Enhanced Reasoning for Biomedical Document-Level Relation Extraction via a Novel Cascade Language Model Framework

Haohua Song[1,2], Wenhao Gu[3], Zhijing Li[1,2], Yunwen Yu[1,2], Tiantian Zhu[1,2,*], Xiao Yang[4], Zexuan Zhu[1,2]

[1]School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen, China
[3]Guangdong Jiyin Biotech Co. Ltd., Shenzhen, China, [4]PromptBio Inc., Pleasanton, USA
{2310533010, 2310533034}@email.szu.edu.cn,{alf.gu, xyang}@promptbio.ai
{lizhijing, tiantianzhu, zhuzx}@szu.edu.cn

Abstract

Biomedical document-level relation extraction poses significant challenges beyond sentence-level tasks, as it necessitates the integration of evidence from entire documents and the ability for coherent cross-sentence reasoning. While pretrained language models (PLMs) demonstrate efficiency in handling local contexts, they often struggle with global dependency modeling. Conversely, large language models (LLMs) exhibit strong reasoning capabilities but tend to generate hallucinations in knowledge-intensive biomedical tasks. This paper introduces CoRE, a novel cascade framework that leverages the complementary strengths of PLMs and LLMs through a *detect-then-rethink* paradigm. The PLM serves as an efficient detector for high-confidence relations, while challenging cases are forwarded to an LLM enhanced with semantic retrieval and iterative reasoning mechanisms. Experimental results on the BioRED and CDR datasets show that CoRE achieves substantial improvements over state-of-the-art baselines, validating the effectiveness of the proposed cascade paradigm for complex biomedical relation extraction.

1 Introduction

Biomedical document-level relation extraction (BioDocRE) is a critical task that identifies semantic relations between entities across entire scientific documents, forming the foundation for deriving structured knowledge from large-scale unstructured literature. This task poses substantially greater challenges than sentence-level extraction, owing to its inherent complexity. Specifically, it demands the integration of long-range semantic dependencies scattered across documents, robustness to noise from domain-specific terminology variations, and the ability to process dense specialized knowledge beyond the reach of general-purpose language models. The accurate extraction of such relations is crucial for advancing biomedical research, as it directly supports knowledge discovery and accelerates progress in areas including drug development, disease mechanism interpretation, and precision medicine (Luo et al., 2022).

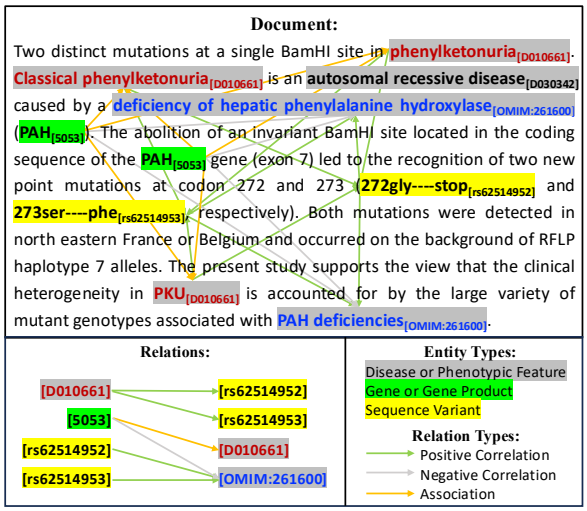

Figure 1: An illustrative example of BioDocRE, in which entities sharing the same ID but varying synonyms are in color. The colored boxes and lines denote different entity types and relation types, respectively.

Pre-trained language models (PLMs) have emerged as the leading paradigm in BioDocRE, surpassing earlier sequence models with self-attention and domain-specific pre-training. These advancements primarily follow three directions: graph-based structural dependency modeling (Li et al., 2025), attention-based entity-pair contextualization (Xu et al., 2021), and novel task reformulation (Wang et al., 2024). Nonetheless, PLMs remain constrained by heavy reliance on fine-tuning and limited reasoning capacity, particularly for implicit relations or low-resource settings. In parallel, large language models (LLMs) offer a new paradigm with strong in-context learning (ICL) (Brown et al., 2020) and reasoning abilities, reducing the need for task-specific annotations (Xu et al., 2024). How-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32865–32879

ever, in the precise biomedical domain, LLMs are prone to hallucination, often generating unsubstantiated relations that inflate recall at the expense of precision (Huang et al., 2025). Their performance still lags behind fully-supervised PLMs (Rehana et al., 2024), and their high computational cost further hinders large-scale deployment.

Consider the example in Figure 1, where determining the relation between the entities `D010661` and `rs62514952` requires cross-sentence reasoning, as the entities are linked indirectly through a bridge entity `5053`. Specifically, the second sentence states that the disease `D010661` is caused by a deficiency of the gene `5053`, while the third sentence notes that the variant `rs62514952` is located on the `5053` gene sequence. Moreover, the disease entity `D010661` appears with multiple synonyms, such as *phenylketonuria* and *PKU*, which necessitates both entity disambiguation and evidence integration.

To simultaneously overcome the reasoning limitations of PLMs on low-confidence instances and curb the hallucination bias of LLMs, we propose an innovative cascade method named CoRE. It synergizes the strengths of domain-specific PLMs and LLMs, and improves accuracy while optimizing computational efficiency in BioDocRE through confidence-based sample routing. Given the document and entity pairs, a domain-specific PLM first generates initial relation predictions and corresponding confidence scores. Instances yielding high-confidence predictions are deemed reliably resolved and accepted, whereas ambiguous, low-confidence samples are selectively routed to an LLM for refined prediction. For the LLM stage, we employ a semantic retriever to select relevant training demonstrations and construct a few-shot prompt, and further introduce a tailored iterative chain-of-thought (ItCoT) mechanism to encourage more grounded, multi-step reasoning. Finally, predictions from both routes are aggregated to form the final output, allowing the PLM to handle the majority of high-confidence cases while reserving the LLM's advanced reasoning only for intricate or long-tail cases.

The main contributions of this work are summarized as follows: (1) We propose CoRE, a novel cascade framework for BioDocRE that leverages the complementary strengths of PLMs and LLMs via a detect-then-rethink paradigm to achieve efficient state-of-the-art (SOTA) performance; (2) We design a robust LLM reasoning mechanism enhanced with semantic retrieval and iterative reasoning, significantly reducing hallucinations for more reliable extraction; (3) Experiments on BioRED and CDR show that CoRE substantially outperforms PLM-only and LLM-only baselines with practical computational costs.

2 Related Work

2.1 Document-level Relation Extraction

Document-level relation extraction (DocRE) aims to identify semantic relations between entities across an entire document, presenting unique challenges beyond sentence-level extraction (Yao et al., 2019). Recent PLM-based approaches have advanced along multiple directions. FILR (Li et al., 2022) performs multi-granularity inference at both mention-pair and entity-pair levels with bridge nodes. KG-DGAN (Li et al., 2025) explicitly incorporates external knowledge by constructing a knowledge-enhanced graph and employs a dynamic generative adversarial network to refine node representations. Regarding attention optimization, BERT-GT (Lai and Lu, 2021) employs graph Transformers (Yun et al., 2019) with neighbor attention to reduce irrelevant noise, and ATLOP (Zhou et al., 2021) utilizes adaptive thresholding and localized pooling for discriminative representations. SSAN (Xu et al., 2021) integrates co-occurrence and coreference dependencies via learnable bias terms. From a paradigm perspective, DocuNet (Zhang et al., 2021) treats the relation matrix as an image input and applies U-Net (Ronneberger et al., 2015) to capture global dependencies, while Bio-RFX (Wang et al., 2024) decomposes DocRE into predicting potential relation types followed by entity pair extraction via a Question-Answering framework. To improve the generalization of RE models, BioREx (Lai et al., 2023) proposes a data-centric framework that aggregates and standardizes multiple heterogeneous, domain-specific datasets into a unified training corpus. To effectively leverage local entity associations within a document, SRF (Zhang et al., 2024) performs secondary relation mining after completing its initial prediction pipeline, aggregating both global and local features of Noun Fragments to target adjacent entities that lack initial relational predictions.

