Defining and Evaluating Cell–Cell Relation Extraction from Biomedical Literature under Realistic Annotation Constraints

Mei Yoshikawa[1,*], Tadahaya Miuzuno[2, 3, *, †], Yohei Ohto[1]
Hiromi Fujimoto[1], Hiroyuki Kusuhara[1]

[1] Laboratory of Molecular Pharmacokinetics, Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, Japan
[2] Laboratory of Molecular Pharmacokinetics, Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, Japan, tadahaya@gmail.com
[3] The Institute of Statistical Mathematics (ISM), Research Organization of Information and Systems, Tachikawa, Tokyo, 190-8562, Japan.
[†]Author to whom correspondence should be addressed.
[*]These authors contributed equally.

Abstract

Extracting cell–cell relations from biomedical literature is essential for understanding intercellular communication in immunity, inflammation, and tissue biology. However, cell–cell relation extraction has not been established as a standalone biomedical relation extraction task, and no benchmark corpus or systematic evaluation framework currently exists. Fully manual corpus construction is costly and difficult to scale, limiting literature-based analyses of cell–cell communication. Here, we define a sentence-level cell–cell relation extraction task and construct complementary manually annotated corpora under realistic annotation constraints. To enable scalable annotation, rule-based literature mining is used solely as an annotation accelerator to identify candidate sentences, while all relation labels are assigned manually. In addition, an independently annotated PubMed corpus without rule-based filtering is constructed to evaluate robustness on natural sentence distributions. Using these resources, we evaluate representative model configurations involving entity indication strategies, classification architectures, and continued pre-training. Our results show that cell–cell relation extraction remains challenging under realistic conditions. Increasing training data size yields consistent performance gains, and specific combinations of entity-aware architectures and continued pre-training provide modest robustness improvements. Nevertheless, performance on unfiltered PubMed sentences remains in the 70% accuracy range, and error analyses indicate that failures cannot be readily explained by simple surface-level factors. Comparisons with general-purpose large language models further suggest that task complexity, rather than model class, is the primary limiting factor. Together, these findings establish a practical foundation for literature-scale cell–cell relation extraction while clarifying its intrinsic limitations.

Keywords: Cell–Cell Communication (CCC), Biomedical Relation Extraction (BioRE), Corpus Construction, Out-of-Distribution Evaluation, Continued Pre-training (CPT), Sentence-level Information Extraction

1 Introduction

Cell–cell communication (CCC) plays a central role in immunity, inflammation, tissue homeostasis, and disease progression[1]. Experimental and computational approaches, including single-cell and spatial transcriptomics, have enabled systematic analysis of intercellular interactions at the molecular level[2–4]. However, interpretation of such data still relies heavily on prior biological knowledge describing how specific cell types interact, regulate, or influence one another. Much of this knowledge is documented in the biomedical literature, making automated extraction of CCC-related information from text an important but largely unmet need[5–9].

In biomedical literature, evidence relevant to CCC is often expressed as explicit or implicit relations between cell types. In this study, we refer to such textual descriptions as cell–cell relations and formulate their extraction from text as a biomedical relation extraction (BioRE) task. While BioRE has been extensively studied for relations involving well-defined entities such as proteins, chemicals, and diseases[10,11], cell–cell relation extraction has not been established as a standalone task, and no benchmark corpus or evaluation framework currently exists[12].

Several factors have hindered the formulation of cell–cell relation extraction. Cell types are described using heterogeneous and context-dependent nomenclature, often exhibit hierarchical organization, and frequently co-occur within the same sentence without implying a direct biological relationship[13,14]. These characteristics complicate both annotation and modeling, making direct transfer of existing BioRE datasets and methods non-trivial. As a result, literature-based approaches to CCC have not been systematically evaluated as an independent relation extraction problem despite their biological relevance. Beyond these conceptual challenges, constructing annotated data for CCC presents substantial practical difficulties. Identifying sentences that potentially describe cell–cell relations requires searching a vast literature, while assigning relation labels demands careful biological interpretation. For CCC, dataset construction therefore benefits from mechanisms that narrow down candidate sentences for manual review, while still allowing explicit labeling of both relation-positive sentences and sentences in which two cell types are mentioned without any relationship described.

Against this background, we aim to enable literature-scale CCC extraction under realistic annotation constraints. We define a cell–cell relation extraction task and establish annotation guidelines to distinguish biologically meaningful relations from simple co-occurrence. Using these guidelines, we construct two complementary manually annotated corpora: a SemMed[15]-based corpus, in which candidate sentences are identified by rule-based filtering and labeled manually, and an independently annotated PubMed corpus created without such filtering to evaluate robustness on natural sentence distributions. Using these resources, we evaluate representative sentence-level BioRE configurations involving different entity representations, classification architectures, and continued pre-training strategies. Rather than proposing a new state-of-the-art model or deriving general design principles for BioRE, our goal is to clarify what can currently be achieved for extracting literature-level evidence of CCC and to characterize the practical limits of sentence-level approaches under realistic data conditions.

Figure 1 provides an overview of the overall study design, including dataset construction, model configurations, and evaluation settings. To make large-scale manual annotation feasible under realistic constraints, we intentionally adopt rule-based literature mining not as a source of labels, but solely as an annotation accelerator to narrow down candidate sentences for expert review. All relation labels are assigned manually, and an independently annotated PubMed corpus without such filtering is used to explicitly assess robustness beyond the filtered setting.

2 Related Work

2.1 Entity marking and replacement strategies in BioRE

How target entities are specified in the input text is a fundamental design choice in relation extraction. Boundary marking approaches, such as those introduced in R-BERT[16], explicitly denote entity spans and allow models to incorporate span-level representations[17–21]. In contrast, entity replacement strategies, which substitute entity names with category labels (e.g., CELL), have been widely adopted in biomedical NLP to reduce vocabulary variability caused by synonymy and terminological diversity.

Both strategies have been reported to be effective in different settings, but systematic comparisons remain limited in the biomedical domain. In particular, little is known about how these strategies behave for relations involving homogeneous entity types, such as cell–cell relations, where surface-level cues and co-occurrence patterns differ substantially from more commonly studied BioRE tasks.

2.2 CLS-only vs. entity-aware classification

BERT-based relation classification models typically fall into two categories: CLS-only architectures that rely on the global [CLS] representation, and entity-aware architectures that explicitly incorporate representations of target entity spans in addition to the [CLS] token. CLS-only models offer simplicity and have been shown to provide strong baselines in various relation extraction tasks[22,23], while entity-aware approaches have been reported to better capture local information around entity mentions[24].

However, the relative effectiveness of CLS-only and entity-aware architectures has not been systematically established across relation extraction tasks, particularly in biomedical text where sentence structures are complex and multiple entity mentions often appear. For cell–cell relation extraction, which involves relations between identical entity types, no comparative evaluation has previously been reported, leaving this design choice largely unexplored in this context.

2.3 Special tokens and continued pre-training

Entity-aware models often rely on special boundary markers that are not present in the original pre-training vocabulary of BERT-based encoders[25]. This mismatch raises the question of whether fine-tuning alone is sufficient to learn meaningful representations for such markers, or whether additional pre-training on marker-containing text is beneficial. Continued pretraining (CPT) has been reported to improve performance in some general-domain and biomedical NLP tasks by adapting models to task-specific input formats[21,26–30].

Despite these reports, the effect of CPT in combination with domain-specific encoders such as BiomedBERT, particularly for boundary-marked BioRE tasks, has not been systematically examined. Its interaction with different classification architectures in cell–cell relation extraction therefore remains an open question.

2.4 Cell–cell relation extraction from literature

BioRE has been extensively studied for interactions such as protein–protein, chemical–disease, and gene–protein relations, supported by established benchmarks and shared tasks[11,31]. In contrast, no dedicated task formulation or annotated corpus exists for extracting cell–cell relations directly from biomedical literature.

As a result, key methodological questions—such as how to represent cell entities in text and which model architectures are most suitable—have not been empirically evaluated for this class of relations. Addressing this gap requires both the construction of appropriate datasets and the evaluation of representative BioRE configurations tailored to cell–cell relations.

2.5 SemMedDB as a candidate sentence filtering resource

SemMedDB[15] consists of PubMed sentences associated with subject–predicate–object tuples automatically extracted by the rule-based SemRep system[32]. Rather than serving as a source of labeled training data, SemMedDB provides a mechanism for identifying candidate sentences in which relations between biomedical entities are likely to be described.

Because SemRep preferentially extracts sentences with relatively explicit predicate structures, sentences derived from SemMedDB may differ in distribution from unfiltered PubMed text. However, the implications of using SemMedDB for candidate sentence selection—followed by full manual annotation—have not been systematically evaluated for cell–cell relation extraction, motivating a direct comparison with independently sampled PubMed sentences.

3 Datasets

Due to the absence of existing benchmarks for cell–cell relation extraction, we constructed datasets specifically for this study (**Supplementary Table 1**). These consist of two manually annotated corpora for model training and evaluation, and an additional unlabeled corpus used exclusively for CPT. The overall dataset construction pipeline is illustrated in **Figure 2**. Explicit annotation guidelines were defined in advance, and all relation labels in the annotated corpora were assigned manually.

3.1 Annotation guideline and process

Definitions of cell–cell interactions vary across literature, necessitating criteria to distinguish substantial relationships from simple co-occurrences. In this study, a relationship was defined as a description explicitly indicating biological influence between two cells and was classified into three types:

1. Unidirectional interaction: Interaction directed from one cell to another.
2. Bidirectional interaction: Interactions are described bidirectionally.
3. Change/transition: Descriptions of cell state changes, such as conversion, differentiation, or transition.

The guidelines included definitions, decision criteria, and both typical and atypical examples. Consistency was ensured through an iterative process of drafting, annotating, reconciling rules, and revising the guidelines. The final annotation was performed based on these established guidelines (refer to **Supplementary Information** for details).

3.2 SemMed-filtered annotated corpus for training and in-distribution evaluation

The annotated corpus for model training and in-distribution evaluation was constructed using the SemMedDB[15]. In this study, SemMedDB was used solely as a candidate sentence filtering resource to identify sentences in which two cell types are mentioned in contexts suggestive of relations.

All candidate sentences were subsequently annotated manually according to the guidelines described in **Supplementary Information**. Both positive instances, where a biologically meaningful cell–cell relation is described, and negative instances, where two cell types are mentioned without an indicated relation, were labeled explicitly. Using this procedure, we obtained a manually annotated corpus of 1,560 sentences, which was split into 1,400 sentences for training and development and 160 sentences for in-distribution (ID) testing.

