 2.82%, respectively. Notably, for ChemDisGene, even though the candidate entity-pair space volume is reduced to just 1.01% of the original, our optimal model successfully extracts approximately 12% of all correct relations in the test set. Similar recall improvements are observed for the other two datasets.

Dataset (test)	Origin	Reduced	Reduction rate	Recall increment
ChemDisGene	250,590	2,534	1.01%	+12.05
BioRED	16,034	486	3.03%	+9.20
DocRED	198,670	5,616	2.82%	+4.50

Table 3: Reduction in candidate entity pair-space volume after Stage 1 filtering and corresponding recall improvement in Stage 2, evaluated across multiple datasets.

Model	Training	LLM Inference	Total
SSR-PU	1h 13m	-	1h 13m
P^{3}M-ATLOP	1h 52m	-	1h 52m
TTM-RE	1h 19m	-	1h 19m
GLiM (Llama3-8B)	1h 14m	1m32s	1h 16m
GLiM (Qwen2.5-7B)	1h 14m	4m10s	1h 18m
GLiM (gpt-4o-mini)	1h 14m	2m2s	1h 16m
GLiM (DeepSeek-V3)	1h 14m	30m	1h 44m

Table 4: Training time and LLM Inference times for different models over 50 epochs with batch size of 2 on BioRED dataset.

These results indicate that our dynamic threshold selection mechanism effectively identifies high-density false-negative samples. In particular, for ChemDisGene, where the automatically labeled training data is of relatively low quality, the first-stage model struggles to correctly identify all valid relations. This further justifies the necessity of Stage 2 LLM inference, which leverages the LLM's internal knowledge and reasoning ability to mitigate noise from low-quality training labels.

By dynamically selecting candidate samples, LLM inference is applied to a much smaller and refined search space, enabling efficient processing and improved recall.

As showed in Table 4, we evaluate multiple LLMs on the BioRED dataset by processing 486 Stage-1-filtered instruction inputs using parallelized API calls with a batch size of 25. The results reveal distinct performance characteristics across models: gpt-4o-mini and Llama3-8B achieve rapid inference times of approximately 2 minutes, Qwen2.5-7B completes processing in 4 minutes and 10 seconds, while DeepSeek-V3 requires around 30 minutes for full inference execution.

Compared to the baseline models, the LLM inference time is minimal and does not impose significant computational overhead. Furthermore, across all three datasets, the total API cost remained under $10, confirming that our framework preserves computational efficiency and remains cost-effective, with all tested LLMs demonstrating practical inference times and budget-friendly usage.

6 Conclusion and Future Work

In this paper, we propose GLiM, a novel framework that combines Transformer-based pre-trained model with LLM inference. This approach significantly improves recall and overall F1 scores for biomedical document-level relation extraction. Notably, by incorporating a Graph Transformer into the Transformer-based model to enhance fine-grained relation handling, GLiM surpasses the performance of current SOTA models. Furthermore, its effectiveness is validated through experiments on three benchmark datasets.

For future work, we plan to further explore the relation reasoning capabilities of LLM, with a focus on improving precision while maintaining high recall to achieve a higher F1 score.

Limitations

Although our model achieves SOTA performance in multi-label biomedical document-level relation extraction, it does not yet surpass existing models in overall F1 score on general-domain datasets.

While GLiM achieves a higher recall score, the increase in misclassified relations prevents it from outperforming current models in terms of F1. This discrepancy is primarily due to the greater number of relation types and more complex interactions in general-domain datasets, which limit the effectiveness of LLM inference for relation completion.

Future work will focus on refining LLM inference methods to improve precision and mitigate misclassification errors.

Ethics Statement

Document-level relation extraction is a widely recognized and well-understood problem. Our model, GLiM, is trained and evaluated solely on publicly available benchmark datasets, which do not involve any personal privacy.

All the documents, external data, and models used in our research are sourced from open domains to ensure transparency and easy access to the information. Therefore, based on the methods we are currently using, we have no significant ethical concerns.

Acknowledgements

This work is supported by the Science and Technology Commission of Shanghai Municipality (No. 23511100602).