Despite these innovations, existing PLM-based methods heavily rely on statistical co-occurrence patterns and shallow semantic matching rather than genuine multi-step reasoning. Consequently, their

performance degrades in low-resource scenarios or when processing samples with implicitly expressed relationships.

2.2 LLM-enhanced Reasoning

LLMs have demonstrated promising performance on Information Extraction (IE) tasks through ICL (Dong et al., 2024). To further augment adaptability and reliability, researchers widely adopt retrieval-augmented generation (Lewis et al., 2020) to mitigate hallucinations by retrieving external knowledge, alongside chain-of-thought prompting (Wei et al., 2022), which induces intermediate reasoning steps. Specific to relation extraction, GPT-RE (Wan et al., 2023) combines task-aware retrieval with gold-label-induced reasoning, and CoT-ER (Ma et al., 2023a) prompts LLMs to generate evidence using task-specific conceptual knowledge and then explicitly integrates this evidence into a CoT prompt.

Despite these advances, their performance on biomedical RE tasks still lags behind fully supervised PLMs (Rehana et al., 2024), especially in BioDocRE, where long-range dependencies, implicit relations, and specialized knowledge pose compounded challenges. Existing retrieval mechanisms fail to adequately aggregate evidence spanning multiple sentences, and standard CoT methods generate reasoning paths in a single forward pass without error detection or correction mechanisms. Additionally, applying LLMs directly to all samples is often infeasible due to prohibitive computational costs and low processing efficiency. While the cascade strategy (Ma et al., 2023b) has shown promise in sentence-level IE tasks, it has not yet been adapted to the complexity of document-level reasoning. To bridge these gaps, we propose CoRE, a cascade framework that synergizes domain-specific PLMs with LLMs through confidence-based routing. Our approach incorporates a document-level semantic retriever optimized for cross-sentence evidence aggregation, and an iterative reasoning mechanism utilizing a dynamic error experience library, enabling reflection and correction during inference.

3 Method

3.1 Task Definition

The BioDocRE task can be formalized as follows: given a biomedical document $\mathcal{D}$, containing a set of entities $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$, each entity e corresponds to a set of synonyms $\mathcal{M}_e = \{m_1, m_2, \ldots, m_k\}$. For a target pair of entities (e_s, e_o), where $e_s, e_o \in \mathcal{E}$, the objective is to infer the corresponding semantic relation $r \in \mathcal{R}$, where $\mathcal{R}$ denotes a predefined set of biomedical relation types. The final output forms a set of triples $\{(e_s, r, e_o)\}$ derived from the semantic information within $\mathcal{D}$.

3.2 Preliminary Study

The proposed method is motivated by an empirical observation regarding the correlation between model confidence and reliability. We define the confidence score as the maximum probability assigned by the model across all possible relation types:

$$\mathcal{C}(e_s, e_o) = \max_{r \in \mathcal{R}} P(r \mid \mathcal{D}, e_s, e_o) \quad (1)$$

As illustrated in Figure 2, the performance of the PLM DocuNet on the CDR development set demonstrates a strong correlation between confidence and reliability. Higher-confidence predictions yield significantly greater accuracy and F1-scores, while lower-confidence ones are notably less reliable. In contrast, LLMs show stable performance across the same confidence intervals, but their performance declines in high-confidence regions due to the prevalence of "no-relation" samples. The LLMs have a bias toward predicting the existence of a relation, which leads to a higher rate of false positives and poorer performance on these specific samples.

3.3 Model Architecture

The overall architecture is illustrated in Figure 3. To synergize the exceptional performance of PLMs on high-confidence predictions with the superior generalization capabilities of LLMs, the proposed CoRE adopts a two-stage prediction paradigm to address the BioDocRE task. The system leverages the PLM for efficient handling of reliable and high-confidence samples, whereas low-confidence or intricate instances are routed to the LLM for deeper analysis. The following subsections describe each stage in detail.

3.4 Initial Prediction by PLM

The first stage employs a biomedical PLM to efficiently process documents and assign confidence scores for subsequent routing. For each target entity pair (e_s, e_o) within the document $\mathcal{D}$, the PLM predicts the relation type r_{os} and the confidence

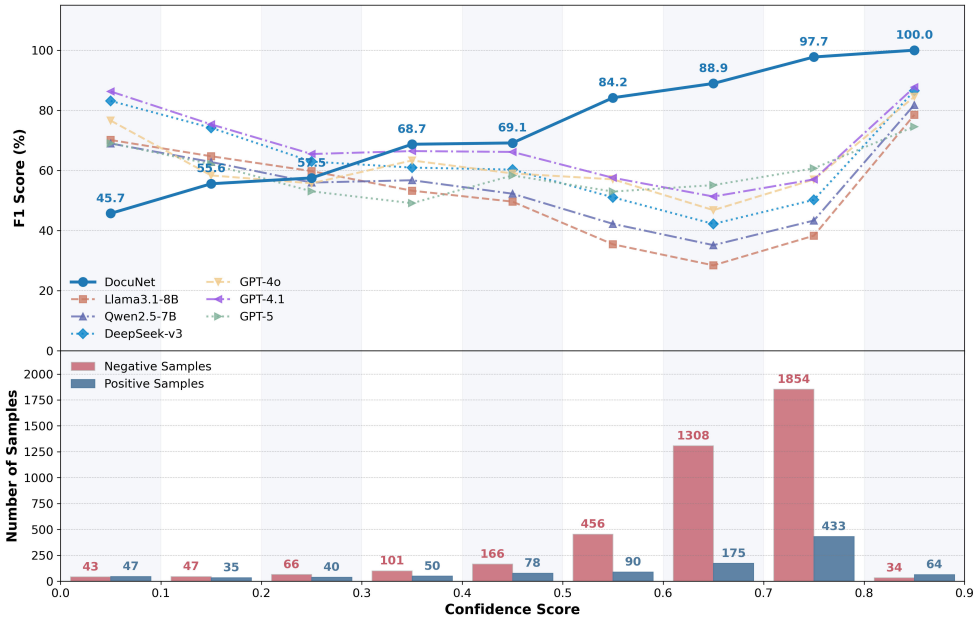

Figure 2: Performance comparison across confidence score intervals on the CDR development set. The top panel illustrates the F1 scores of DocuNet (solid line) and various LLMs (dashed lines) across different confidence score intervals, revealing the correlation between performance and confidence. The bottom panel displays the distribution of positive and negative samples within each confidence interval for DocuNet, highlighting the data class imbalance and its impact on model performance.

score c_{os}. To decide which predictions should be trusted, we introduce a confidence threshold τ, which is empirically optimized on the development set via grid search to maximize the overall system performance. For high-confidence samples with a confidence score c_{os} satisfying $c_{os} \geq \tau$, the fully supervised PLM's predictions are generally reliable, indicating that the model has successfully captured the relevant relational patterns from the training data. Thus, these predictions can be accepted directly. Conversely, instances with scores below τ are considered challenging for the PLM and are routed to the second stage for enhanced reasoning.

3.5 Mention-Aggregation-based Semantic Retrieval

To mitigate the inherent over-prediction bias of LLMs, we introduce a novel context-aware semantic retriever that augments the LLM with contextually relevant demonstrations from the training set. In BioDocRE, a significant challenge arises from the dispersed nature of entity mentions, where a single entity often appears multiple times in different sentences, each providing partial evidence. Relying on isolated mention contexts fails to capture the global relational semantics. To address this, we propose a Mention-Aggregation-based strategy that encodes all occurrences of the two target entities and aggregates them into a unified relation vector, thereby naturally integrating cross-sentence evidence for retrieval.