Supplementary Figure 1 shows the distribution of the above three types in this corpus.

3.3 PubMed-based natural text corpus for test

To evaluate model robustness beyond the SemMed-based setting, we constructed an independent test corpus from PubMed abstracts without SemMedDB or SemRep filtering. PubMed abstracts were segmented into sentences, and sentences containing at least two cell-type mentions were identified using a cell-type dictionary. From this pool, 680 sentences were randomly sampled and manually annotated in advance using fixed random seeds.

All sentences in this corpus were annotated manually using the same guidelines as in **Supplementary Information**. Because this corpus is derived without rule-based filtering, it reflects the linguistic and structural diversity encountered in practical literature mining and serves as an out-of-distribution evaluation set. Note that although all 680 sentences were fully annotated, model evaluation was performed by randomly sampling 300 sentences per run from this annotated pool, using fixed random seeds for reproducibility. This design allows robustness to be assessed across different subsets of naturally occurring PubMed sentences, rather than relying on a single fixed test set.

3.4 PubMed-based unlabeled corpus for continued pretraining

For CPT, we constructed an unlabeled corpus from PubMed abstracts. Approximately two million sentences containing cell-type mentions were extracted using the same dictionary as in Section 3.3. For each sentence, two cell mentions were selected, and entity markers were inserted to match the input format used by entity-aware models.

This corpus was used to expose the encoder to marker-containing sentence structures prior to fine-tuning, thereby adapting the model to the downstream input format. The CPT procedure followed standard masked language modeling, with marker tokens excluded from masking.

4 Methods

In this study, we examine representative modeling choices that are required to instantiate a sentence-level cell–cell relation extraction task, focusing on how different design decisions affect performance under realistic data conditions (**Figure 1**). Specifically, we consider entity indication methods, classification architectures, and the use of CPT for handling special tokens.

4.1 Problem formulation

The input consists of a sentence x extracted from scientific literature and two cell-type entities, e_1 and e_2, appearing within it. The output is a binary label $y \in \{0, 1\}$ indicating whether a cell–cell interaction is described between e_1 and e_2. In cell–cell relation extraction, both entities belong to the same semantic category, and the relationship can be symmetric. Moreover, cell names exhibit high lexical diversity and strong context dependence, and simple co-occurrence of two cell types does not necessarily imply a meaningful interaction. As a result, accurate prediction requires careful interpretation of the surrounding textual context.

These characteristics directly influence the design of relation extraction models, including choices related to entity indication, representation extraction, and pre-training strategies.

4.2 Entity marking, model architecture, and pretraining strategies

Multiple models were constructed by combining entity indication methods, classification architectures, and the presence or absence of CPT.

Replacement (Category substitution)

In this method, cell names in the sentence are replaced with category labels such as [CELL0] and [CELL1] (**Supplementary Figure 2a**).

Macrophages activate T cells. → [CELL0] activate [CELL1].

This approach reduces vocabulary variance and mitigates the out-of-vocabulary (OOV) problem. It may discard fine-grained biological semantics, but can improve robustness under lexical variability.

Boundary marking (R-BERT style)
Target entities are enclosed by markers, such as <E0>...</E0> and <E1>...</E1> (**Supplementary Figure 2b**).

Macrophages activate T cells. → <E0>Macrophages</E0> activate <E1>T cells</E1>.

This allows the model to identify explicit entity spans and extract representations from the surrounding context. A challenge, however, is that these special tokens are absent during the original pre-training of BiomedBERT; thus, the learning of their embeddings depends solely on fine-tuning.

CLS-based classifier
This architecture inputs the global [CLS] representation into the classifier (**Figure 2a**).

$$\hat{y} = softmax(Wh_{CLS} + b)$$

While the structure is simple and training is stable, it does not explicitly utilize local entity information, which may limit the model's ability to explicitly leverage local information around target entities in context-dependent tasks such as cell–cell RE.

Entity-aware classifier (R-BERT)
This architecture incorporates entity-specific representations in addition to the global [CLS] token (**Figure 2b**). The extraction method depends on the entity indication strategy:

When boundary marking is utilized, entity representations are extracted via average pooling of the tokens enclosed by the markers:

$$h_{e1}{}' = \frac{1}{|S_1|}\sum_{i \in S_1} h_i \, , h_{e2}{}' = \frac{1}{|S_2|}\sum_{i \in S_2} h_i$$

When replacement is utilized, the final hidden states corresponding to the special tokens [CELL0] and [CELL1] are directly used as $h_{e1}{}'$ and $h_{e2}{}'$.

Next, these averaged representations are passed through a separate fully connected layer with weights W_e and bias b_e, followed by a tanh activation to obtain the final entity representations:

$$h_{e1} = W_{e1} \tanh h'_{e1} + b_{e1}$$
$$h_{e2} = W_{e2} \tanh h'_{e2} + b_{e2}$$

The final prediction is obtained by concatenating the [CLS] vector and the two span vectors, followed by input into an MLP.

$$\hat{y} = softmax(W[h_{CLS}; h_{e1}; h_{e2}] + b)$$

The advantage of this approach is the ability to directly handle local information specific to the target entities.

Continued Pretraining
The motivation for CPT was to expose the encoder to input formats containing entity markers prior to fine-tuning, thereby adapting the model to the structural characteristics of the downstream task. CPT was evaluated for both CLS-only and entity-aware architectures to assess whether such adaptation affects performance in cell–cell relation extraction (**Supplementary Figure 3**).

4.3 Fine-tuning

For fine-tuning, the cell–cell RE corpus of 1,400 sentences constructed from SemMed Predications was used. Implementation was carried out using the Hugging Face Transformers library[33]. The parameters of BiomedBERT (base or CPT version) were updated to optimize for the relation classification task.

Optimization was performed using the RAdamScheduleFree[34] optimizer with a batch size of 16 for fine-tuning and 128 for CPT, and a learning rate of 2×10^{-5} for both phases. Training was run for up to 20 epochs during fine-tuning and up to 2 epochs during CPT, with early stopping applied based on development set loss. To ensure reproducibility, all experiments were conducted with multiple random seeds.

4.4 Evaluation

The test corpus of 680 cell–cell RE sentences extracted and annotated from PubMed was used as the source for evaluation. Since this dataset is derived from a natural text distribution different from SemMed, it allows for the measurement of out-of-distribution (OOD) generalization performance.

In addition to BERT-based models, we evaluated several general-purpose large language models (LLMs) on the same OOD PubMed test set to provide a contextual reference for task performance. LLMs were prompted to perform sentence-level binary classification without any task-specific fine-tuning or additional training (**Supplementary Figure 4**). This comparison is intended to contextualize the performance of task-specific models rather than to establish direct competitiveness with LLMs optimized for different objectives.

Performance was measured using accuracy, F1 score, and Matthews correlation coefficient (MCC). Among these, MCC was used as the primary evaluation metric due to its robustness to class imbalance[35].

4.5 Model configurations

The models compared in this study consist of 8 conditions derived from the combination of three axes: Entity Indication, Classification Architecture, CPT (**Supplementary Table 2**). This configuration allows us to examine the effects of each design element and their interactions within the scope of cell–cell relation extraction.

5 Results

5.1 In-distribution and out-of-distribution performance

We first compared model performance under ID and OOD evaluation settings to examine how design choices manifest under different data selection regimes. The ID test set was derived from SemMedDB[15], where candidate sentences were pre-selected by the rule-based SemRep system[32]. The OOD test set, in contrast, was sampled directly from PubMed using dictionary-based matching without predicate-based filtering. Importantly, all labels in both settings were manually annotated using identical guidelines, ensuring comparable ground truth quality.

Across all configurations, models achieved consistently high performance on the ID test set, with accuracy values clustered around 0.82 (**Table 1**). Although modest differences among

configurations were observed, the overall performance range under ID evaluation was narrow. On the OOD PubMed test set, performance decreased across all configurations, with accuracy values spanning a broader range, approximately from 0.70 to 0.76 (**Table 2**).

Comparing ID and OOD evaluations reveals a systematic performance degradation when models are applied to data drawn from a different source, with ID performance consistently exceeding OOD performance across configurations (**Table 3**). In addition, while differences among configurations are present under both settings, the range of mean performance values is wider under OOD evaluation than under ID evaluation. This observation suggests that the OOD setting provides a less constrained evaluation regime, in which variations in model behavior become more apparent than in the highly structured, rule-based filtered ID setting.

5.2 Interaction between classification architecture and continued pre-training

We next examined the interaction between classification architecture and CPT, explicitly contrasting their behavior under ID and OOD evaluation. Under ID evaluation, differences attributable to CPT or architecture were limited (**Table 1**). In both CLS-only and entity-aware architectures, models trained with and without CPT showed largely overlapping performance distributions, and the observed differences were small relative to the corresponding standard deviations. Consequently, the ID setting did not allow for a clear separation of model configurations based on CPT or architectural choice.

Under OOD evaluation, clearer differences emerged (**Table 2**). For CLS-only architectures, the introduction of CPT resulted in minimal changes in performance. In contrast, entity-aware architectures consistently showed improved performance when combined with CPT. In the Replacement setting, CPT increased OOD accuracy from 0.731±0.015 to 0.757±0.009, while in the Boundary setting, accuracy increased from 0.703±0.008 to 0.745±0.012. Similar trends were observed for MCC and F1.

These results indicate that the effect of CPT depends on the classification architecture. Rather than providing uniform improvements, CPT primarily benefits architectures that explicitly incorporate entity-level representations, and this interaction becomes apparent under OOD evaluation.

5.3 Interaction between entity indication strategy and continued pre-training

We next examined the effect of entity indication strategies, focusing on their interaction with CPT across evaluation settings. Under ID evaluation, differences between Replacement and Boundary marking were small regardless of whether CPT was applied, and performance distributions largely overlapped (**Table 1**). As a result, the effect of entity indication strategy was not clearly distinguishable under the source-matched ID setting.

In contrast, clearer differences emerged under OOD evaluation, particularly in entity-aware architectures (**Table 2**). When CPT was not applied, Replacement achieved substantially higher OOD accuracy than Boundary marking (0.731±0.015 vs. 0.703±0.008), with non-overlapping standard deviations. This result indicates that, in the absence of CPT, Boundary marking is more sensitive to source-shifted evaluation than Replacement. When CPT was applied, the performance gap between the two entity indication strategies was reduced. With CPT, Replacement achieved an OOD accuracy of 0.757±0.009 compared to 0.745±0.012 for Boundary marking, and the corresponding performance distributions partially overlapped. This suggests that CPT mitigates, but does not fully eliminate, the sensitivity of Boundary marking under OOD evaluation.