References

Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. 2021. Comparative toxicogenomics database (ctd): update 2021. *Nucleic acids research*, 49(D1):D1138–D1143.

Chufan Gao, Xuan Wang, and Jimeng Sun. 2024. TTM-RE: Memory-augmented document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 443–458, Bangkok, Thailand. Association for Computational Linguistics.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2018. Accurate, large minibatch sgd: Training imagenet in 1 hour. *Preprint*, arXiv:1706.02677.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Jinwoo Kim, Dat Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. 2022. Pure transformers are powerful graph learners. *Advances in Neural Information Processing Systems*, 35:14582–14595.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Yangming Li, lemao liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *International Conference on Learning Representations*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, RISHITA ANUBHAI, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *International Conference on Learning Representations*.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2023. Class-adaptive self-training for relation extraction with incompletely annotated training data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8630–8643, Toronto, Canada. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting DocRED - addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. 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 conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore. Association for Computational Linguistics.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4123–4135, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ye Wang, Huazheng Pan, Tao Zhang, Wen Wu, and Wenxin Hu. 2024. A positive-unlabeled metric learning framework for document-level relation extraction with incomplete labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19197–19205.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2024. Chatie: Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. AutoRE: Document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 211–220, Bangkok, Thailand. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

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, Florence, Italy. Association for Computational Linguistics.

Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. 2022. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1073–1082, Marseille, France. European Language Resources Association.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

A Parameter Settings

We apply a linear warm-up (Goyal et al., 2018) for the first 6% of steps, followed by a linear decay to zero. For the ChemDisGene, BioRED, and DocRED datasets, we set the learning rates to 1e-5, 2e-5, and 1e-5, the batch sizes to 4, 2, and 4, the epochs to 2, 50, and 40, and the e values to 3, 1, and 3, respectively.

Baseline models are trained using the optimal hyperparameters reported in their original papers, ensuring consistency in pre-trained encoder models, batch sizes, and epochs.

B Effect of $\boldsymbol{k}$

The effect of $\boldsymbol{k}$ on model performance is illustrated in Figures 5 and 6. All experiments in this section use Qwen2.5-7B. For DocRED (Figure 5), Precision increases as $\boldsymbol{k}$ rises, while Recall decreases, leading to a peak in F1 and IgnF1 scores around $\boldsymbol{k} = 1.52$. For ChemDisGene (Figure 6), Precision rises and Recall falls as $\boldsymbol{k}$ increases, with F1 peaking near $\boldsymbol{k} = 0.9$. These results highlight the trade-off between Precision and Recall, emphasizing the need to tune $\boldsymbol{k}$ for each dataset.

Through experiments, we find that the optimal $\boldsymbol{k}$ value primarily depends on the inherent data distribution of the dataset. To validate this, we conduct tests on three datasets (ChemDisGene, BioRED, and DocRED) by randomly sampling 5% of the data from each. Based on the results, we select the optimal $\boldsymbol{k}$ for each dataset.

Table 5 demonstrates that the optimal $\boldsymbol{k}$ values closely align with those reported in the original paper (0.9, 1.4, and 1.52, respectively). These findings indicate that the selection of $\boldsymbol{k}$ is largely independent of dataset size and is instead primarily determined by the dataset's inherent data distribution.

C LLMs Hallucination Risks

One potential drawback of leveraging LLM inference in Stage 2 is hallucination, where the model incorrectly infers relations that do not exist in the text. To investigate this issue, we evaluate precision changes across different LLMs. The results in Table 6 reveal three findings:

First, while recall improves significantly after introducing LLM inference, the associated drop in precision limits the potential gains in F1 score. This precision drop is mainly attributed to LLM hallucinations, where the model mistakenly identifies non-existent relations.