Given a document $\mathcal{D}$ and a target entity pair (e_s, e_o), we first construct a document-level semantic representation for each entity. We segment $\mathcal{D}$ into sentences, and let $\mathcal{M}_e = \{m_1, m_2, \ldots, m_k\}$ denote the set of mentions for entity e within these sentences. For each mention m_i, we extract its local context as s_i. Subsequently, we employ a pre-trained Sentence-Transformer (Reimers and Gurevych, 2019) model all-mpnet-base-v2[1] to encode the local context sentence of each mention. This generates a context-aware mention vector:

$$\mathbf{h}_i = \text{Encoder}(s_i), \quad \forall m_i \in \mathcal{M}_e \tag{2}$$

where $\mathbf{h}_i \in \mathbb{R}^d$ encapsulates the local semantic features of the i-th mention. While a target entity may appear multiple times within a document, decisive relational evidence is often sparse. Standard average pooling dilutes these specific signals. To address this, we adopt a LogSumExp (LSE) pooling strategy to synthesize the mention vectors into a unified entity representation $\mathbf{v}_e \in \mathbb{R}^d$. LSE serves as a smooth approximation of the max operator, enabling the model to selectively emphasize the most salient contextual features while maintaining

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

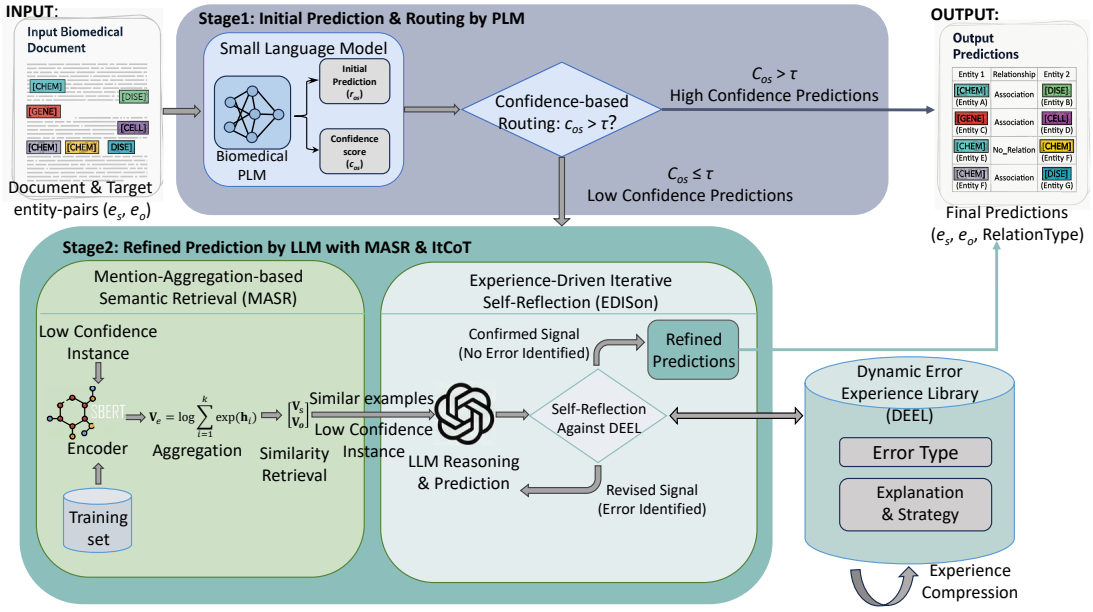

Figure 3: The overall architecture of CoRE.

numerical stability:

$$\mathbf{v}_e = \mathrm{LSE}(\{\mathbf{h}_1, \ldots, \mathbf{h}_k\}) = \log\left(\sum_{i=1}^{k} \exp(\mathbf{h}_i)\right) \quad (3)$$

For document $\mathcal{D}$ and target entities e_s and e_o, we concatenate the aggregated vectors $\mathbf{v_s}$ and $\mathbf{v_o}$, then apply L2 normalization to form the final relation representation, which removes magnitude differences and ensures a consistent basis for similarity measurement:

$$\mathbf{v}_{os} = \frac{[\mathbf{v_s} \oplus \mathbf{v_o}]}{\left\|[\mathbf{v_s} \oplus \mathbf{v_o}]\right\|_2} \quad (4)$$

During the inference phase, $\mathbf{v}_{os}$ is used to query the pre-computed index of the training set. We retrieve the Top-K instances based on cosine similarity, which serve as grounded demonstrations to guide the LLM's subsequent reasoning.

3.6 Iterative Chain-of-Thought Reasoning

Although the semantic retriever identifies contextually relevant demonstrations from the training set, standard ICL and naive CoT prompting tend to treat these demonstrations merely as formatting templates rather than logical guides. This superficial utilization not only fails to exploit the full instructional value of the demonstrations but also renders the reasoning process fragile in complex contexts, where a single logical deviation can easily lead to error propagation.

To explicitly guide the LLM in leveraging the demonstrations while robustly mitigating hallucination in BioDocRE, we propose an iterative ItCoT mechanism. This approach introduces an iterative refinement loop that empowers the LLM to engage in self-reflection and actively rectify its initial conclusions. The core idea is to construct a reasoning system capable of dynamically learning from error experiences, consisting of two collaborating components: a Dynamic Error Experience Library (DEEL) and an Experience-Driven Iterative Self-Reflection (EDISon) process.

3.6.1 Dynamic Error Experience Library

A distinctive feature of the ItCoT mechanism is its capacity for self-improvement. Concretely, before predicting on query instances, we first prompt the LLM to perform a probing inference on the retrieved exemplars and compare its predictions with the gold labels. Upon detecting a deviation from the gold label, we provide the LLM with the document context, its erroneous reasoning process, and the gold label. Then it is tasked with re-reasoning and summarizing the failure by identifying the "Error Type" and formulating a concise "Explanation & Avoidance Strategy". These distilled insights are then structured into a knowledge entry and added to the global Error Experience Library, transforming individual errors into generalizable principles that can guide future reasoning, endowing the system with a capacity for continuous learning in BioDocRE.

To prevent library redundancy, control unwieldy growth, and mitigate noise, we introduce an LLM-based experience compression mechanism. It is

triggered when the number of entries exceeds a predefined threshold, invoking the LLM to cluster, deduplicate, and summarize existing experiences. This process yields a compact set of error experiences with improved coverage and reduced noise.

3.6.2 Experience-driven Iterative Self-reflection

Building upon the constructed Error Experience Library, we further instantiate these experiences as inference constraints via an experience-driven iterative self-reflection procedure. Unlike standard CoT prompting, which is strictly feed-forward, our approach induces an iterative "Generate-Reflect-Revise" loop within a single model instance.

Specifically, after obtaining the initial reasoning and prediction, the LLM is prompted to review its reasoning trace against the entries in the error library, checking for logical errors that match known error patterns. This process compels the LLM to perform a secondary validation of its own thought process and make one of two decisions based on its self-reflection:

- **Confirm**: If the model determines that its reasoning is sound and free of any known error patterns, it outputs a "Confirm" signal. The reasoning process then terminates, and the current prediction is adopted as the final result.
- **Revise**: If the model diagnoses a flaw matching an entry in the library, it is prompted to generate a new, corrected reasoning trace and provide an updated prediction.

This "Generate-Reflect-Revise" loop can be iterated multiple times until the model outputs a "Confirm" signal or a maximum number of iterations is reached. This iterative refinement improves the reliability of the LLM's prediction by ensuring that only thoroughly validated reasoning is ultimately adopted.