Taken together, these results indicate that the effect of entity indication strategy cannot be interpreted independently of pre-training strategy. Rather, CPT plays a compensatory role for Boundary marking, reducing the performance gap relative to Replacement when models are evaluated under source-shifted conditions.

5.4 Generalization gap analysis

To summarize the impact of source shift across evaluation settings, we report the generalization gap, defined as the difference between ID and OOD performance for each model configuration (**Table 3**). Across all configurations, performance decreased when models were evaluated on the OOD test set, resulting in a non-zero generalization gap for all metrics. This observation indicates that models trained on SemMed-filtered annotated data do not fully retain their performance when applied to PubMed sentences sampled without rule-based filtering.

Because ID and OOD evaluations were conducted with the same number of runs (N=5), the uncertainty of the generalization gap can be estimated by standard error propagation. Under this estimation, the uncertainty of the gap is comparable in magnitude to the gap itself for most configurations. Consequently, we do not perform statistical comparisons of gap magnitude across models, and instead treat the reported gaps as descriptive summaries of performance change between evaluation settings.

When interpreted alongside the OOD performance results in Sections 5.2 and 5.3, the gap estimates are consistent with the observation that CPT primarily benefits entity-aware architectures under OOD evaluation, whereas CLS-only architectures show limited sensitivity to CPT. At the same time, substantial gaps remain across all configurations, indicating that architectural and pre-training choices alone are insufficient to fully address the challenges posed by source-shifted evaluation in sentence-level cell–cell relation extraction.

To further examine whether the observed performance limitations could be attributed to simple surface-level factors, we conducted additional analyses stratified by training data size, sentence length, entity distance, and the presence of negation or speculative expressions (**Supplementary Figures 5–8, Supplementary Tables 3–4**). As expected, increasing the size of the manually annotated training data resulted in consistent performance improvements. However, no clear or systematic performance stratification was observed with respect to sentence length, entity distance, or the presence of negation and speculative cues. These results suggest that CCC extraction performance cannot be readily explained by individual surface-level properties alone.

5.5 Comparison with large language models

Finally, we compared the best-performing BERT-based configuration—an entity-aware architecture with Replacement and CPT—with several general-purpose large language models (LLMs) evaluated on the same OOD test set (**Figure 4, Supplementary Table 5**). This comparison is intended to contextualize performance rather than to establish a direct benchmark, as the models differ substantially in scale, training objectives, and inference cost.

The proposed model achieved higher accuracy (0.757±0.009) than GPT-4o (0.715±0.012) and GPT-4o-mini (0.710±0.022), and performance comparable to GPT-4. More recent models, such as GPT-4.1 (0.795±0.016) and GPT-5.1 (0.789±0.012), achieved higher accuracy, but require substantially greater computational resources and rely on API-based access, which limits reproducibility and scalability in large-scale literature mining.

These results indicate that lightweight, domain-adapted discriminative models can remain competitive for structured information extraction from biomedical literature, particularly in large-scale settings where computational cost, scalability, and independence from external APIs are practical considerations.

6 Discussion

6.1 Enabling literature-scale cell–cell relation extraction under realistic constraints

The primary contribution of this study is to enable literature-scale cell–cell relation extraction under realistic annotation constraints. To our knowledge, no established task definition or

annotated corpus has previously existed for extracting cell−cell relations directly from biomedical literature, despite their central importance in immunology, inflammation, and tissue biology.

By defining an explicit CCC task, constructing annotation guidelines, and assembling complementary annotated corpora, we demonstrate that CCC extraction is feasible in practice. Candidate sentences are identified using rule-based literature mining to reduce annotation cost, and all relations are annotated manually. Although overall performance remains limited, lightweight discriminative models and general-purpose large language models reach a similar performance range on this task, indicating that CCC extraction is non-trivial even for large models while remaining operationally tractable without reliance on API-based systems (**Figure 4, Supplementary Table 5**)[36].

In practical terms, the present study establishes that literature-scale CCC extraction can be operationalized under realistic constraints, and that remaining performance limitations are driven more by intrinsic task complexity than by model class alone.

6.2 Rule-based candidate filtering as a practical annotation accelerator

A central methodological choice in this study was the use of SemMedDB[15] as a source of candidate sentences for corpus construction. SemRep[32], the rule-based system underlying SemMedDB, extracts predicate–argument structures using predefined linguistic and semantic patterns. While such rule-based extraction is necessarily limited and does not capture the full linguistic diversity of biomedical literature, it offers a practical mechanism for identifying candidate sentences and reducing annotation cost.

In the context of CCC, rule-based candidate filtering should be understood as an annotation accelerator rather than a labeling mechanism. Leveraging SemMedDB enabled efficient identification of candidate sentences, while all relations were annotated manually. In parallel, we constructed an independently annotated PubMed corpus without rule-based filtering to serve as a complementary evaluation benchmark. Although smaller in size, this corpus is comparable to those used in many existing BioRE benchmarks and is sufficient to reveal substantial performance differences between filtered and unfiltered sentence distributions.

6.3 What our results imply specifically for CCC

Our evaluation shows that common BioRE design choices exhibit conditional effects in the context of CCC. CPT improved robustness for entity-aware architectures under OOD evaluation, while its impact on CLS-only models was limited (**Table 1**). Similarly, the effect of entity indication strategies depended on the presence of CPT, with boundary-based representations showing increased sensitivity in its absence.

Importantly, these observations are specific to CCC and to the data construction regime examined in this study. They should not be interpreted as general design principles for biomedical relation extraction. Rather than identifying optimal architectures in a broad sense, the present results narrow the space of plausible design choices for CCC and clarify which directions are unlikely to yield substantial gains through sentence-level modeling alone.

6.4 Task-specific challenges in cell–cell relation extraction

Beyond methodological considerations, CCC presents intrinsic challenges that distinguish it from more established relation extraction settings. Cell types are often described using variable nomenclature and exhibit hierarchical organization. For example, a term such as “T cell” may subsume multiple subtypes, including $CD4^+$ and $CD8^+$ T cells, depending on context[13,14]. This variability in granularity and naming complicates sentence-level extraction, as models must implicitly resolve ontological ambiguity without explicit access to hierarchical knowledge.

The persistent performance limitations observed in this study suggest that further architectural refinement at the sentence level may yield diminishing returns for CCC

extraction. Instead, explicitly incorporating cell-type hierarchies, resolving granularity mismatches, or leveraging document-level context and background biological knowledge may represent more promising directions for advancing CCC extraction from biomedical literature.

Consistent with this view, supplementary analyses indicate that CCC extraction errors are not strongly associated with simple surface-level factors such as sentence length, entity distance between cell mentions, or the presence of negation and speculative expressions (**Supplementary Figures 5–8, Supplementary Tables 3–4**). Even when the size of the manually annotated training data is increased, performance improvements remain incremental. Together, these observations reinforce the notion that the difficulty of CCC extraction arises from a combination of factors, including semantic variability and ontological ambiguity, rather than from isolated linguistic phenomena that can be addressed by straightforward modeling adjustments.

7 Conclusion

This study establishes a practical framework for extracting cell–cell relations from biomedical literature by defining the task, constructing complementary manually annotated corpora under realistic annotation constraints, and providing systematic baseline evaluations. Candidate sentences are identified using rule-based literature mining to reduce annotation cost, while all relation labels are assigned through manual annotation.

Rather than proposing a new state-of-the-art model, we clarify what can currently be achieved for CCC extraction using sentence-level BioRE approaches. Our results show that, despite incremental gains from specific design combinations, substantial performance limitations persist when models are evaluated on unfiltered PubMed sentences, indicating that CCC extraction remains intrinsically challenging.

By clarifying these limits and releasing task definitions, annotation guidelines, datasets, and baseline results, this work provides a concrete foundation for future studies on CCC extraction and related literature mining tasks involving complex and weakly defined biomedical entities.

Author Contribution

Mei Yoshikawa: Methodology, Software, Investigation, Writing – Original Draft, Visualization.
Tadahaya Mizuno: Conceptualization, Resources, Supervision, Project administration, Writing – Original Draft, Writing – Review & Editing, Funding acquisition.
Yohei Ohto: Methodology, Software.
Hiromi Fujimoto: Investigation
Hiroyuki Kusuhara:Writing - Review

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Availability

Code and models are available at https://github.com/mizuno-group/cc-bert.

Acknowledgement

We thank the contributors to the SemMedDB database utilized in this study. This work was supported by AMED under Grant Number JP22mk0101250h and 23ak0101199h0001. Additionally, we utilized ChatGPT and Gemini to improve the clarity and readability of the English language in this manuscript. The authors reviewed and revised the output and take full responsibility for the content of the publication.

References

1. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell-cell interactions and communication from gene expression. *Nat. Rev. Genet.* **22**, 71–88 (2021).
2. Troulé, K. *et al.* CellPhoneDB v5: inferring cell-cell communication from single-cell multiomics data. *Nat. Protoc.* (2025) doi:10.1038/s41596-024-01137-1.
3. Jin, S., Plikus, M. V. & Nie, Q. CellChat for systematic analysis of cell-cell communication from single-cell transcriptomics. *Nat. Protoc.* **20**, 180–219 (2025).
4. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods* **17**, 159–162 (2020).
5. Landhuis, E. Scientific literature: Information overload. *Nature* **535**, 457–458 (2016).
6. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
7. Ji, Z. *et al.* Survey of hallucination in natural Language Generation. *ACM Comput. Surv.* (2022) doi:10.1145/3571730.
8. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are few-shot clinical information extractors. in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* 1998–2022 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2022).
9. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
10. Pyysalo, S. *et al.* BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* **8**, 50 (2007).
11. Li, J. *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)* **2016**, baw068 (2016).
12. Diaz-Garcia, J. A. & Lopez, J. A. D. A survey on cutting-edge relation extraction techniques based on language models. *Artif. Intell. Rev.* **58**, 287 (2025).