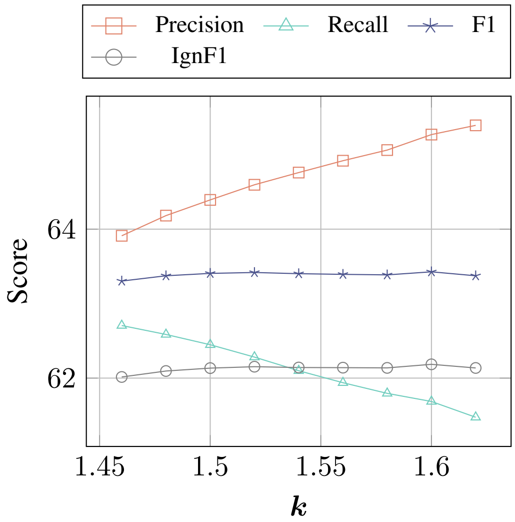


Figure 5: Effect of different $\boldsymbol{k}$ values on DocRED.

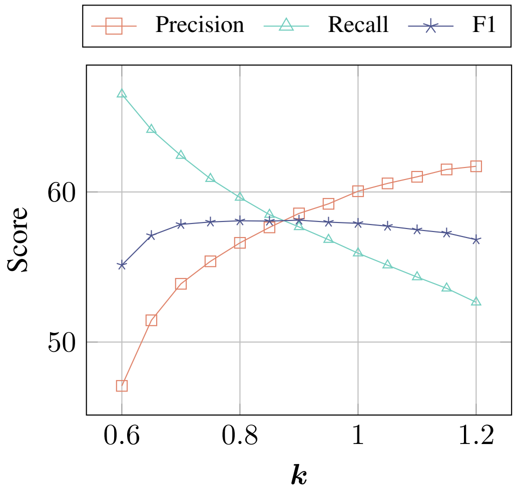


Figure 6: Effect of different $\boldsymbol{k}$ values on ChemDisGene.

Second, we observe missing annotations in the test set, which results in correctly inferred relations being mislabeled as incorrect. This highlights a potential confounding factor when evaluating hallucination.

Third, different LLMs exhibit varying degrees of hallucination. Among the models tested, DeepSeek-V3 incurs the smallest precision reduction, followed by gpt-4o-mini.

D LLM Prompt

The prompts used for LLM inference consist of four main parts:

Instruction: The LLM is guided with the instruction "You are a good reasoner", to ensure it follows the provided task and reasoning process.

ChemDisGene	***k***	0.6	0.7	0.8	**0.85**	0.9	1.0	1.1
	F1	55.53	60.77	61.05	**61.36**	60.53	59.94	57.94
BioRED	***k***	1.1	1.2	1.3	**1.4**	1.5	1.6	1.7
	F1	48.08	48.54	48.54	**49.02**	47.52	47.52	47.42
DocRED	***k***	1.40	1.46	1.50	1.52	**1.54**	1.56	1.60
	F1	59.36	59.64	59.78	59.74	**59.80**	59.69	59.52

Table 5: Effect of different k values on 5% test sets of ChemDisGene, BioRED, and DocRED.

Model	Predictions	Correct Predictions	Precision	Final Precision	Reduction
Llama3-8B	1059	397	37.49	57.01	-7.66
Qwen2.5-7B	1008	426	42.26	58.57	-6.10
gpt-4o-mini	944	411	43.54	59.19	-5.48
DeepSeek-V3	995	455	45.73	59.56	-5.11

Table 6: Precision of LLMs inference (Correct Predictions / Predictions) and corresponding final Precision reduction on ChemDisGene test dataset, compared to the base model without LLM Inference (precision: 64.67).

Supplementary Text: This contains the relevant document content for context.

Directed Candidate Relation: This part defines the candidate relation using external knowledge base augmentation.

Task Definition: This part instructs the LLM: "Given the supplementary text, judge whether the directed candidate relation directly exists from the head entity to the tail entity without other entities."

Table 7 shows the detailed prompt template used for LLM inference. Chain-of-Thought (CoT) reasoning is employed to support relation extraction inference, and regularization techniques are applied to help the model reach a final judgment. An illustrative example is shown in Figure 7.

