Designing Cell–Cell Relation Extraction Models: A Systematic Evaluation of Entity Representation and Pre-training Strategies

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.

Abstract

Extracting cell–cell communication (CCC) from biomedical literature is essential for supporting experimental and computational analyses of intercellular interactions. Although Large Language Models (LLMs) have advanced many NLP tasks, their generative nature raises reliability concerns for scientific fact extraction. This highlights the need for transparent, domain-adapted relation extraction (RE) models. However, despite extensive studies in biomedical RE (BioRE), the fundamental design choices that determine performance—how to indicate entities, how to structure classification models, and whether continued pre-training (CPT) is required—have not been systematically evaluated, particularly for CCC extraction. To address this methodological gap, we conducted the first systematic analysis of core BioRE design decisions using a newly constructed CCC dataset derived from SemMedDB and an independently annotated PubMed test set for out-of-distribution (OOD) evaluation. CPT had minimal effect on CLS-only models but substantially improved entity-aware architectures. The replacement strategy generalized better to natural text than boundary marking, likely due to reduced reliance on lexical identity. Our best model achieved an OOD accuracy of 0.757±0.009, outperforming widely used models such as GPT-4o (0.715±0.012) while remaining lightweight and reproducible. This work provides the first dedicated resources and design principles for CCC extraction and establishes a practical foundation for robust BioRE model development.

Keywords: Relation Extraction, Cell–Cell Communication, BioNLP, BioRE, BERT, Continued Pre-training

1 Introduction

Cell–cell Communication (CCC) play a pivotal role in biological processes, including immune response, inflammation, tissue remodeling, development, and disease progression[1]. Advances in single-cell and spatial transcriptomics have further intensified interest in CCC, and tools such as CellPhoneDB[2], CellChat[3], and NicheNet[4] aim to infer intercellular interactions from omics data. Yet these methods still rely heavily on prior knowledge extracted manually by researchers from an ever-expanding body of literature. At the same time, the volume of CCC-related literature has grown rapidly, and the exponential expansion of PubMed[5,6] makes sustainable manual curation increasingly impractical.

Consequently, computational approaches are essential to systematically extract and structure CCC information from the vast body of scientific literature. While Large Language Models (LLMs) excel in many tasks, their generative nature poses risks of hallucination and incorrect citation[7–9], necessitating caution in scientific fact extraction. Therefore, the established Relation Extraction (RE) framework remains a robust foundation for this task. Encoder-based models like BERT[10] offer high interpretability by focusing on representation learning rather than generation. Domain-specific models pre-trained on biomedical text, such as BiomedBERT[11] and BioBERT[12], have demonstrated stability in BioNLP tasks and provide reasonable baselines for CCC extraction.

Despite these advantages, several fundamental design choices for biomedical RE (BioRE) remain unresolved. First, it is unclear how target entities should be indicated in text: whether they should be replaced with category labels or explicitly marked using special boundary tokens[11–14]. Second, there is no consensus on how relation classification models should represent sentences—either relying solely on the global [CLS] embedding or incorporating explicit representations of the entity span[15]. Third, boundary markers require introducing special tokens not seen during the original pre-training of biomedical language models, raising the question of whether fine-tuning alone is sufficient or whether additional continued pre-training (CPT) is needed to properly learn these tokens[16,17]. Each of these design decisions has the potential to substantially affect both model performance and generalization.

Furthermore, cell–cell RE is a particularly underdeveloped subfield of BioRE. Existing research predominantly focuses on gene–protein[18] or chemical–disease[19] relations; to our knowledge, no dedicated corpus for cell–cell interactions exists. Cell types possess diverse classification systems[20,21] and are highly context-dependent, making generalization difficult. Addressing this requires both rigorous dataset construction and optimized model design.

To enable reliable extraction of cell–cell interactions from biomedical literature, we conducted a systematic analysis of the key design decisions involved in building cell–cell RE models (**Figure 1**). Specifically, we evaluated three fundamental components that directly influence model performance and generalization:

1. Entity Indication: Replacement vs. Boundary Marking.
2. Classification Architecture: CLS-only vs. Entity-aware approaches.
3. Pre-training Strategy: The necessity of CPT when introducing special tokens.