13. Tan, S. Z. K. *et al.* The Cell Ontology in the age of single-cell omics. *ArXiv* arXiv:2506.10037v2 (2025).
14. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, (2016).
15. Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G. & Rindflesch, T. C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* **28**, 3158–3160 (2012).
16. Wu, S. & He, Y. Enriching pre-trained language model with entity information for relation classification. *arXiv [cs.CL]* (2019) doi:10.48550/arXiv.1905.08284.
17. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2022).
18. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
19. Zhong, Z. & Chen, D. A frustratingly easy approach for entity and relation extraction. in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 50–61 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2021).
20. Sahu, S. K. & Anand, A. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *J. Biomed. Inform.* **86**, 15–24 (2018).
21. Baldini Soares, L., FitzGerald, N., Ling, J. & Kwiatkowski, T. Matching the blanks: Distributional similarity for relation learning. in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 2895–2905 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019).
22. Shi, P. & Lin, J. Simple BERT models for relation extraction and semantic role labeling. *arXiv [cs.CL]* (2019) doi:10.48550/arXiv.1904.05255.
23. Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. What does BERT look at? An analysis of BERT's attention. in *Proceedings of the 2019 ACL Workshop BlackboxNLP:*

Analyzing and Interpreting Neural Networks for NLP 276–286 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019).

24. Hu, Y., Chen, Y., Qin, Y. & Huang, R. Learning entity-oriented representation for biomedical relation extraction. *J. Biomed. Inform.* **147**, 104527 (2023).
25. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. in *Proceedings of the 2019 Conference of the North* (eds. Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2019).
26. Gerritse, E. J., Hasibi, F. & de Vries, A. P. Entity-aware transformers for entity search. *arXiv [cs.IR]* (2022) doi:10.48550/arXiv.2205.00820.
27. Yamada, I., Asai, A., Shindo, H., Takeda, H. & Matsumoto, Y. LUKE: Deep contextualized entity representations with entity-aware self-attention. *arXiv [cs.CL]* (2020) doi:10.48550/arXiv.2010.01057.
28. Yang, D., Wang, X. & Celebi, R. Expanding the vocabulary of BERT for Knowledge Base Construction. *arXiv [cs.CL]* (2023) doi:10.48550/arXiv.2310.08291.
29. Tai, W., Kung, H. T., Dong, X., Comiter, M. & Kuo, C.-F. ExBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. in *Findings of the Association for Computational Linguistics: EMNLP 2020* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020). doi:10.18653/v1/2020.findings-emnlp.129.
30. Gururangan, S. *et al.* Don't stop pretraining: Adapt language models to domains and tasks. in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (eds. Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J.) 8342–8360 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020).
31. Taboureau, O. *et al.* ChemProt: a disease chemical biology database. *Nucleic Acids Res.* **39**, D367-72 (2011).

32. Rindflesch, T. C. & Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.* **36**, 462–477 (2003).

33. Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2020). doi:10.18653/v1/2020.emnlp-demos.6.

34. Defazio, A. *et al.* The road less scheduled. *arXiv [cs.LG]* (2024) doi:10.48550/arXiv.2405.15682.

35. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).

36. OpenAI *et al.* GPT-4 Technical Report. *arXiv [cs.CL]* (2023).

Figures and Tables

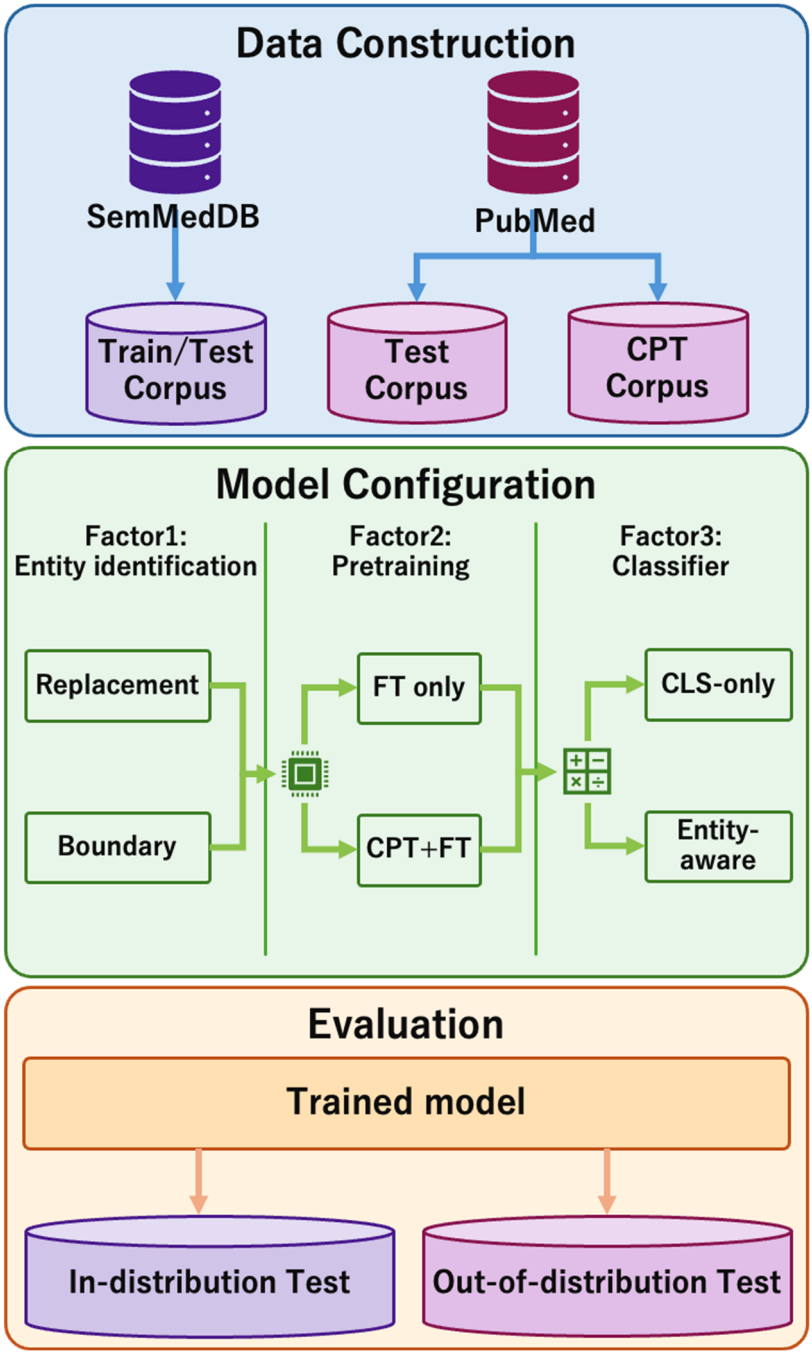


Figure 1. Schematic overview of the study design.
The experimental workflow consists of three phases: (A) Dataset Construction, where purpose-specific datasets are constructed from SemMedDB and PubMed; (B) Model Configuration, systematically comparing combinations of Entity Indication (Replacement vs. Boundary), Pre-training Strategy (with/without CPT), and Classification Architecture; and (C) Evaluation, assessing performance on both in-distribution (SemMed-derived) and out-of-distribution (PubMed-derived) datasets.

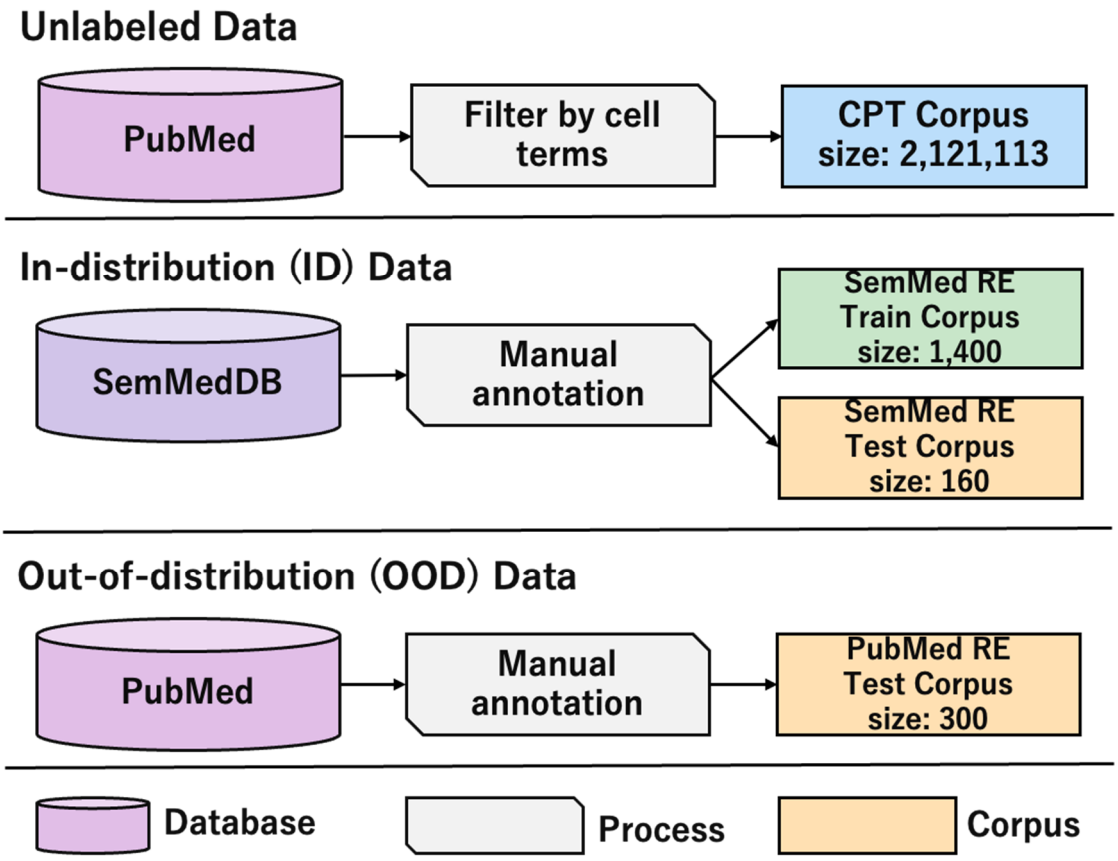


Figure 2. Dataset construction pipeline.
The figure illustrates the process of generating final corpora from their respective data sources. Top: Construction of the unlabeled corpus for Continued Pre-training (CPT) via dictionary-based filtering from PubMed. Middle: Construction of the supervised training set and In-Distribution (ID) test set utilizing Predications (triples) from SemMedDB. Bottom: Construction of the Out-of-Distribution (OOD) test set via dictionary-based matching and random sampling from PubMed. "Size" denotes the number of sentences included in each final dataset.