4 Experiments

4.1 Datasets

We evaluate CoRE on two publicly available and widely recognized BioDocRE benchmarks: CDR (Li et al., 2016) and BioRED (Luo et al., 2022). The detailed statistics of these datasets are summarized in Table 1. CDR consists of 1,500 PubMed abstracts and serves as a benchmark for identifying relations between chemicals and diseases. BioRED is a more recent dataset comprising 600 PubMed documents. Unlike prior datasets restricted to a single entity pair or relation type, BioRED features a comprehensive schema with six entity categories and multiple relation types.

Table 1: Statistics of the CDR and BioRED datasets. We report the numbers of documents and relation instances in the training, development, and test sets, together with the numbers of entity types and relation types.

Dataset	Types		Documents			Relation Instances		
	Ent.	Rel.	Train	Dev	Test	Train	Dev	Test
CDR	2	2	500	500	500	1,038	1,012	1,066
BioRED	6	8	400	100	100	4,178	1,162	1,163

4.2 Experimental Settings

We benchmark CoRE against a diverse set of state-of-the-art methods listed in Table 2 and Table 3, and adopt DocuNet and e2eBioMedRE (Sarol et al., 2024) as base PLMs. Their predictions and corresponding confidence scores serve as the basis for the subsequent routing process. For LLMs, we employ Llama3.1-8B (Dubey et al., 2024) and Qwen2.5-7B (Yang et al., 2024), along with larger-scale models including Deepseek-v3.2 (Liu et al., 2024), GPT-4o (Achiam et al., 2023), GPT-4.1 (OpenAI, 2025a), and GPT-5 (OpenAI, 2025b). All LLMs are deployed using default parameters to ensure reproducibility. The optimal routing threshold τ is determined via grid search on the development set to maximize the hybrid F1-score and is fixed for the test set. We utilize *all-mpnet-base-v2* to encode relation instances into vector representations.

4.3 Main Results

Table 2 and Table 3 present the comparative performance of baseline PLMs, standalone zero-shot LLMs, and our CoRE method.

Performance of Zero-shot LLMs. All LLMs exhibit high recall but significantly lower precision compared to supervised PLMs. This confirms the observation in Section 3.2: LLMs have a strong bias towards predicting the existence of a relation, leading to massive false positives, especially on prevalent "no_relation" samples. Consequently, their F1-scores lag far behind domain-specific SOTA PLMs.

Effectiveness of CoRE. Compared to the original predictions, CoRE consistently boosts the performance of each LLM. Specifically, on the CDR dataset, CoRE achieves an F1 score of 88.20%, surpassing DocuNet by 1.88%. On the BioRED

dataset, our method reaches 66.35% F1 score, exceeding e2eBioMedRE by 1.91%. Crucially, these improvements are observed across the full spectrum of LLMs, from the 7B-parameter open-source model to the most advanced proprietary model GPT-5. This underscores that the coordination paradigm is a powerful and model-agnostic strategy, rather than an artifact dependent only on a specific LLM's capability. Notably, even smaller LLMs like Llama3.1-8B, when integrated into CoRE, can outperform the original results of much larger standalone models, highlighting the framework's effectiveness.

Table 2: Performance comparison on the CDR subsets. "*" denotes results produced using officially released code. "Δ" denotes the improvement over the second-best method. CoRE is implemented with DocuNet as the PLM backbone.

Method	CDR						
	Dev			Test			
	Precision	Recall	F1	Precision	Recall	F1	Δ
Base Group							
BioGPT	–	–	–	–	–	46.17	–
Bio-RFX	–	–	–	–	–	74.49	–
SSAN (SciBert)	–	–	68.40	–	–	68.70	–
KG-DGAN	–	–	–	78.00	74.00	76.00	–
FILR	–	–	–	–	–	85.70	–
ATLOP (SciBert)*	–	–	–	–	–	68.99	–
BioREx*	–	–	–	73.45	68.68	71.03	–
BERT-GT	–	–	–	64.49	71.79	65.99	–
PubMedBERT	–	–	–	57.84	53.57	55.63	–
DocuNet (SciBert)*	86.38	85.87	86.12	89.03	83.77	86.32	–
e2eBioMedRE*	68.11	75.13	71.45	69.97	71.79	70.87	–
Zero-shot LLMs							
Llama3.1-8B	26.16	91.01	40.64	27.34	89.21	41.84	–
Qwen2.5-7B	31.05	89.43	46.09	28.98	92.21	44.10	–
Deepseek-v3.2	37.66	93.67	53.72	39.81	92.02	55.58	–
GPT-4o	45.07	80.53	57.80	48.01	78.14	59.47	–
GPT-4.1	31.64	**95.95**	47.59	39.75	**93.34**	55.76	–
GPT-5	56.32	62.55	59.27	62.11	61.82	61.97	–
CoRE Group (Ours)							
Llama3.1-8B + CoRE	88.40	87.35	87.87	88.89	85.55	87.19	+0.87
Qwen2.5-7B + CoRE	88.42	87.55	87.98	**90.09**	85.27	87.61	+1.29
Deepseek-v3.2 + CoRE	88.09	89.92	**89.00**	88.75	87.34	88.04	+1.72
GPT-4o + CoRE	**90.36**	84.29	87.22	89.64	86.02	87.79	+1.47
GPT-4.1 + CoRE	88.87	87.55	88.20	89.39	86.12	87.72	+1.40
GPT-5 + CoRE	88.54	89.33	88.93	89.09	87.34	**88.20**	+1.88

4.4 Stage-wise Performance Analysis

To understand how each stage contributes to the observed improvements, we conduct a granular analysis of the confidence routing mechanism. Under the confidence-based routing mechanism, the PLM handles the majority of straightforward cases efficiently. On CDR, 4,425 out of 5,286 instances (83.71%) are resolved by the PLM with high confidence, while 861 instances (16.29%) are routed to the LLM. On BioRED, the corresponding ratio is 87.37% (high confidence) versus 12.63% (low confidence). Table 4 demonstrates the performance across these groups. By routing this specific subset to the LLM, CoRE achieves substantial improvements over the baseline PLMs, yielding F1 score increases of 15.95% on CDR and 8.07% on BioRED

Table 3: Performance comparison on the BioRED subsets. "*" denotes results produced using officially released code. "Δ" denotes the improvement over the second-best method. CoRE is implemented with e2eBioMedRE as the PLM backbone.

Method	BioRED						
	Dev			Test			
	Precision	Recall	F1	Precision	Recall	F1	Δ
Base Group							
BERT-GT	–	–	–	–	–	56.50	–
PubMedBERT	–	–	–	–	–	58.90	–
BioREx*	–	–	–	65.99	60.66	63.21	–
DocuNet (SciBert)*	59.19	43.84	50.37	64.29	48.93	55.57	–
e2eBioMedRE*	64.59	59.02	61.68	67.81	61.39	64.44	–
Zero-shot LLMs							
Llama3.1-8B	11.40	46.42	18.30	12.13	53.57	19.77	–
Qwen2.5-7B	17.52	53.75	26.43	17.25	59.76	26.77	–
Deepseek-v3.2	18.17	48.40	26.42	21.44	49.96	30.00	–
GPT-4o	19.29	52.11	28.16	20.98	52.88	30.04	–
GPT-4.1	19.22	56.51	28.68	21.03	58.21	30.90	–
GPT-5	27.14	63.68	38.06	28.19	62.42	38.84	–
CoRE Group (Ours)							
Llama3.1-8B + CoRE	65.75	60.14	62.82	68.68	61.65	64.98	+0.54
Qwen2.5-7B + CoRE	**67.94**	59.62	63.51	**70.82**	61.56	65.87	+1.43
Deepseek-v3.2 + CoRE	64.98	65.49	65.23	67.65	65.09	**66.35**	+1.91
GPT-4o + CoRE	64.89	66.01	65.44	67.98	63.71	65.78	+1.34
GPT-4.1 + CoRE	64.42	65.92	65.16	66.03	**65.35**	65.69	+1.25
GPT-5 + CoRE	67.10	**67.04**	**67.07**	67.81	64.66	66.20	+1.76

Table 4: Performance based on confidence routing. High-confidence instances are directly resolved by the PLM in the CoRE framework.