In addition, we adopted an efficient annotation strategy using SemMedDB[22], which provides natural PubMed sentences paired with automatically extracted predications. This enables scalable acquisition of weakly labeled data with minimal manual effort. We employed a two-stage evaluation design: models were trained on SemMedDB-derived sentences and evaluated on raw PubMed sentences sampled independently without SemRep[23] filtering. This setup allows a direct assessment of how well models trained on SemMedDB generalize to the broader diversity of natural biomedical text, helping to clarify core design principles for BioRE and establish a foundation for automated CCC extraction.

2 Related Works

2.1 Entity marking and replacement methods

In RE, the method of specifying target entities significantly influences classification performance. R-BERT introduced boundary markers such as $ and # to extract and utilize entity span representations[15]. It has also been demonstrated that entity marker-based approaches help Transformers learn entity-centric representations[24].

Conversely, in biomedical NLP (BioNLP), replacing entity names with category labels (e.g., <cell>, CHEM) has been practically adopted to address terminological diversity and synonymy[12,14]. While this approach reduces vocabulary variance, systematic comparisons between replacement and boundary marking remain limited in the biomedical domain. Consequently, knowledge regarding the optimal strategy for domain-specific models like BiomedBERT remains insufficient. Validating these design choices is particularly critical for tasks involving homogeneous entities, such as cell–cell RE.

2.2 CLS-only vs. entity-aware classification

For BERT-based relation classification, two primary architectures exist: [CLS]-only models, which utilize the global [CLS] token representation, and entity-aware models (e.g., R-BERT), which explicitly handle entity spans defined by boundary markers in addition to the [CLS] token. CLS-only models offer simplicity and versatility, and the [CLS] representation has been demonstrated to serve as a strong baseline for RE tasks[25]. On the other hand, R-BERT-style approaches are reported to be advantageous for capturing explicit local information of entity spans[15,24].

However, the relative efficacy of CLS-only versus entity-aware models is task-dependent, and no consensus has been established for BioRE, where sentence structures are complex and entity expressions are diverse. Furthermore, for cell–cell RE, which handles relations between identical entity types, no comparative validation has been conducted, leaving significant room for investigation.

2.3 Special tokens and continued pretraining (CPT)

When using boundary marking, special tokens such as <e1> and <e2> are absent from the original BERT pre-training. Consequently, a debate exists regarding whether fine-tuning alone suffices to learn marker embeddings or if CPT is necessary. CPT involves exposing the model to sentences containing markers during pre-training to learn an input format consistent with downstream tasks; improvements have been reported in the general domain[16,17,24,26,27]. In the context of domain adaptation, additional pre-training on task-relevant data is also known to contribute to performance gains[28].

However, systematic research on the efficacy of CPT in the biomedical domain, specifically regarding the combination of domain-specific encoders like BiomedBERT with boundary marking, remains scarce. Therefore, verifying the extent to which CPT improves performance in BioRE, and how its contribution differs between CLS-only and entity-aware models, holds significant methodological value.

2.4 Cell–cell relation extraction

Biomedical RE has been widely explored for interactions such as gene–protein, protein–protein, and chemical–disease, supported by established corpora and benchmarks like CHEMPROT[29] and the BioCreative V CDR[19]. In contrast, no dedicated dataset or task formulation exists for cell–cell interactions, despite their biological importance.

Consequently, key design choices—such as how to represent cell entities or which model architectures are most effective—remain unexamined for this specific class of relations. Addressing this gap requires constructing an appropriate corpus and systematically evaluating model configurations tailored to cell–cell RE.

2.5 SemMedDB as a weakly supervised resource

SemMedDB[22] consists of natural PubMed sentences paired with subject–predicate–object tuples automatically extracted by the rule-based SemRep[23] system. This makes it a potentially valuable source of weakly supervised data for BioRE. However, SemRep preferentially extracts sentences with clearer predicate structures, introducing selection biases that may differ from the broader linguistic diversity of PubMed. As a result, the generalizability of models trained on SemMedDB-derived sentences to unfiltered natural text has not been systematically evaluated, motivating an explicit comparison between SemMedDB and independently sampled PubMed sentences.

3 Datasets

Due to the lack of existing benchmarks for cell–cell RE, four specific datasets were constructed for this study (**Table 1**). These comprise (1) an unlabeled dataset for CPT, (2) a supervised dataset for fine-tuning, (3) a test set derived from the same database, and (4) an independent test set for evaluating generalization performance. The dataset construction pipeline is illustrated in **Figure 2**. To ensure consistency and reliability, explicit guidelines were defined, and manual annotation was conducted from scratch.