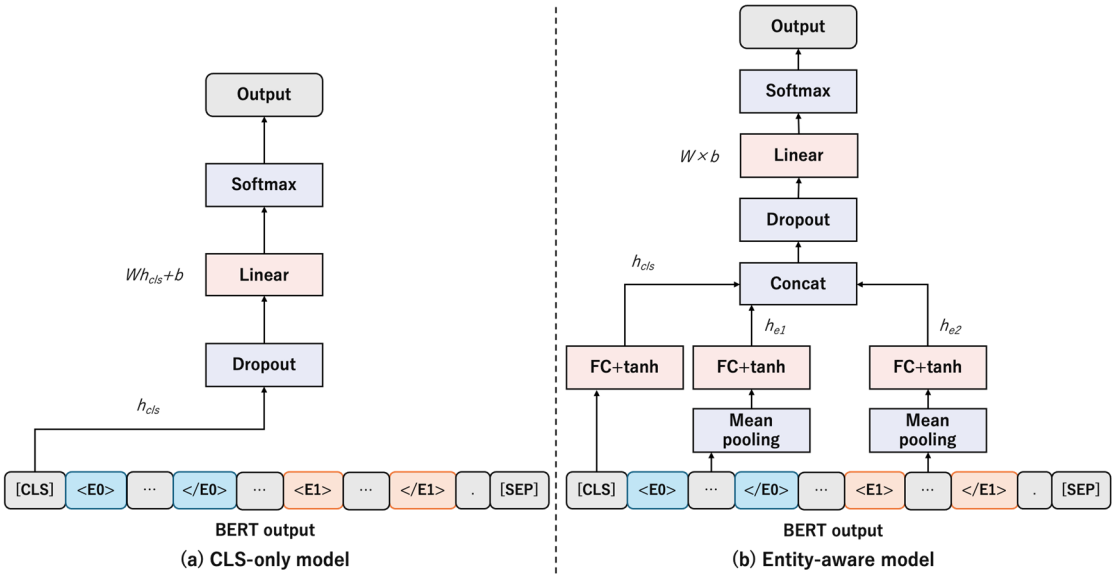


Figure 3. Schematic of classification architectures.
(a) CLS-only model: Performs classification utilizing only the global representation corresponding to the [CLS] token. (b) Entity-aware model: In addition to the [CLS] token, entity-specific representations are extracted via mean pooling of token embeddings within the entity spans (or special tokens in the Replacement method). These vectors are concatenated and passed through a fully connected (FC) layer and a softmax function for binary classification.

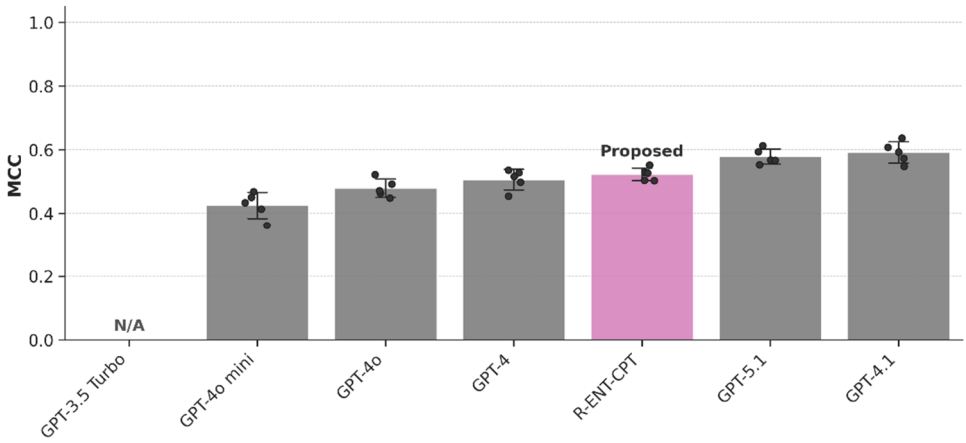


Figure 4. Performance comparison with Large Language Models (LLMs).
The figure compares the proposed task-specific model (R-ENT-CPT) with several general-purpose LLMs on the out-of-distribution (OOD) PubMed test set using Matthews Correlation Coefficient (MCC). LLMs are evaluated in a prompt-based manner without task-specific fine-tuning and are included solely as reference baselines for contextual comparison. This analysis is not intended to claim superiority of the proposed model over LLMs, but rather to illustrate that sentence-level cell–cell relation extraction remains challenging even for large, general-purpose models under the same evaluation setting. Error bars indicate the standard deviation across multiple runs.

Table 1. Performance under in-distribution (ID) evaluation on the SemMed dataset. The table reports Accuracy, Matthews Correlation Coefficient (MCC), and F1-score for all model configurations evaluated on the in-distribution (ID) SemMed test set. Results are presented as mean ± standard deviation across multiple runs.

Model	Accuracy	MCC	F1
R-ENT-CPT	**0.843±0.041**	**0.690±0.081**	**0.852±0.037**
R-ENT-base	0.840±0.053	0.682±0.106	0.845±0.052
R-CLS-CPT	0.816±0.030	0.637±0.059	0.825±0.028
R-CLS-base	0.818±0.030	0.638±0.059	0.826±0.028
B-ENT-CPT	0.825±0.057	0.655±0.115	0.836±0.053
B-ENT-base	0.820±0.034	0.642±0.068	0.825±0.036
B-CLS-CPT	0.829±0.038	0.662±0.076	0.839±0.035
B-CLS-base	0.840±0.023	0.684±0.046	0.848±0.021

Table 2. Performance under out-of-distribution (OOD) evaluation on the PubMed dataset.
The table reports Accuracy, Matthews Correlation Coefficient (MCC), and F1-score for all model configurations evaluated on the out-of-distribution (OOD) PubMed test set. Results are presented as mean ± standard deviation across multiple runs.

Model	Accuracy	MCC	F1
R-ENT-CPT	**0.757±0.009**	**0.523±0.020**	**0.734±0.009**
R-ENT-base	0.731±0.015	0.474±0.032	0.699±0.017
R-CLS-CPT	0.715±0.021	0.435±0.044	0.695±0.018
R-CLS-base	0.713±0.025	0.431±0.053	0.694±0.019
B-ENT-CPT	0.745±0.012	0.498±0.024	0.719±0.015
B-ENT-base	0.703±0.008	0.415±0.017	0.668±0.016
B-CLS-CPT	0.738±0.023	0.480±0.051	0.722±0.017
B-CLS-base	0.743±0.011	0.492±0.023	0.724±0.013

Table 3. Generalization gap between in-distribution (ID) and out-of-distribution (OOD) evaluations.
The table summarizes the performance gap between ID and OOD evaluations for each model configuration. Δ represents the generalization gap, calculated as $\text{Score}_{\text{SemMed}} - \text{Score}_{\text{PubMed}}$. Variance estimates are not reported, as the gap is used as a descriptive indicator of robustness rather than as a statistical estimator.

Model	ΔAccuracy	ΔMCC	ΔF1
R-ENT-CPT	0.085	0.167	0.118
R-ENT-base	0.109	0.208	0.145
R-CLS-CPT	0.101	0.202	0.130
R-CLS-base	0.104	0.207	0.131
B-ENT-CPT	0.080	0.157	0.117
B-ENT-base	0.117	0.227	0.157
B-CLS-CPT	0.091	0.182	0.116
B-CLS-base	0.097	0.192	0.124

Appendix

Supplementary Information

To determine the relationships between two cell names (entities) appearing in the text, we established the following annotation criteria.

1. General Principles

1.1 Entity Eligibility and Basic Rules

- **Suitability of Entities:**
  - Instances where the cell name serves as a modifier for a non-cellular noun (e.g., "**[Mast cell]** colony-forming unit") are considered invalid.
  - Phrases like "CD8 positive **[T cell]**" are acceptable.
  - Phrases connecting distinct cells, such as "B and **[T cells]**," are counted as two independent entities.
  - Phrases grouping subtypes under a single head noun, such as "CD8+ and CD4+ **[T cells]**," are treated as a single cell entity.
- **Definition of Directness and Mediation:**
  - **Substance Mediation:** Cases where Cell A acts on Cell B via a substance secreted by Cell A or via receptors are considered **Direct Relationships**.
  - **Indirect Mediation:** Any mediation **other than the direct mechanisms defined above** (e.g., involving a third cell or intervening events) is considered indirect and annotated as **No Relation**.
- **Marker Expression:**
  - Statements merely indicating marker possession (e.g., "positive for cell A marker") are not considered relationships.
  - However, if interaction is implied via the marker (e.g., "binding via marker"), it is considered a relationship.

1.2 Contextual Judgment

- **Presence and Strength of Relationship:**
  - Negative expressions suggesting a lack of relationship (e.g., "poorly," "few," "less") are annotated as **No Relation**.
  - Positive or suggestive expressions (e.g., "a few," "suggested") are annotated as **Relation Exists**.
- **Changes and Effects:**
  - Contexts describing changes in count, activity, or binding (e.g., "decreases," "increases") due to the action of one cell are considered relationships.
- **Semantic Interpretation:**
  - Even if the grammatical subject or object is not the cell itself, a relationship is determined if the meaning involves the cell's function or behavior.
  - **Example:** "Increased binding of T cells to **[B cells]** increases the apoptosis rate of **[macrophages]**." (Treated as a relationship between B cells and Macrophages).

2. Classification Categories

The relationships between extracted cell pairs are classified into the following categories:

2.1 Bidirectional Effect

Refers to undirected or bidirectional physical/functional contact.

- **Definition:** Binding, recognition, or cross-regulation.
- **Examples:**
  - Binding (where direction is unspecified).
  - Recognition leading to a change.

- Bidirectional regulation (cross-regulation).
- Action of B in the presence of A (if directionality is emphasized, "Regulation" may be used; otherwise, "Interaction").

2.2 Unidirectional Effect

Refers to functional influence with clear directionality.

- **Definition:** One cell causes a change in state, activation, or inhibition of another cell.
- **Examples:**
  - Expressions indicating direction, such as "effect on."
  - "A mediated by B" (Interpreted as B controlling A).

2.3 Change

Refers to the process where one cell changes into another.

- **Definition:** Differentiation or transformation from Cell A to Cell A' (or B).
- **Examples:** Differentiation from progenitor cells to mature cells.

2.4 No Relation

Refers to cases that do not fit the above categories or where a relationship is explicitly denied.