Dataset (Ratio)	Model	Precision	Recall	F1
High-Confidence Subset				
CDR (83.71%)	DocuNet	**96.29**	**94.42**	**95.34**
	GPT-5	60.91	64.55	62.67
BioRED (87.37%)	e2eBioMedRE	**77.43**	**67.53**	**72.14**
	Deepseek-v3.2	19.37	62.46	29.57
Low-Confidence Subset				
CDR (16.29%)	DocuNet	56.94	56.08	56.51
	GPT-5	**66.12**	54.73	59.89
	CoRE (Ours)	63.45	**84.46**	**72.46**
BioRED (12.63%)	e2eBioMedRE	47.97	47.14	47.55
	Deepseek-v3.2	46.17	58.57	51.64
	CoRE (Ours)	**51.19**	**60.90**	**55.62**

for these intricate cases. These results demonstrate that the overall improvement of CoRE is concentrated precisely on the instances where existing models fail.

4.5 Ablation Study

To analyze the contribution of each component in our method, we visualize the granular performance across seven entity pairs as detailed in Figure 4(c). The results indicate that naïve CoT with a standard "think step-by-step" instruction does not always improve LLM performance on the BioDocRE task. For instance, on ⟨Gene-Chemical⟩ pairs, F1 drops from 40.78% to 31.52%. Lacking task-specific guidance, an extended reasoning context can occasionally introduce noise, leading to performance degradation. LoRA fine-tuning markedly improves LLM outputs (e.g., Llama3.1-8B F1 rises from 41.84% to 68.47% on CDR), but incurs high com-

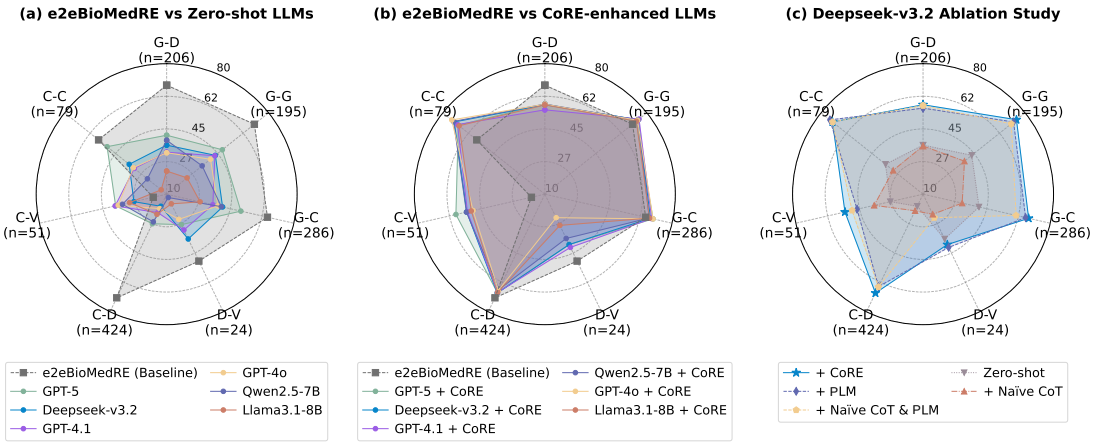

Figure 4: Radar plots comparing F1 scores across seven entity pairs on the BioRED test set. Radial axes correspond to: Gene–Disease (G-D), Gene–Gene (G-G), Gene–Chemical (G-C), Disease–Variant (D-V), Chemical–Disease (C-D), Chemical–Variant (C-V), and Chemical–Chemical (C-C), with sample sizes (n) indicated in parentheses. (a) e2eBioMedRE versus six zero-shot LLMs. (b) e2eBioMedRE versus CoRE-enhanced LLMs. (c) Ablation study of CoRE components on Deepseek-v3.2. Shaded regions indicate coverage area for each method; the radial scale ranges from 10% to 80%.

putational cost and remains inferior to domain-specific PLMs. This validates our premise that direct adaptation of LLMs is not the most efficient path for this task.

In contrast, leveraging PLMs in combination with LLMs, with or without CoT, yields substantial gains, surpassing both individual PLM and LLM baselines. Since neither model undergoes further fine-tuning, this underscores the effectiveness of the routing-based coordination method. Compared to the full CoRE method, variants lacking semantic mention aggregation and iterative reasoning show an average drop of 6.66% F1 score, confirming the necessity of these tailored components.

4.6 Analysis on Long-tail Relations

To further probe the model's capability on long-tail distributions, we additionally report the performance of models across seven major entity type pairs of BioRED in Figure 4 (detailed in Table 7), enabling a more granular evaluation under complex scenarios involving multiple entity and relation types. As shown in Figure 4(a), the state-of-the-art PLM e2eBioMedRE demonstrates a significant performance imbalance. While e2eBioMedRE performs well on common pairs like ⟨Gene, Disease⟩, its performance drops precipitously on rare ones such as ⟨Chemical, Variant⟩ (17.39% F1) due to the lack of sufficient training data. In contrast, CoRE (with Deepseek-v3.2) shows substantial improvements in these rare categories (e.g., +35.57% F1 on ⟨Chemical, Variant⟩), often doubling the performance, as shown in Figure 4(b). This demonstrates that the LLM stage, augmented with semantic retrieval and iterative reasoning, is particularly effective for cases requiring nuanced inference, while the PLM suffers from data scarcity.

4.7 Comparison with Sentence-level RAG

We focus our evaluation on a low-confidence subset, comprising the most challenging instances typically routed to the LLM stage. We benchmark CoRE against (1) LLMs without CoT (Base); (2) LLMs with naïve CoT; and (3) GPT-RE (2-shot), a representative method employing sentence-level retrieval. Figure 5 shows that GPT-RE generally improves over the Base setting, +2.06% and +1.21% F1 on average for CDR and BioRED, respectively, suggesting that RAG+CoT can be beneficial for BioDocRE. CoRE further yields consistent gains over GPT-RE across all settings, +5.39% and +5.30% average F1 on CDR and BioRED. We attribute this improvement to two factors: first, CoRE retrieves demonstrations that capture document-wide relational semantics, whereas GPT-RE's sentence-level retrieval may miss critical cross-sentence evidence; second, the iterative self-reflection process encourages the LLM to actively validate its inference and learn from transferable reasoning patterns.

To verify that CoRE is not a trivial concatenation of cascade and retrieval paradigms, we expand the evaluation to the overall system level in Table 5. While integrating GPT-RE into a cascade improves upon standalone inference, it remains inferior to CoRE, particularly on the complex BioRED dataset.

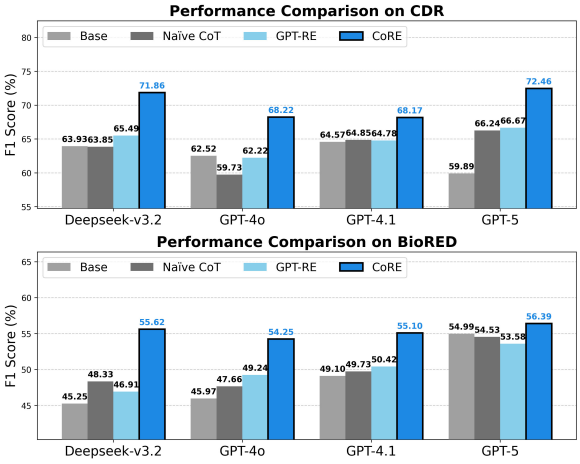

Figure 5: Performance comparison (F1 score) on the low-confidence subsets of CDR (top) and BioRED (bottom). The grouped bars display the performance of four distinct methods: Base, Naïve CoT, GPT-RE (sentence-level retrieval with guided reasoning), and the proposed CoRE method with document-level retrieval and iterative reasoning.