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.

3.2 Training/Test corpus (SemMed-based supervised dataset)

The supervised dataset for fine-tuning was constructed using the Semantic MEDLINE Database (SemMedDB)[22]. SemMedDB contains subject-predicate-object tuples (Predications) extracted from PubMed abstracts using SemRep. This structured information was utilized to retrieve PubMed sentences where cell types appear as subjects or objects. Leveraging Predications allowed for high-precision extraction of candidate relations and significantly improved annotation efficiency.

These candidate sentences were manually labeled based on the guidelines described above. Consequently, a supervised corpus of 1,560 sentences (with a nearly balanced positive/negative ratio and detailed interaction subtypes; **Figure 3**) was obtained, allocated as 1,400 for training and 160 for testing. The training data was further split into train/dev sets for fine-tuning. SemMed-derived sentences tend to be structured with relatively explicit cell–cell relations, which facilitated efficient and reliable annotation.

3.3 Test corpus (PubMed-based natural text dataset)

To properly evaluate model generalization, measuring performance in natural contexts differing from structured SemMed contexts is crucial. Sentence tokenization was performed on the entire PubMed database, and sentences containing cell names were extracted based on a

cell-type dictionary. From these, 680 sentences were randomly sampled. To ensure reproducibility, random seeds were fixed, and 300 sentences were selected as the final test set.

Natural sentences extracted via dictionary matching without SemMed exhibit high lexical and syntactic diversity, often containing implicit relationship descriptions. These characteristics reflect the difficulties faced in actual literature search environments. Fine-tuning on SemMed (structured text) and testing on PubMed (natural text) allows for a direct evaluation of out-of-distribution (OOD) performance. Annotation was conducted using the identical guidelines to those in the previous section.

3.4 CPT corpus (PubMed-based unlabeled dataset)

CPT was conducted to adapt the model to sentence structures containing boundary markers. Approximately 2 million sentences containing cell names were extracted from the entire PubMed database using the cell-type dictionary.

For these sentences, two cell mentions identified by the dictionary were randomly selected, and markers (e.g. <E1>, </E1>, [CELL0]) were inserted. These markers were excluded from the Masked Language Modeling (MLM) masking targets. BiomedBERT pre-training was continued using this processed corpus, designed to stabilize marker embeddings during fine-tuning with entity-aware models.

4 Methods

In this study, three primary design elements influencing the performance and behavior of cell–cell RE were systematically compared (**Figure 1**):

1. Entity indication methods (replacement vs. boundary marking)
2. Classification architectures (CLS-only vs. entity-aware)
3. Pre-training strategies regarding special tokens (with or without CPT)

This section describes the task formulation, model configurations, CPT, fine-tuning, and evaluation procedures.

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. Cell–cell RE possesses the following characteristics:

- Both entities belong to cell types, and the relationship can be symmetric.
- Cell names exhibit high lexical diversity and context dependence.
- Simple co-occurrence does not imply a relationship, necessitating context interpretation.

These properties directly influence the design of relation extraction methods, including entity indication, representation extraction, and pre-training strategies.

4.2 Entity marking, model architectures, 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] (**Figure 4a**).

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

This approach reduces vocabulary variance and mitigates the out-of-vocabulary (OOV) problem. However, it may result in the loss of biological nuances contained in the actual cell names, potentially reducing the information available for context interpretation.

Boundary marking (R-BERT style)
Target entities are enclosed by markers, such as <E0>...</E0> and <E1>...</E1> (**Figure 4b**).

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 5a**).

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

While the structure is simple and training is stable, it does not explicitly utilize local entity information, which may limit performance in context-dependent tasks like cell−cell RE.

Entity-aware classifier (R-BERT)
This architecture incorporates entity-specific representations in addition to the global [CLS] token (**Figure 5b**). 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:

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

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 (CPT)
To address the instability of embeddings when introducing boundary markers, MLM pre-training was continued using approximately 2 million processed sentences from PubMed, into which markers were inserted (**Figure 6**). The markers were excluded from the masking targets to expose the model to the structural format consistent with the downstream task.

The objectives of CPT were:

- Stabilization of marker embeddings.
- Performance improvement of models utilizing span information.
- Ensuring structural consistency during fine-tuning.

The impact of CPT was evaluated for both CLS-only and entity-aware models.

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[30]. The parameters of BiomedBERT (base or CPT version) were updated to optimize for the relation classification task.