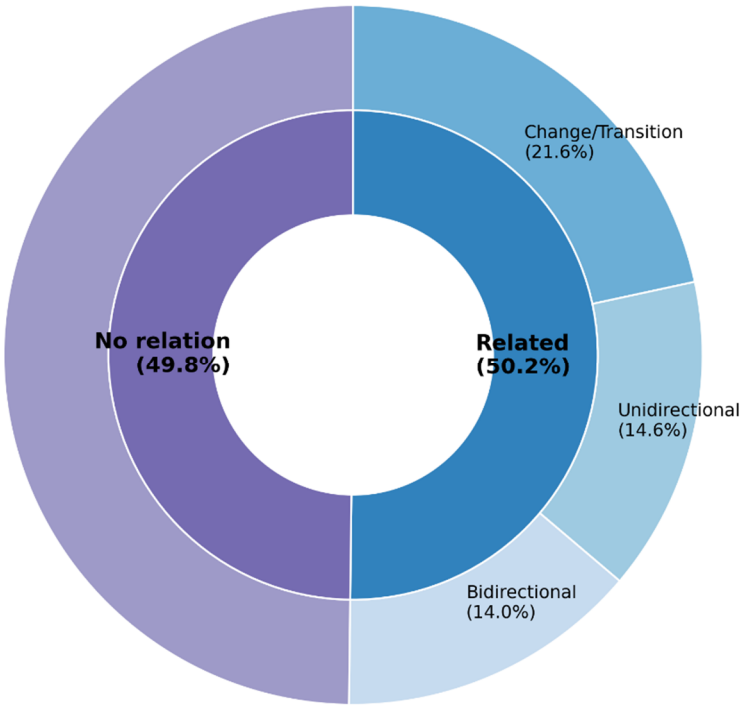
- **Definition:**
  - Simple co-occurrence without described interaction.
  - Taxonomical relationships like "A is a B".
  - Explicitly indirect relationships mediated by another cell.

2.5 Invalid

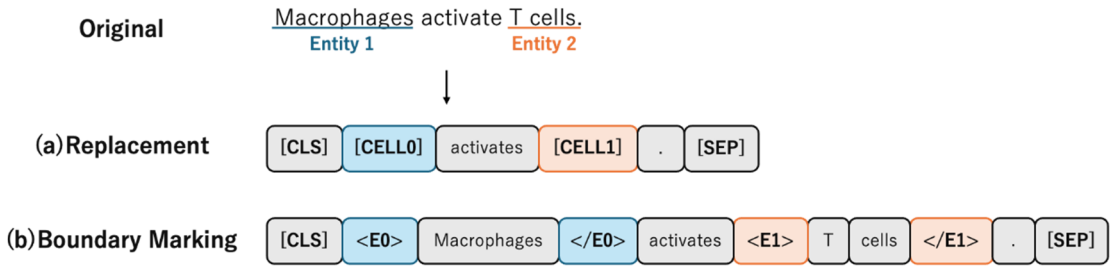
Refers to cases unsuitable for judgment in this study.

- **Examples:**
  - The number of target cells is not two (applicable only to SemMed data).
  - Descriptions of study purpose or methods.
  - Explicit descriptions of an in vitro setting.

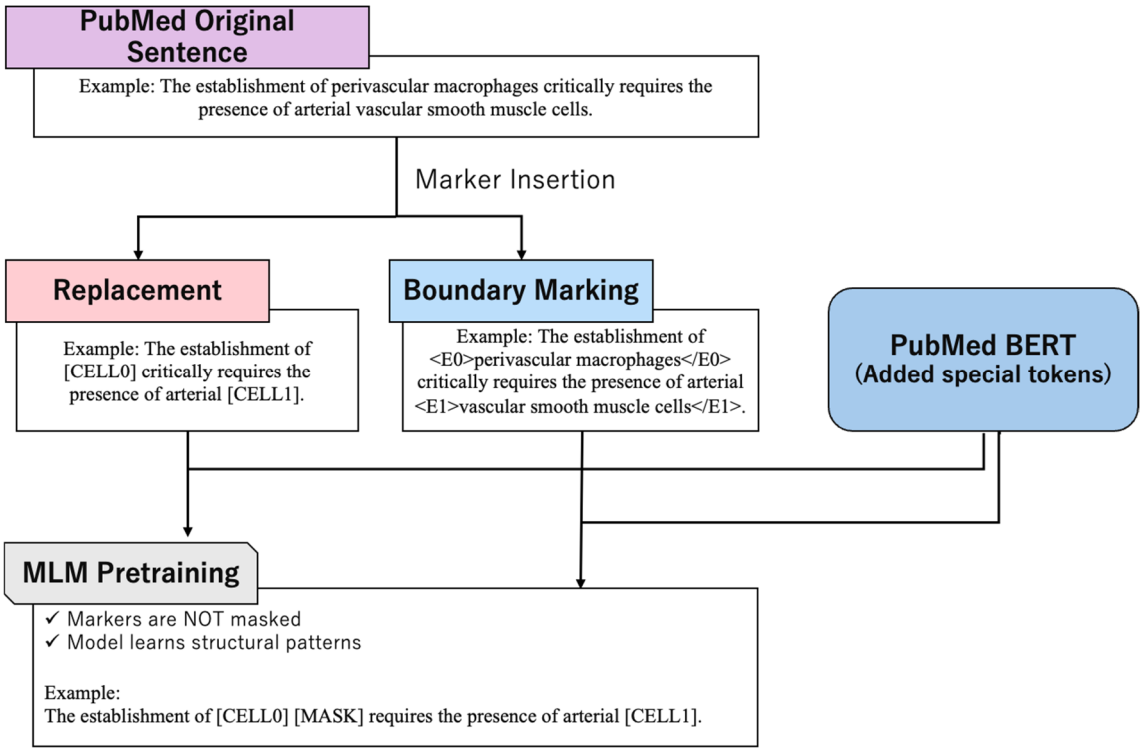
Supplementary Figures and Tables



Supplementary Figure 1. Distribution of relation labels in the SemMed-based corpus. The inner ring represents the proportion of binary classification labels (Related vs. No relation) used in this study. The outer ring displays the breakdown of the positive ("Related") instances into three detailed interaction types: Change/Transition, Unidirectional, and Bidirectional.



Supplementary Figure 2. Comparison of Entity Marking Methods.
"Original" represents the source sentence. **(a) Replacement:** A method where target cell names are completely replaced with abstract special tokens (e.g., [CELL0], [CELL1]), thereby removing entity-specific lexical information from the input. **(b) Boundary Marking:** A method where boundary tags (e.g., <E0>, </E0>) are inserted surrounding the cell names. This approach explicitly indicates entity positions while preserving the original lexical information.



Supplementary Figure 3. Overview of the Continued Pre-training (CPT) pipeline. Special tokens are inserted into original PubMed sentences using either the Replacement or Boundary Marking method. Subsequently, Masked Language Modeling (MLM) is performed with a model where these special tokens have been added to the vocabulary. Notably, the inserted special tokens are explicitly excluded from masking targets and remain visible. This design allows the model to be exposed to the structural patterns associated with entity markers during pre-training.

Your task is to determine the type of direct relationship between two cell entities, [CELL1] and [CELL2], mentioned in a given sentence.

If the sentence describes **any of the following relationship types**, output `1`:

1. **Interaction (A↔B or A–B)**
- The relationship has **no directionality** (bidirectional or mutual).
- Typical cases include:
    - Physical binding or contact
    - Mutual recognition or communication
    - Cross-regulation (e.g., both cells influence each other)
    - Situations where the presence of [CELL1] affects [CELL2] without clear directionality
- Example keywords: *interact, associate, bind, recognize, cross-talk*

2. **Regulation / Control (A→B)**
- The relationship has a **direction** (one cell influences or regulates the other).
- Typical cases include:
    - Induction, suppression, or conversion of one cell type by another
    - "Effect on", "mediated by", or similar phrases indicating causal influence
    - A affects B's behavior, differentiation, or activity
- Example keywords: *induce, promote, inhibit, mediate, regulate, convert*

3. **Differentiation (A→A')**
- One cell type differentiates into another cell type.
- Typical cases include:
    - "A differentiates into B"
    - "A becomes B"
- Example keywords: *differentiate, develop into, transform into*

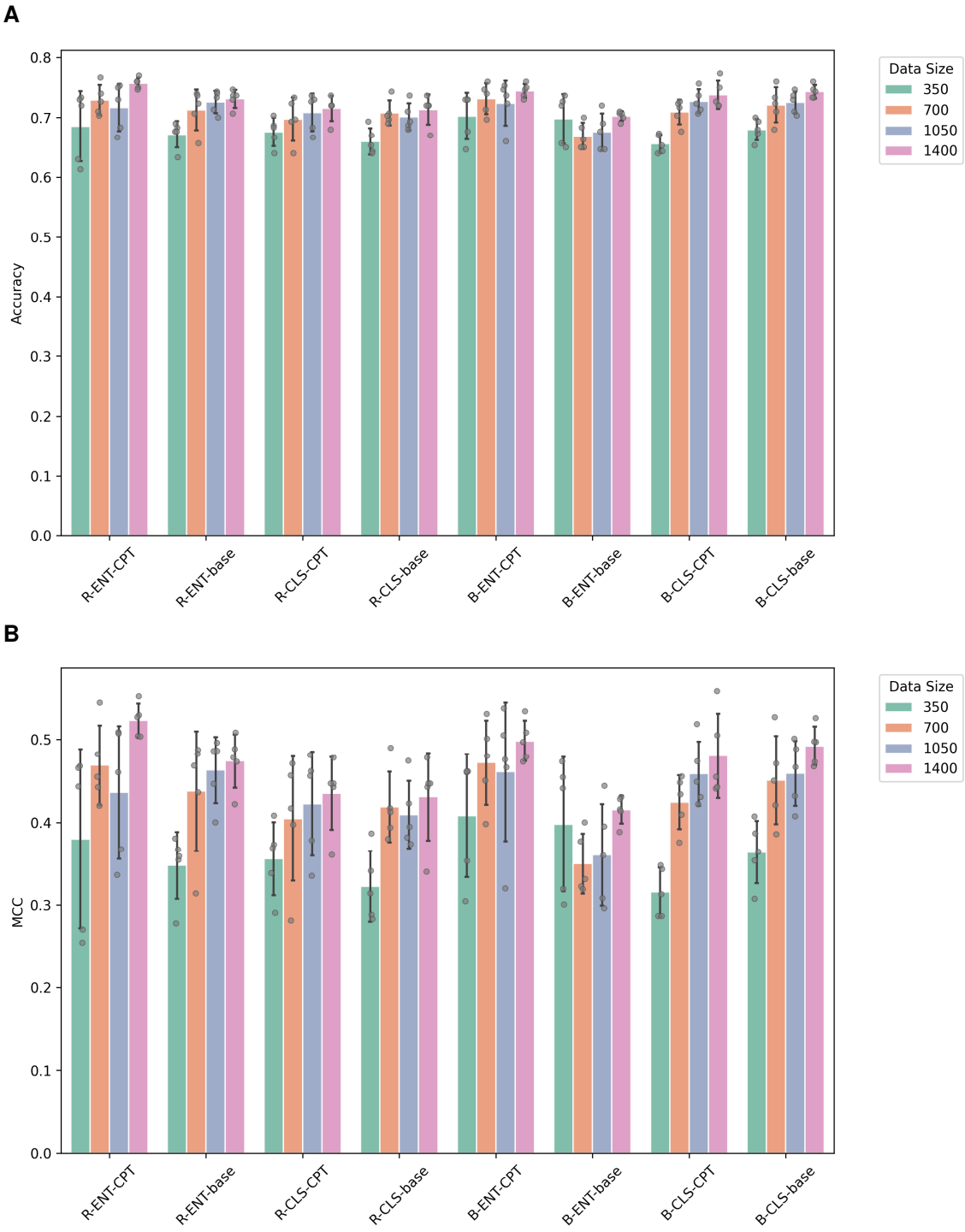
If the sentence describes **no direct biological relationship**, output `0`.
Typical cases include:
- The sentence does **not describe a biological relationship** between [CELL1] and [CELL2].
- Typical cases include:
    - "A expresses marker B" (is-a or marker description only)
    - Methodological or experimental context, not a biological interaction
    - When explicitly stated as in vitro
    - Indirect relationships
    - Statements unrelated to cell–cell relationships