This confirms that the tailored architectural design in CoRE is essential for BioDocRE.

Table 5: Overall performance comparison demonstrating the necessity of task-specific designs within the cascade framework.

Dataset	Methods	Precision	Recall	F1
CDR	GPT-RE	60.66	73.99	66.67
	Cascade + GPT-RE	87.40	84.70	86.03
	CoRE (Ours)	**89.09**	**87.34**	**88.20**
BioRED	GPT-RE	39.80	52.86	45.34
	Cascade + GPT-RE	62.47	63.11	62.79
	CoRE (Ours)	**67.65**	**65.09**	**66.35**

4.8 Quantitative Analysis of Hallucination Mitigation

To quantify the mitigation of hallucinations, we analyze the False Positive Rate (FPR) and False Discovery Rate (FDR) alongside standard metrics on the challenging low-confidence subset of the CDR dataset. To ensure the findings are robust, evaluations are conducted using two distinct model families, Deepseek-v3.2 and GPT-4o. The empirical results in Table 6 reveal that applying standard chain-of-thought prompting increases the false positive rate compared to the zero-shot baseline. This indicates that unguided language models frequently generate plausible but flawed rationales to justify non-existent relations. By grounding the inference process with the structured error experience library and iterative reflection, CoRE substantially reduces false positive rates. For Deepseek, FPR drops from 41.42% to 26.19% and FDR from 48.45% to 37.28%. For GPT-4o, FPR drops from 25.84% to 22.83% and FDR from 42.07% to 36.96%. On no-relation instances, where hallucination is most directly manifested, CoRE improves F1 scores primarily through enhanced recall, confirming that the task-specific mechanisms in CoRE are essential for suppressing false positive predictions.

Table 6: Quantitative hallucination analysis on the CDR low-confidence test subset. Lower False Positive Rate (FPR) and False Discovery Rate (FDR) indicate reduced hallucination.

Methods	Split	Precision	Recall	F1	FPR ↓	FDR ↓
Deepseek-v3.2						
Base	Total	51.55	84.12	63.93	41.42%	48.45%
	No-rel.	87.57	58.58	70.20		
Naïve CoT	Total	50.49	**86.82**	63.85	44.60%	49.51%
	No-rel.	88.92	55.40	68.27		
CoRE (Ours)	Total	**62.72**	84.12	**71.86**	**26.19%**	**37.28%**
	No-rel.	**89.87**	**73.81**	**81.05**		
GPT-4o						
Base	Total	57.93	67.91	62.52	25.84%	42.07%
	No-rel.	81.52	74.16	77.66		
Naïve CoT	Total	53.32	67.91	59.73	31.15%	46.68%
	No-rel.	80.37	68.85	74.17		
CoRE (Ours)	Total	**63.04**	**74.32**	**68.22**	**22.83%**	**36.96%**
	No-rel.	**85.16**	**77.17**	**80.97**		

4.9 Cost and Efficiency Analysis

Applying LLMs to entire document collections incurs high computational costs. The confidence-based routing mechanism addresses this limitation by selectively invoking the LLM only for challenging instances. Compared to full-dataset inference on the CDR test set, the selective routing approach yields a 45.55% reduction in total API cost and a 46.89% decrease in latency. Specifically, the cost drops from $54.7 to $29.8, and the latency decreases from 1,220 minutes to 648 minutes, underscoring its substantial efficiency gains for deployment. The complete statistics are in Appendix E.

5 Conclusion

This paper introduces CoRE, a cascade framework that synergizes PLMs with LLMs via confidence-aware routing. By integrating mention-aggregation semantic retrieval and iterative reasoning, CoRE effectively addresses PLM limitations on hard and long-tail cases while mitigating LLM hallucinations. Experiments on CDR and BioRED confirm that CoRE substantially outperforms SOTA baselines, validating the efficacy of the detect-then-rethink paradigm.

Limitations

Despite the empirical success of CoRE, several limitations remain. First, the effectiveness of our retrieval-augmented prompting relies on the availability of relevant examples in the training set. Second, the confidence-based routing strategy requires the calibration of an initial threshold. While empirical analysis indicates that the optimal threshold region is relatively wide and stable across data distributions, the framework still depends on manual selection via a validation set. Future work will explore adaptive thresholding mechanisms to alleviate this dependency. Third, while CoRE reduces computational overhead compared to LLM-only approaches, the cascade architecture introduces additional latency for low-confidence samples, potentially limiting its applicability in real-time settings.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62406199, 62471310, 62502318), the Scientific Foundation for Youth Scholars of Shenzhen University (No. 868-000001032413), the Internal Fund of National Engineering Laboratory for Big Data System Computing Technology (No. SZU-BDSC-IF2024-10), Guangdong Basic and Applied Basic Research Foundation(No. 2024A1515110169), and the Intelligent Computing Center of Shenzhen University.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Abhinanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Po-Ting Lai and Zhiyong Lu. 2021. Bert-gt: cross-sentence n-ary relation extraction with bert and graph transformer. *Bioinformatics*, 36(24):5678–5685.

Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. 2023. Biorex: improving biomedical relation extraction by leveraging heterogeneous datasets. *Journal of Biomedical Informatics*, 146:104487.

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

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.

Lishuang Li, Jing Hao, Hongbin Lu, and Jingyao Tang. 2025. Document-level biomedical relation extraction via knowledge-enhanced graph and dynamic generative adversarial networks. *IEEE Transactions on Computational Biology and Bioinformatics*.

Lishuang Li, Ruiyuan Lian, Hongbin Lu, and Jingyao Tang. 2022. Document-level biomedical relation extraction based on multi-dimensional fusion information and multi-granularity logical reasoning. In *Proceedings of the 29th international conference on computational linguistics*, pages 2098–2107.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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.

Xilai Ma, Jing Li, and Min Zhang. 2023a. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples!

In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601.

OpenAI. 2025a. Introducing GPT-4.1 model family. https://openai.com/index/gpt-4-1/. Accessed: 2025-07-09.

OpenAI. 2025b. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-08-07.

Hasin Rehana, Nur Bengisu Çam, Mert Basmaci, Jie Zheng, Christianah Jemiyo, Yongqun He, Arzucan Özgür, and Junguk Hur. 2024. Evaluating gpt and bert models for protein–protein interaction identification in biomedical text. *Bioinformatics Advances*, 4(1):vbae133.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

M Janina Sarol, Gibong Hong, Evan Guerra, and Halil Kilicoglu. 2024. Integrating deep learning architectures for enhanced biomedical relation extraction: a pipeline approach. *Database*, 2024:baae079.

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.

Minjia Wang, Fangzhou Liu, Xiuxing Li, Bowen Dong, Zhenyu Li, Tengyu Pan, and Jianyong Wang. 2024. Bio-rfx: refining biomedical extraction via advanced relation classification and structural constraints. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10524–10539.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14149–14157.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Frontiers of Computer Science*, 18(6):186357.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv: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.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. *Advances in neural information processing systems*, 32.

Fu Zhang, Qi Miao, Jingwei Cheng, Hongsen Yu, Yi Yan, Xin Li, and Yongxue Wu. 2024. Srf: enhancing document-level relation extraction with a novel secondary reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15426–15439.