Optimization Details:

- Optimizer: RAdamScheduleFree[31]
- Batch size: 16 (fine-tuning), 128 (CPT)
- Learning rate: 2e-5 (fine-tuning), 2e-5 (CPT)
- Epochs: Max 20 (fine-tuning), Max 2 (CPT)
- Early stopping: Based on development set loss
- Reproducibility: Training was conducted with multiple random seeds.

4.4 Evaluation

The test corpus of 680 cell–cell RE sentences extracted and annotated from PubMed was used 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.

The following metrics were used:

- Accuracy
- F1 score
- Matthews correlation coefficient (MCC)

MCC was selected as the primary metric due to its robustness against class imbalance.[32]

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 (**Table 2**). This configuration allows for a comprehensive analysis of the effects of each design element and their interactions.

5 Results and Discussion

5.1 Impact of Classification Architecture and CPT

To clarify the dominant factors affecting model performance, we first examined the interaction between classification architecture and CPT. Using the Replacement method as a representative setting (**Table 3**), CLS-only models showed little sensitivity to CPT (0.713±0.025 without CPT vs. 0.715±0.021 with CPT). In contrast, entity-aware models benefited substantially from CPT, improving from 0.731±0.015 to 0.757±0.009.

A similar pattern was observed under the Boundary method: CPT had no meaningful effect on CLS-only models (0.743±0.011 vs. 0.738±0.023), whereas it markedly improved the performance of entity-aware models (0.703±0.008 to 0.745±0.012).

These consistent trends indicate that CPT primarily enhances architectures that rely on local entity information. Because the [CLS] representation aggregates global sentence-level

context[33], it is less affected by the introduction of special entity markers. Entity-aware models, however, depend on span-level signals and therefore benefit from pre-training that exposes the encoder to sentences containing these markers.

5.2 Effect of Entity Marking Methods

Having established that architecture and CPT exert the strongest influence on performance, we next evaluated whether the choice of entity marking method contributes additional differences (**Table 3**). Under the best-performing condition (entity-aware + CPT), Replacement achieved slightly higher accuracy (0.757±0.009) than Boundary (0.745±0.012). Without CPT, the performance gap widened, as the entity-aware Boundary model dropped to 0.702 compared with 0.731±0.015 for Replacement.

These results suggest that boundary marking may increase reliance on specific surface forms of cell names, potentially leading to overfitting to cell-pair co-occurrence patterns. In contrast, Replacement obscures lexical identity and encourages models to focus more on syntactic and contextual cues, which appears to support better generalization to diverse natural text.

5.3 Evaluating generalization beyond weakly supervised training data

To clarify how well models trained on SemMedDB generalize beyond this weakly supervised source, we compared performance on the SemMedDB-derived test set with that on the independently sampled PubMed dataset (**Table 4**). As expected, all models achieved substantially higher accuracy on the SemMed set (typically >0.8), reflecting the fact that SemRep preferentially extracts sentences with clearer predicate–argument structure. This yields a more regular linguistic distribution that is closely aligned with the training data.

In contrast, performance on the PubMed dataset revealed larger gaps between model configurations, indicating that the more diverse and syntactically varied natural sentences expose differences in robustness. The pronounced drop in accuracy from SemMed to PubMed highlights a key limitation of relying solely on SemMedDB for weak supervision: although it enables efficient dataset construction and supports high in-distribution performance, it does not fully represent the complexity of real biomedical text. Consequently, evaluating models on unfiltered PubMed sentences is essential for assessing their practical utility in literature-wide CCC extraction.

5.4 Comparison with Large Language Models

Figure 7 and **Table 5** compare our best-performing model (R-ENT-CPT) with several Large Language Models (LLMs) accessed via the OpenAI API[34]. This comparison is intended not as a direct benchmarking exercise—LLMs differ substantially in scale, training objectives, and inference cost—but rather to contextualize the performance of our lightweight, domain-adapted model relative to widely used general-purpose systems.

Clear differences in extraction behavior were observed across model generations. GPT-3.5 was excluded because it frequently failed to follow prompt instructions, making quantitative evaluation unreliable. The currently deployed GPT-4o-mini (Accuracy 0.710±0.022) and GPT-4o (0.715±0.012) showed a conservative extraction style with many false negatives, leading to lower accuracy than our model. GPT-4 achieved slightly higher accuracy (0.744) but tended to generate more false positives. In contrast, our model (0.757±0.009) achieved a balanced trade-off and outperformed these general-purpose models.