Examples
- "Thus, [CELL1] from adult human pancreas can differentiate to [CELL2]." → 1
- "[CELL1] could differentiate into [CELL2] in vitro and have been shown to reconstitute the impaired myocardium in vivo." → 0
- "Living tissue imaging provides evidence of interactions between differentiating [CELL1] and [CELL2]." → 1
- "[CELL1] targeting [CELL2] were monitored by specific markers." → 1

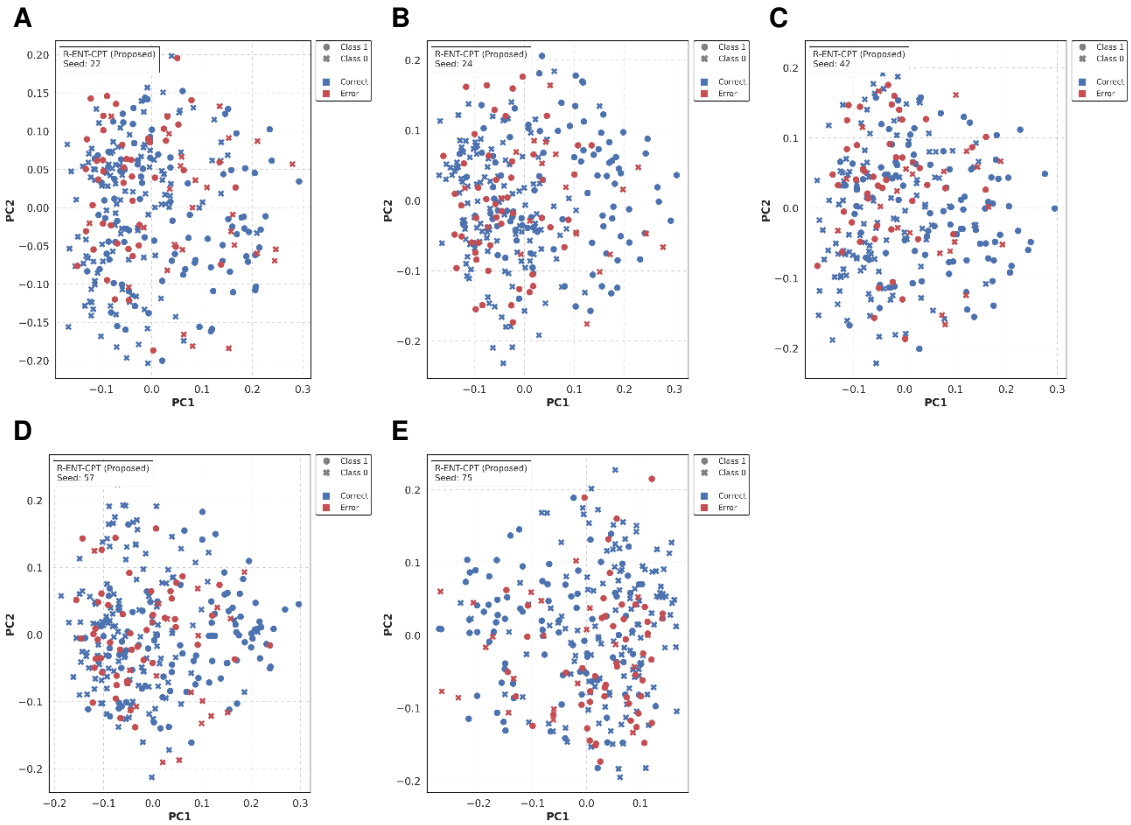
Input
Sentence: {sentence}

Output
Output `1` if there is a direct biological relationship (Interaction, Regulation / Control, or Differentiation).
Output `0` if there is no relationship.
Output only the number.

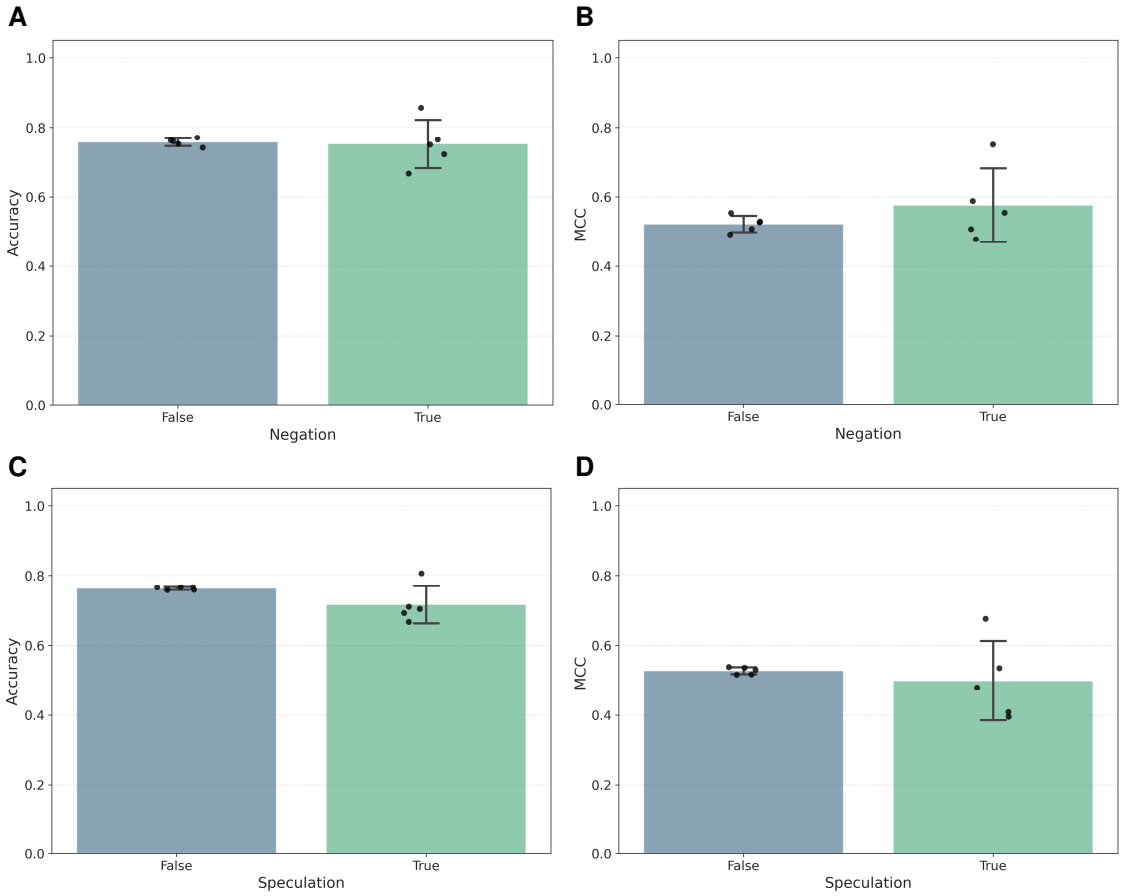
Supplementary Figure 4. Prompt Design for Large Language Models (LLMs)
A standardized prompt was used for all LLMs to perform sentence-level binary classification (Interaction vs. No Relation) given a sentence and two target entities. The prompt instructs the model to perform a binary classification task (Interaction vs. No Relation) given a sentence and two target entities.



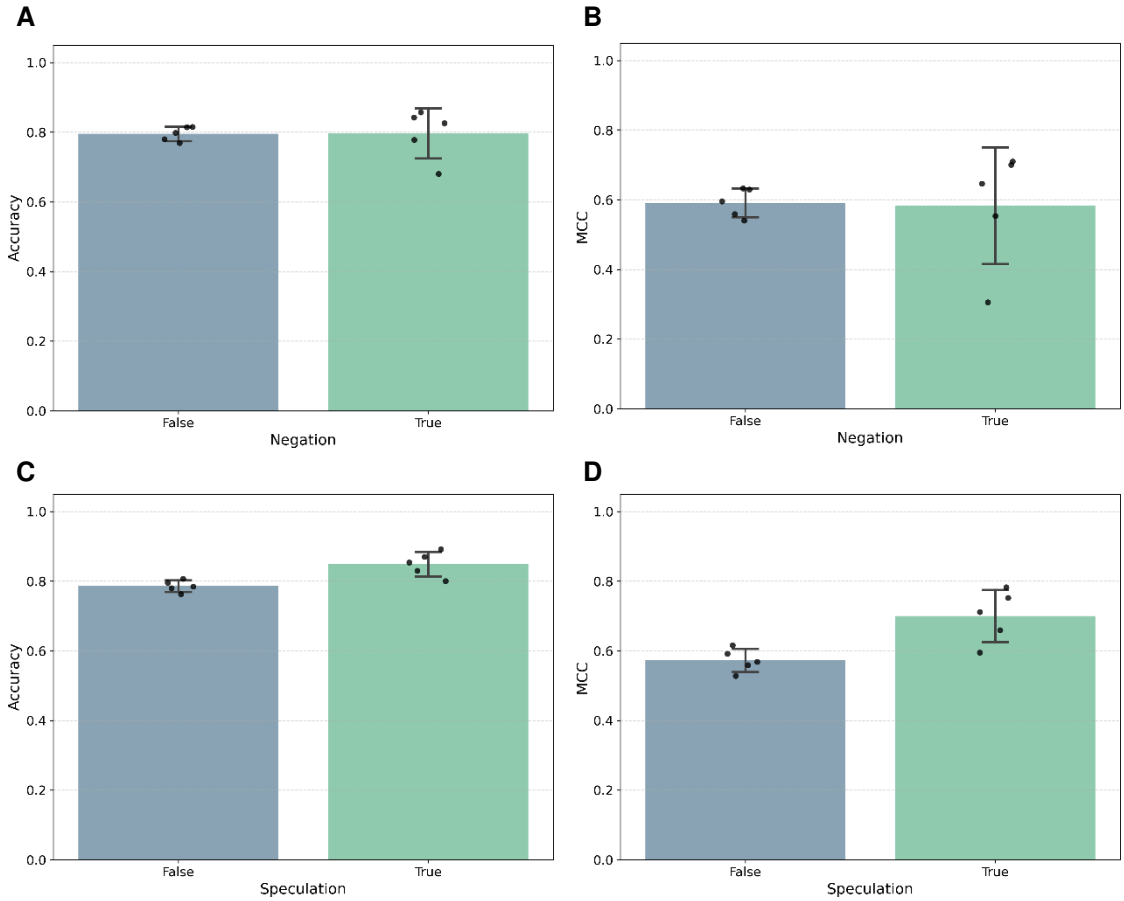
Supplementary Figure 5. Model performance across different training data sizes. Comparisons of Accuracy (A) and Matthews correlation coefficient (MCC) (B) for different model configurations as a function of training data size. The x-axis represents combinations of classification architecture and pre-training setting (with or without continued pre-training), while the legend indicates the number of training samples used (350, 700, 1050, and 1400).