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*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization.

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 Detailed Performance Results

Table 7: Overall performance on CDR and BioRED datasets. Methods marked with an asterisk (*) indicate our reproduction of the original models. **Yellow background** indicates the best baseline performance (SOTA). **Green background** indicates the best performance achieved by our CoRE method. The Δ columns show the improvement over the respective best baselines (DocuNet F1 score=86.32 for CDR, e2eBioMedRE F1 score=64.44 for BioRED). Teal indicates positive improvement; Red indicates negative difference. Abbreviations: *G=Gene, D=Disease, C=Chemical, V=Variant*.

Method	Variant	CDR Test Set				BioRED Test Set				BioRED Test Set (Detailed F1)						
		Precision	Recall	F1	Δ	Precision	Recall	F1	Δ	G-D	G-G	G-C	D-V	C-D	C-V	C-C
Base Group																
BioGPT	–	–	–	46.17	–	–	–	–	–	–	–	–	–	–	–	–
Bio-RFX	–	–	–	74.49	–	–	–	–	–	–	–	–	–	–	–	–
SSAN		–	–	68.70	–	–	–	–	–	–	–	–	–	–	–	–
KG-DGAN	–	78.00	74.00	76.00	–	–	–	–	–	–	–	–	–	–	–	–
FILR	–	–	–	85.70	–	–	–	–	–	–	–	–	–	–	–	–
ATLOP	*	–	–	68.99	–	–	–	–	–	–	–	–	–	–	–	–
DocuNet	*	89.03	83.77	86.32	–	64.29	48.93	55.57	–	–	–	–	–	–	–	–
e2eBioMedRE	*	69.97	71.79	70.87	–	67.81	61.39	64.44	–	68.46	70.19	65.47	49.84	71.70	17.39	56.70
BERT-GT	–	64.49	71.79	65.99	–	–	–	56.50	–	54.80	63.50	60.20	42.50	67.00	11.80	52.90
PubMedBERT	–	57.84	53.57	55.63	–	–	–	58.90	–	56.60	66.40	59.90	50.80	65.80	25.80	54.40
Zero-shot LLMs																
Llama3.1-8B	Zero-shot	27.34	89.21	41.84	–	12.13	53.57	19.77	–	22.39	24.01	28.40	15.75	21.97	30.39	13.75
Qwen2.5-7B	Zero-shot	28.98	92.21	44.10	–	17.25	59.76	26.77	–	38.91	34.36	40.08	12.00	26.48	34.31	23.21
Deepseek-v3.2	Zero-shot	39.81	92.02	55.58	–	21.44	49.96	30.00	–	36.30	43.41	40.78	36.67	17.19	28.03	35.70
GPT-4o	Zero-shot	48.01	78.14	59.47	–	20.98	52.88	30.04	–	32.22	40.06	38.18	25.00	18.60	36.65	32.84
GPT-4.1	Zero-shot	39.75	93.34	55.76	–	21.03	58.21	30.90	–	32.53	43.07	35.43	31.33	20.96	38.34	32.19
GPT-5	Zero-shot	62.11	61.82	61.97	–	28.19	62.42	38.84	–	41.61	48.21	50.83	28.13	27.60	38.06	50.85
CoRE Group (Ours)																
Llama3.1-8B	+ CoRE	88.89	85.55	87.19	+0.87	68.68	61.65	64.98	+0.54	57.87	72.78	67.63	28.57	68.57	50.49	68.83
Qwen2.5-7B	+ CoRE	90.09	85.27	87.61	+1.29	70.82	61.56	65.87	+1.43	58.33	73.91	67.41	36.36	66.67	53.16	72.61
Deepseek-v3.2	+ CoRE	88.75	87.34	88.04	+1.72	67.65	65.09	**66.35**	+1.91	57.97	74.09	68.10	40.00	68.87	52.96	72.15
GPT-4o	+ CoRE	89.64	86.02	87.79	+1.47	67.98	63.71	65.78	+1.34	58.25	73.63	69.53	24.00	67.25	49.52	73.87
GPT-4.1	+ CoRE	89.39	86.12	87.72	+1.40	66.03	65.35	65.69	+1.25	55.24	74.62	68.77	41.67	68.49	52.60	69.89
GPT-5	+ CoRE	89.09	87.34	**88.20**	+1.88	67.81	64.66	66.20	+1.76	58.33	72.63	69.53	40.00	68.63	58.93	69.36
Ablation Study																
Llama3.1-8B	+ Naïve CoT	27.28	94.37	42.32	–	16.79	60.88	26.32	–	30.77	33.49	35.40	20.56	25.21	34.23	21.40
	+ LoRA	56.58	86.68	68.47	–	–	–	–	–	–	–	–	–	–	–	–
	+ SLM	86.64	86.96	86.80	+0.48	66.29	60.71	63.38	−1.06	56.28	70.59	64.98	26.09	68.34	49.84	66.26
	+ Naïve CoT&SLM	86.84	86.68	86.76	+0.44	66.12	62.25	64.13	−0.31	56.28	72.49	65.73	34.78	66.19	49.52	70.29
Qwen2.5-7B	+ Naïve CoT	28.99	81.89	42.83	–	16.15	55.03	24.97	–	37.99	28.66	40.16	16.39	21.77	33.49	20.29
	+ LoRA	62.32	79.27	69.78	–	–	–	–	–	–	–	–	–	–	–	–
	+ SLM	87.41	86.59	86.99	+0.67	64.29	62.85	63.57	−0.87	58.29	71.98	68.55	25.00	65.64	45.88	69.94
	+ Naïve CoT&SLM	87.81	85.18	86.48	+0.16	62.92	62.17	62.54	−1.90	56.16	68.04	68.53	25.00	64.96	51.05	66.25
Deepseek-v3.2	+ Naïve CoT	38.57	91.74	54.31	–	19.52	57.70	29.16	–	35.82	38.41	31.52	21.98	19.86	36.89	30.51
	+ SLM	87.34	87.34	87.34	+1.02	66.12	62.42	64.22	−0.22	56.16	72.11	66.43	41.67	65.04	46.30	73.89
	+ Naïve CoT&SLM	87.08	87.24	87.16	+0.84	62.93	63.63	63.27	−1.17	57.55	70.53	61.22	24.00	65.19	49.07	72.13
GPT-4o	+ Naïve CoT	44.44	76.54	56.23	–	24.11	55.55	33.63	–	38.89	38.85	43.52	32.84	22.90	38.20	39.94
	+ SLM	89.36	85.08	87.17	+0.85	63.73	61.65	62.67	−1.77	52.43	70.47	66.43	16.67	62.85	52.44	69.66
	+ Naïve CoT&SLM	88.17	85.27	86.70	+0.38	65.94	62.77	64.32	−0.12	55.17	71.54	66.20	25.00	64.22	51.71	73.98
GPT-4.1	+ Naïve CoT	38.08	90.24	53.55	–	24.49	55.80	34.04	–	37.80	45.57	39.19	29.85	23.93	38.62	37.71
	+ SLM	88.16	86.59	87.36	+1.04	64.09	63.37	63.73	−0.71	54.55	71.25	66.67	40.00	65.64	51.71	68.75
	+ Naïve CoT&SLM	88.96	86.21	87.57	+1.25	65.38	62.68	64.00	−0.44	56.31	73.85	65.96	32.00	66.20	49.52	68.66
GPT-5	+ Naïve CoT	55.71	81.43	66.16	–	28.35	59.85	38.47	–	41.08	51.22	48.18	34.38	26.44	40.38	46.76
	+ SLM	90.39	84.71	87.46	+1.14	64.63	64.57	64.60	+0.16	59.62	73.50	69.18	50.00	65.75	49.08	68.62
	+ Naïve CoT&SLM	89.09	85.83	87.43	+1.11	65.91	64.66	65.28	+0.84	57.00	75.00	71.08	32.00	65.40	50.62	71.04

B Additional Ablation Studies

To investigate the effectiveness of the task-specific designs in the proposed framework, individual modules are evaluated on the low-confidence subsets of the test data.