The latest large-scale models, GPT-4.1 (0.795±0.016) and GPT-5.1 (0.789±0.012), surpassed our model in accuracy. However, their computational cost, latency, and API-based access make them impractical for large-scale literature mining. In contrast, our approach provides a transparent, reproducible, and cost-efficient solution that can be deployed locally and scaled to millions of PubMed sentences—an essential requirement for CCC extraction at corpus level.

6 Conclusion

This study provides the first dedicated resource and systematic analysis for extracting CCC from biomedical literature by constructing a high-quality dataset and systematically evaluating the key design choices that underlie BioRE. Our analysis demonstrated that classification architecture and CPT are the primary determinants of performance: entity-aware models benefited substantially from CPT, whereas CLS-only models showed minimal gains. Entity-indication strategies had smaller effects overall, although the replacement method consistently provided stronger generalization to diverse natural PubMed sentences.

The resulting model achieved accuracy competitive with, and in several cases surpassing, widely used general-purpose LLMs while remaining lightweight, transparent, and deployable for large-scale literature mining—properties essential for practical CCC extraction. These findings provide actionable design principles for building robust BioRE systems and establish a foundation for automated mining of cell–cell interactions to support downstream computational analyses of intercellular communication.

A remaining limitation is that the weakly supervised SemMedDB-derived training data do not fully capture the linguistic diversity of biomedical text, due to selection effects introduced by the SemRep extraction pipeline. Future work will address this by constructing more varied corpora and developing methods that better accommodate distributional differences between weakly supervised and natural-text settings.

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.

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.

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. 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).
11. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **3**, 1–23 (2022).
12. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
13. 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).

14. 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).
15. 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.
16. 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.
17. 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.
18. Pyysalo, S. *et al.* BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* **8**, 50 (2007).
19. Li, J. *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)* **2016**, baw068 (2016).
20. Tan, S. Z. K. *et al.* The Cell Ontology in the age of single-cell omics. *ArXiv* arXiv:2506.10037v2 (2025).
21. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, (2016).
22. 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).
23. 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).
24. 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).

25. Shi, P. & Lin, J. Simple BERT models for relation extraction and semantic role labeling. *arXiv [cs.CL]* (2019) doi:10.48550/arXiv.1904.05255.

26. 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.

27. 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.

28. 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).

29. Taboureau, O. *et al.* ChemProt: a disease chemical biology database. *Nucleic Acids Res.* **39**, D367-72 (2011).

30. 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.

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

32. 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).

33. 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).