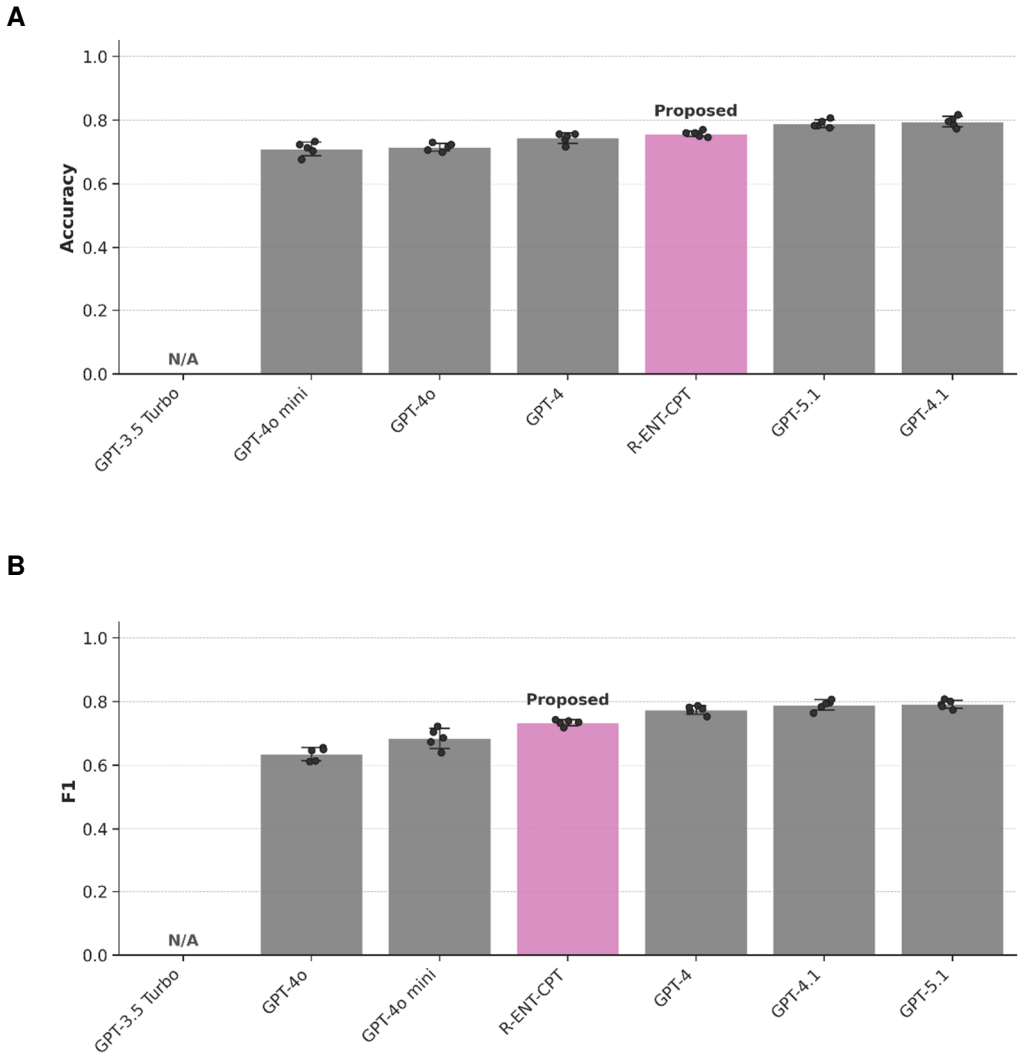
Supplementary Figure 6. PCA visualization of test embeddings and prediction outcomes. Principal component analysis (PCA) visualization of test data embeddings and prediction results for the proposed model (R-ENT-CPT). Feature vectors are projected into two-dimensional space for visualization purposes. Marker shapes indicate ground truth labels (circle: positive; cross: negative), and colors denote prediction correctness (blue: correct; red: incorrect). Each panel corresponds to a different random seed.



Supplementary Figure 7. Performance stratified by negation and speculative expressions on the proposed model.
Performance of the proposed model (R-ENT-CPT) stratified by the presence of negation (top row) and speculative expressions (bottom row). Distributions of Accuracy (left column) and MCC (right column) are shown. “False” indicates sentences without the corresponding linguistic feature, and “True” indicates sentences containing it.



Supplementary Figure 8. Performance stratified by negation and speculation on GPT-4.1 performance.
Performance of GPT-4.1 stratified by the presence of negation (top row) and speculative expressions (bottom row). Accuracy (left column) and Matthews correlation coefficient (MCC) (right column) are shown for sentences with ("True") and without ("False") the corresponding linguistic feature.



Supplementary Figure 9. Performance comparison with Large Language Models (LLMs). Evaluation results of the proposed model (R-ENT-CPT) and various OpenAI GPT models on the PubMed test set (OOD data). The panels display (A) Accuracy and (B) F1-score. The pink bar represents the proposed model, while grey bars represent general-purpose LLMs accessed via the OpenAI API. Error bars indicate the standard deviation across multiple runs. Note that GPT-3.5 Turbo is marked as "N/A" as it failed to generate valid outputs following the prompt instructions. This figure provides additional performance metrics corresponding to the comparison shown in **Figure 4**.

Supplementary Table 1. Summary of the datasets constructed in this study.
The table lists the size (number of sentences), data source, experimental purpose, and text characteristics for each corpus. **CPT:** Continued Pre-training. **ID**: In-Distribution. **OOD:** Out-of-Distribution.

Corpus	Size	Source	Purpose	Characteristics
CPT Corpus	2,121,113	PubMed	CPT	Marker insertion, large scale
SemMed RE Train Corpus	1,400	SemMed	Fine-tuning	Structured text, simple syntax
SemMed RE Test Corpus	160	SemMed	ID Test	Structured text, simple syntax
PubMed RE Test Corpus	300	PubMed	OOD Test	Natural text, complex syntax

Supplementary Table 2. List of model configurations compared in this study.
The table lists the combinations of entity marking method, classifier architecture, and the presence or absence of continued pre-training (CPT). The symbol ✓ indicates that CPT was applied. "Model Name" denotes the abbreviations used throughout the experimental results, following the naming convention [Entity Marking]–[Classifier]–[Pre-training].

Entity Marking	Classifier	CPT	Model Name
Replacement	CLS-only		R-CLS-base
Replacement	CLS-only	✓	R-CLS-CPT
Replacement	Entity-aware		R-ENT-base
Replacement	Entity-aware	✓	R-ENT-CPT
Boundary	CLS-only		B-CLS-base
Boundary	CLS-only	✓	B-CLS-CPT
Boundary	Entity-aware		B-ENT-base
Boundary	Entity-aware	✓	B-ENT-CPT

Supplementary Table 3. Performance stratified by entity distance.
Performance is reported for three entity distance bins (Short: 0–5 tokens, Medium: 6–15 tokens, Long: 16 or more tokens). Values represent mean ± standard deviation for Accuracy and Matthews Correlation Coefficient (MCC).

	Proposed		GPT-4.1	
	Accuracy	MCC	Accuracy	MCC
Short (0-5)	0.736±0.022	0.482±0.047	0.823±0.016	0.647±0.028
Medium (6-15)	0.736±0.028	0.510±0.055	0.778±0.029	0.558±0.060
Long (16+)	0.796±0.035	0.596±0.066	0.788±0.053	0.574±0.108

Supplementary Table 4. Performance stratified by sentence length.
Performance is stratified by the total number of tokens in the sentence (Short: ≤30 tokens, Medium: 31–45 tokens, Long: 46 or more tokens). Values represent mean ± standard deviation for Accuracy and Matthews Correlation Coefficient (MCC).

	Proposed		GPT-4.1	
	Accuracy	MCC	Accuracy	MCC
Short (0-5)	0.743±0.023	0.493±0.066	0.801±0.027	0.597±0.054
Medium (6-15)	0.775±0.037	0.560±0.089	0.786±0.045	0.572±0.092
Long (16+)	0.753±0.034	0.497±0.065	0.798±0.027	0.590±0.059

Supplementary Table 5. Performance comparison with Large Language Models (LLMs). The table reports the performance of the proposed task-specific model (R-ENT-CPT) and several general-purpose LLMs on the out-of-distribution (OOD) PubMed test set. Accuracy, Matthews Correlation Coefficient (MCC), and F1-score are reported. LLMs were evaluated using prompt-based inference without task-specific fine-tuning and are included as reference baselines.

Model	Accuracy	MCC	F1
GPT-4o mini	0.710±0.022	0.425±0.041	0.685±0.032
GPT-4o	0.715±0.012	0.480±0.029	0.636±0.021
GPT-4	0.744±0.017	0.506±0.032	0.774±0.013
R-ENT-CPT	0.757±0.009	0.523±0.020	0.734±0.009
GPT-5.1	0.789±0.012	0.579±0.024	**0.792±0.013**
GPT-4.1	**0.795±0.016**	**0.592±0.033**	0.789±0.016