E Relation Definitions

To enhance LLM performance in relation extraction, we enrich the dataset with detailed relation definitions and example samples for each relation. Comprehensive relation definitions are presented in Tables 8 and 9, which provide a structured overview of the relations used in the ChemDisGene dataset.

Instruction: You are a good reasoner. Give you a text. You need to understand the task definition, and answer the question.
Supplementary Text: {text}
Directed Candidate Relation: {"relation type": relation definitions}
Task Definition: Given the supplementary text, you need to judge whether the directed Candidate relation directly exists from the head entity ("Type": "mention") to the tail entity ("Type": "mention") without other entities. You must carefully judge based on the information provided in the supplementary text. Let's think step by step.
Answer:

Table 7: Prompt for relation inference

Example for LLM Inference Prompt

Instruction:
You are a good reasoner. Give you a text. You need to understand the task definition, and answer the question.

Supplementary Text:
Sulfated glucosamine inhibits MMP-2 and MMP-9 expressions in human fibrosarcoma cells. In the present study, sulfated glucosamine (SGlc) that has been reported to relieve joint pain and inflammation in many arthritis patients was studied for its inhibitory effects on MMP-2 and MMP-9 in human fibrosarcoma cells...

Directed Candidate Relation:
"chem_disease:therapeutic": A chemical(head entity) that has a known or potential therapeutic role in a disease(tail entity) (e.g., chemical X is used to treat leukemia).

Task Definition:
Given the supplementary text, you need to judge whether the directed Candidate relation directly exists from the head entity("Type-Chemical": "glucosamine") to the tail entity("Type-Disease": "arthritis") without other entities. You must carefully judge based on the information provided in the supplementary text. Let's think step by step.

Answer:
Let's break down the task step by step:

1. Identify the Head Entity and Tail Entity:
- Head Entity: "glucosamine" (Type-Chemical)
- Tail Entity: "arthritis" (Type-Disease)

2. Understand the Directed Candidate Relation:
- The relation "chem_disease:therapeutic" means that the chemical (glucosamine) has a known or potential therapeutic role in the disease (arthritis).

3. Analyze the Supplementary Text:
- The text mentions that sulfated glucosamine (SGlc) has been reported to relieve joint pain and inflammation in many arthritis patients.
- It also states that SGlc inhibits MMP-2 and MMP-9 expressions in human fibrosarcoma cells, which are involved in joint inflammation.
- The text suggests that the inhibition of MMP-2 and MMP-9 can be one of the mechanisms of SGlc to be an effective drug in relieving the symptoms of osteoarthritis.

4. Determine if the Relation Exists:
- The text directly states that SGlc (a form of glucosamine) relieves joint pain and inflammation in arthritis patients.
- It also implies that SGlc can be an effective drug in relieving the symptoms of osteoarthritis, which is a type of arthritis.

5. Conclusion:
- Based on the information provided in the supplementary text, the directed candidate relation "chem_disease:therapeutic" does exist from the head entity "glucosamine" to the tail entity "arthritis". Therefore, the directed candidate relation "chem_disease:therapeutic" directly exists from "glucosamine" to "arthritis" based on the information provided in the supplementary text.