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

Figures and Tables

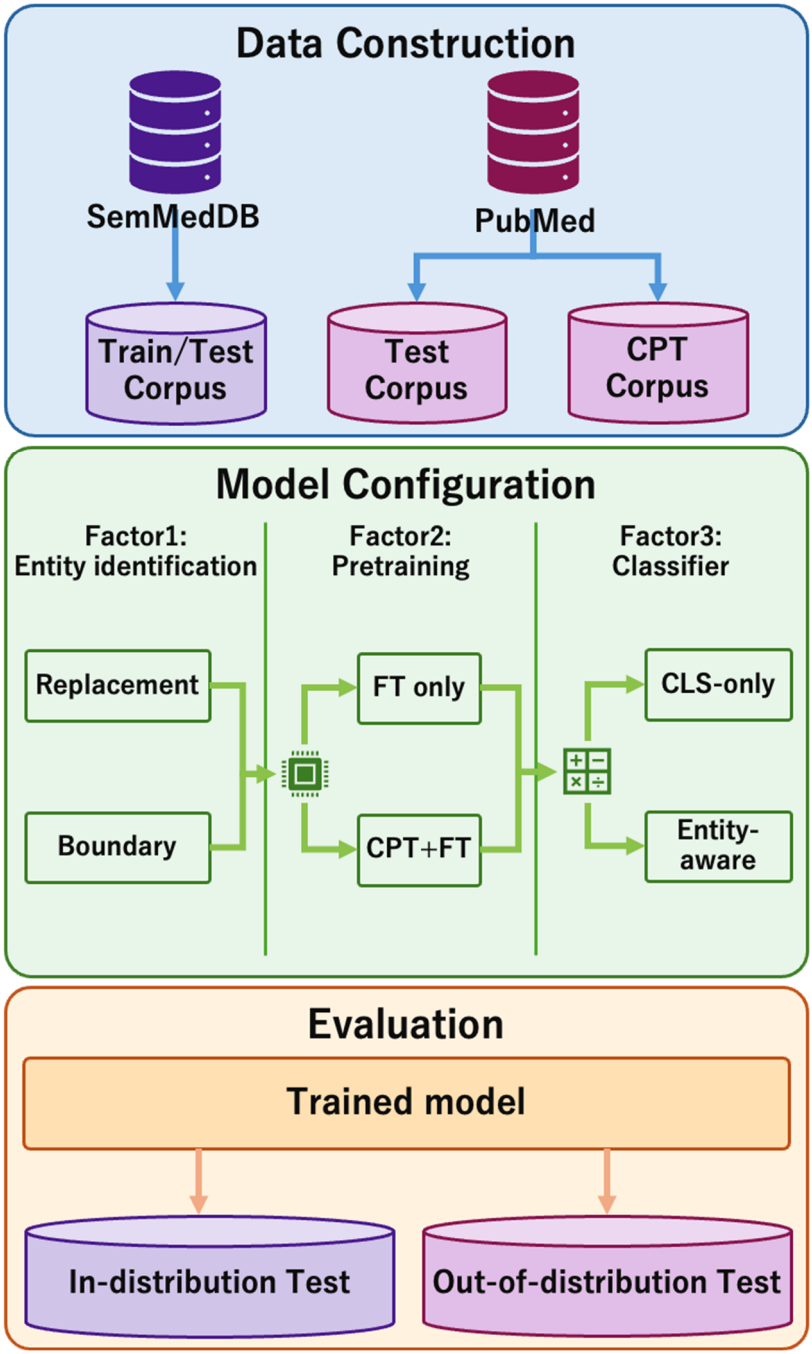

Figure 1. Schematic overview of the study design.
The experimental workflow consists of three phases: (A) Dataset Construction, where purpose-specific datasets are constructed from SemMedDB and PubMed; (B) Model Configuration, systematically comparing combinations of Entity Indication (Replacement vs. 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.