Retrieval Strategy Analysis. Table 8 compares the proposed semantic retrieval against keyword matching and sentence-level embeddings. Keyword matching frequently retrieves negative instances due to vocabulary overlap, whereas sentence-level embeddings struggle to capture long-range relational evidence. The mention-aggregation strategy effectively addresses both issues by contextualizing global mention information.

Table 8: F1 scores on the low-confidence subset utilizing different retrieval strategies.

Dataset	Model	BM25	PURE	MASR (Ours)
CDR	Deepseek	66.87	68.22	**71.86**
	GPT-5	66.42	69.47	**72.46**
BioRED	Deepseek	50.72	53.60	**55.62**
	GPT-5	56.24	53.65	**56.39**

Pooling Strategy in Semantic Retrieval. Table 9 shows that the LogSumExp pooling function yields superior performance compared to max and average pooling. Max pooling tends to discard diverse context information, while average pooling dilutes critical relational signals.

Table 9: Impact of different mention aggregation pooling methods on extraction performance.

Dataset	Model	Max	Average	LSE (Ours)
CDR	Deepseek	71.03	68.98	**71.86**
	GPT-5	70.44	69.33	**72.46**
BioRED	Deepseek	53.82	52.71	**55.62**
	GPT-5	54.13	53.58	**56.39**

Structured Error versus Raw Exemplars. During the iterative reflection stage, providing the language model with raw erroneous exemplars introduces redundant context noise. By extracting concise and structured avoidance strategies, the proposed mechanism achieves improved accuracy over the raw exemplar baseline, as detailed in Table 10.

C Threshold Stability and Generalization

To evaluate the stability of the confidence routing threshold τ, Table 11 details the F1 score variations on the CDR development set. The results indicate a stable performance range. Within the interval of 0.21 to 0.31, the maximum performance fluctuation remains below 0.2 F1 score. This stability suggests that the optimal threshold derived from the development set can generalize reliably to unseen test distributions without severe overfitting.

Table 10: Performance comparison of different error experience formats.

Dataset	Model	Raw Erroneous Exemplars	DEEL (Ours)
CDR	Deepseek	64.78	**71.86**
	GPT-5	67.61	**72.46**
BioRED	Deepseek	49.24	**55.62**
	GPT-5	48.70	**56.39**

Table 11: F1 scores across different routing thresholds on the CDR development set.

τ	0.21	0.22	0.23	0.24	0.25	0.26	0.27	0.28	0.29	0.30	0.31
F1	88.75	88.78	88.85	88.92	88.91	88.93	88.90	88.87	88.84	88.90	88.79

D Baseline Evaluation Details

The evaluation of BioREx (Lai et al., 2023) was conducted utilizing the official code repository and provided weights. The original evaluation protocol of BioREx employs a simplified multi-class setting for the BioRED dataset, where variant and gene entities are merged into a single concept. To ensure a rigorous comparison, we aligned the evaluation of the proposed framework with this specific entity-merging protocol. As illustrated in Table 12, CoRE maintains higher performance across the majority of specific relation pairs under identical evaluation constraints.

Table 12: Detailed performance breakdown on BioRED aligned with the BioREx evaluation setting.

Model	Overall	C-C	C-D	C-G	D-G	G-G
e2eBioMedRE	64.44	56.70	71.70	61.79	61.88	68.17
BioREx	63.21	53.48	68.68	65.26	**69.78**	52.85
CoRE (Ours)	**66.35**	**57.97**	**74.09**	**65.79**	63.17	**69.94**

E Efficiency Analysis

Table 13 reports token usage, monetary cost, and latency on the CDR test set (GPT-5, 2-shot). CoRE routes only 16.29% of instances to the LLM, reducing total cost and latency by 45.55% and 46.89%, respectively, relative to Standard CoT, while achieving higher F1 scores.

Table 13: Efficiency comparison on the CDR test set (GPT-5, 2-shot). Metrics may fluctuate slightly due to network conditions and API retry mechanisms. Abbreviations: #Total (count of total instances), #LLM (count of instances processed by LLM), InTok (input tokens), OutTok (output tokens), Avg (average per instance), and Lat (latency).

Method	#Total	#LLM	InTok	OutTok	Avg. InTok	Avg. OutTok	Cost ($)	Avg. Cost ($)	Lat (min)	Avg. Lat (min)
PLM (DocuNet)	5,286	–	–	–	–	–	–	–	1	0.0002
Standard CoT (2-shot)	5,286	5,286	23,675,935	2,507,412	4,479	474.3	54.7	0.0103	1,220	0.2308
CoRE (2-shot)	5,286	861	13,168,836	1,331,667	2,491	251.9	29.8	0.0056	648	0.1226

F LLM Prompts for Relation Extraction

The following prompts illustrate the primary inference (few-shot) and iterative reasoning.

F.1 Primary Inference Prompt

This prompt concatenates the generated demonstrations with the target task.

Listing 1: Few-Shot Prompt for Main Inference

```
[SYSTEM]
### Task: Given the following biomedical text and two entities, determine the
    relationship between them.

[USER]
Entity 1: E359 K
Entity 2: ventricular septal defect|VSD
Text: <...Demonstration Text 1 (GATA4)...>
... <Relationship definitions and Options omitted for brevity> ...
Output format: MUST begin with 'Let's think step by step:', end with 'Final answer
    :'.

[ASSISTANT]
Let's think step by step: The text discusses ... <...Model Reasoning...> ...
    Therefore, the best fit among the provided options is 'Association'.
Final answer: Association

[USER]
Entity 1: S429 T
Entity 2: ventricular septal defect|VSD
Text: <...Demonstration Text 1 (GATA4)...>
... <Relationship definitions and Options omitted for brevity> ...
Output format: MUST begin with 'Let's think step by step:', end with 'Final answer
    :'.

[ASSISTANT]
Let's think step by step:
1. The text describes a genetic study... <...Model Reasoning...>
Final answer: Association

[USER]
Entity 1: tachycardia
Entity 2: V1763M
Text: A novel SCN5A mutation manifests as a malignant form of long QT syndrome...
    <...Target Text...> ...
### Relationship definitions:
Positive_Correlation: The variant causally increases disease risk...
Negative_Correlation: The variant is protective...
Association: A general or unclear connection...
no_relation: The sentence mentions both...

### Output format
Answer MUST begin with 'Let's think step by step: ', end with 'Final choice: '.
```

F.2 Iterative Reasoning Prompt

This prompt asks the model to review its own reasoning from the previous step.

Listing 2: Prompt for iterative reasoning

```
[USER]
### Task: Review the prediction for biomedical relation extraction. Your goal is to
    critically identify if the prediction is correct or if it contains any errors.

### Context:
- Entity 1: tachycardia
- Entity 2: V1763M
- Text: <...Target Text Content...>

### Relationship definitions:
<Same Definitions as above>

### Current Prediction:
Positive_Correlation

### Reasoning:
Let's think step by step: The text describes a newborn with ventricular tachycardia
    ... <...Generated CoT from Previous Step...> ...Final choice:
    Positive_Correlation

### Instructions
    1.  Analyze the prediction's reasoning objectively.
    2.  Compare it against the list of common errors.
    < Error types, Explanations and strategies>
    3.  Provide your response in one of the following two formats ONLY:

    'Let's check the reasoning: ..... Final answer: Confirmed.'

    or 'Let's check the reasoning: ..... Revised: <new reasoning and prediction.>'

Response:
```