Figure 7: Example for LLM inference prompt

Relation	**Description**
chem_gene: affects^expression	A chemical(head entity) that alters the expression of a gene(tail entity), potentially leading to changes in the amount or activity of its gene product, without specifying the direction of the change (e.g., compound X affects the expression of gene Y, leading to unpredictable variations in protein levels and cellular functions).
chem_disease: therapeutic	A chemical(head entity) that has a known or potential therapeutic role in a disease(tail entity) (e.g., chemical X is used to treat leukemia).
chem_disease: marker/mechanism	A chemical(head entity) that correlates with a disease(tail entity) (e.g., increased abundance in the brain of chemical X correlates with Alzheimer disease) or may play a role in the etiology of a disease (e.g., exposure to chemical X causes lung cancer).
chem_gene: affects^binding	A chemical(head entity) that interacts with a gene(tail entity) or its product, potentially influencing the binding affinity or stability of molecular complexes, leading to functional alterations (e.g., compound X affects the binding of transcription factor Y to gene Z, resulting in changes to gene expression and cellular activity).
chem_gene: increases^activity	A chemical(head entity) that increases the activity of a gene(tail entity), potentially enhancing its function or contributing to disease progression (e.g., chemical X enhances the expression of gene Y, leading to increased inflammatory responses in autoimmune diseases).
gene_disease: marker/mechanism	A gene(head entity) that may be a biomarker of a disease(tail entity) (e.g., increased expression of gene X correlates with breast cancer) or play a role in the etiology of a disease (e.g., mutations in gene X causes liver cancer).
chem_gene: decreases^activity	A chemical(head entity) that decreases the activity of a gene(tail entity), potentially leading to therapeutic effects in disease treatment (e.g., administration of drug Y reduces the expression of gene Z, resulting in decreased tumor growth).
chem_gene: increases^metabolic _processing	A chemical(head entity) that increases the metabolic processing of a gene(tail entity) or its product, potentially enhancing its biochemical modifications and activity (e.g., chemical Y increases the metabolic processing of protein X, leading to enhanced phosphorylation and activation of signaling pathways).
chem_gene: decreases^expression	A chemical(head entity) that decreases the expression of a gene(tail entity), potentially leading to a reduction in the production of its gene product, which could have therapeutic implications (e.g., drug Z reduces the expression of gene A, leading to lower levels of protein B, which helps mitigate inflammation in autoimmune diseases).

Table 8: 18 relation definitions in ChemDisGene.

Relation	Description
chem_gene: increases^expression	A chemical(head entity) that increases the expression of a gene(tail entity), potentially enhancing the production of its gene product and influencing biological pathways (e.g., chemical Y boosts the expression of gene X, leading to increased production of protein Z, which may promote tissue repair following injury).
chem_gene: decreases^transport	A chemical(head entity) that decreases the transport of a gene(tail entity) product into or out of a cell, potentially limiting its bioavailability and impairing its biological function (e.g., drug Z decreases the transport of protein A into the nucleus, leading to reduced transcriptional activity).
chem_gene: affects^activity	A chemical(head entity) that modifies the activity of a gene(tail entity), potentially altering the gene's functional outcome without specifying the direction of the change (e.g., chemical Z alters the expression of gene A, leading to unpredictable changes in cellular responses).
chem_gene: decreases^metabolic _processing	A chemical(head entity) that decreases the metabolic processing of a gene(tail entity) or its product, potentially leading to reduced modifications or activity of the molecule (e.g., drug Z decreases the metabolic processing of protein A, resulting in reduced post-translational modifications and lower enzymatic activity).
gene_disease: therapeutic	A gene(head entity) that is or may be a therapeutic target in the treatment a disease(tail entity) (e.g., targeted reduction of gene X expression reduces susceptibility to emphysema).
chem_gene: affects^metabolic _processing	A chemical(head entity) that alters the metabolic processing of a gene(tail entity) or its product, potentially modifying its biochemical structure without affecting expression, stability, folding, localization, splicing, or transport (e.g., compound X affects the metabolic processing of protein Y, leading to changes in its post-translational modifications and functional activity).
chem_gene: affects^localization	A chemical(head entity) that alters the localization of a gene(tail entity) or its gene product within the cell, potentially influencing its functional role in specific cellular compartments (e.g., compound X affects the localization of protein Y, leading to its redistribution from the cytoplasm to the nucleus, which may modulate gene expression and cellular signaling pathways).
chem_gene: increases^transport	A chemical(head entity) that increases the transport of a gene(tail entity) product into or out of a cell, potentially enhancing its functional activity and cellular effects (e.g., chemical Y increases the transport of protein B into the mitochondria, promoting enhanced energy production and cellular respiration).
chem_gene: affects^transport	A chemical(head entity) that alters the transport of a gene(tail entity) product into or out of a cell, potentially influencing its availability and activity within cellular compartments (e.g., compound X affects the transport of protein Y across the cell membrane, leading to changes in its intracellular concentration and function).

Table 9: 18 relation definitions in ChemDisGene (Continued).