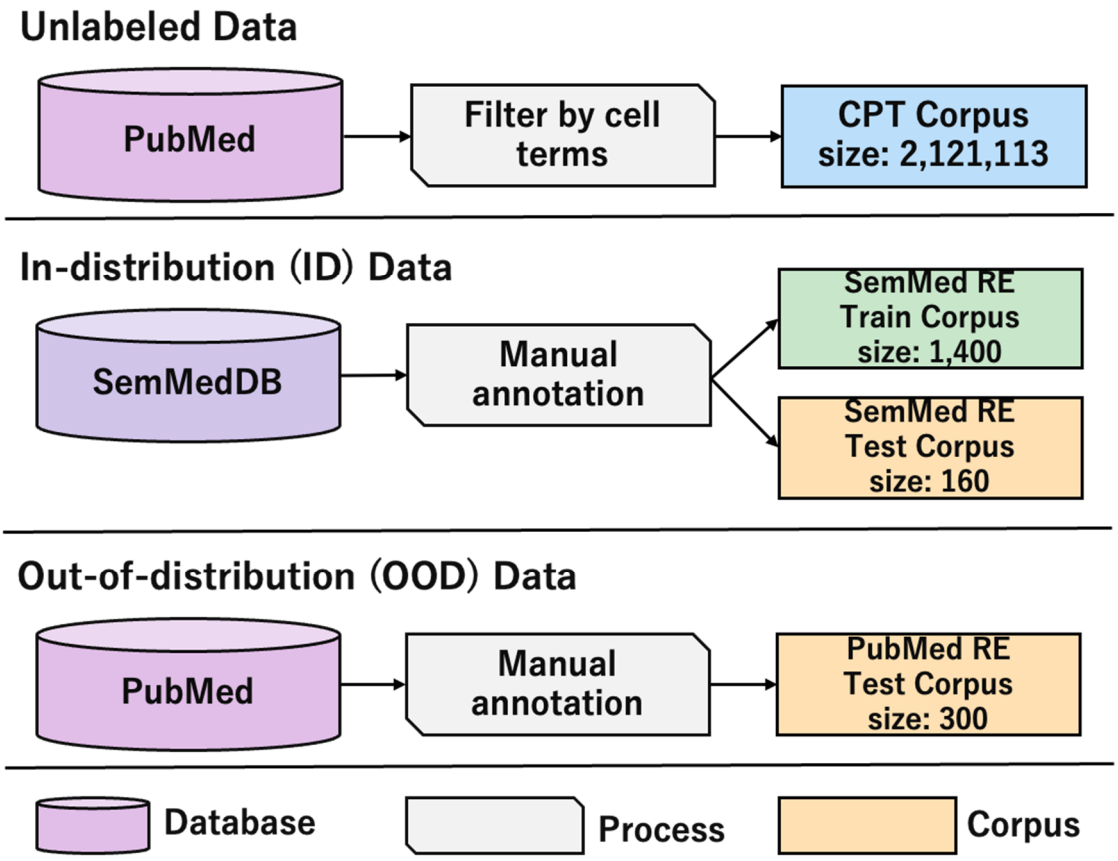

Figure 2. Dataset construction pipeline.
The figure illustrates the process of generating final corpora from their respective data sources. Top: Construction of the unlabeled corpus for Continued Pre-training (CPT) via dictionary-based filtering from PubMed. Middle: Construction of the supervised training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

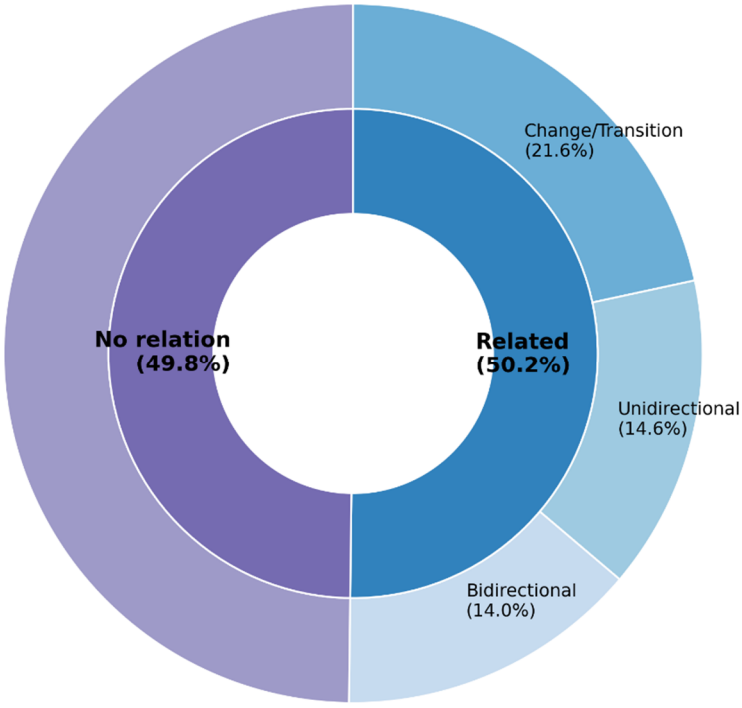

Figure 3. Distribution of relation labels in the SemMed-based corpus.
The inner ring represents the proportion of binary classification labels (Related vs. No relation) used in this study. The outer ring displays the breakdown of the positive ("Related") instances into three detailed interaction types: Change/Transition, Unidirectional, and Bidirectional.

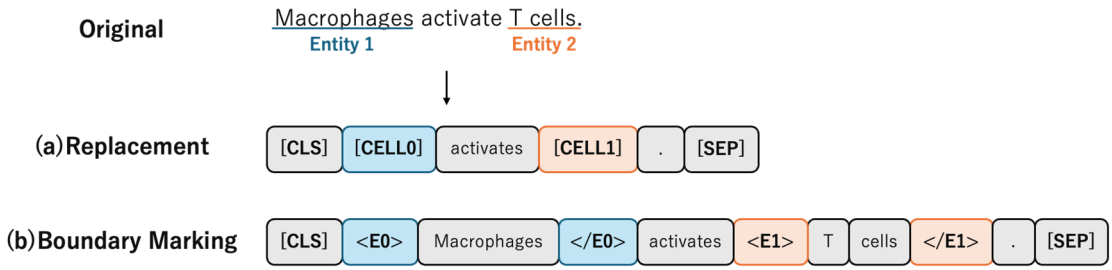

Figure 4. Comparison of Entity Marking Methods.
"Original" represents the source sentence. **(a) Replacement:** A method where target cell names are completely replaced with abstract special tokens (e.g., [CELL0], [CELL1]), resulting in the loss of entity-specific lexical information. **(b) Boundary 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.

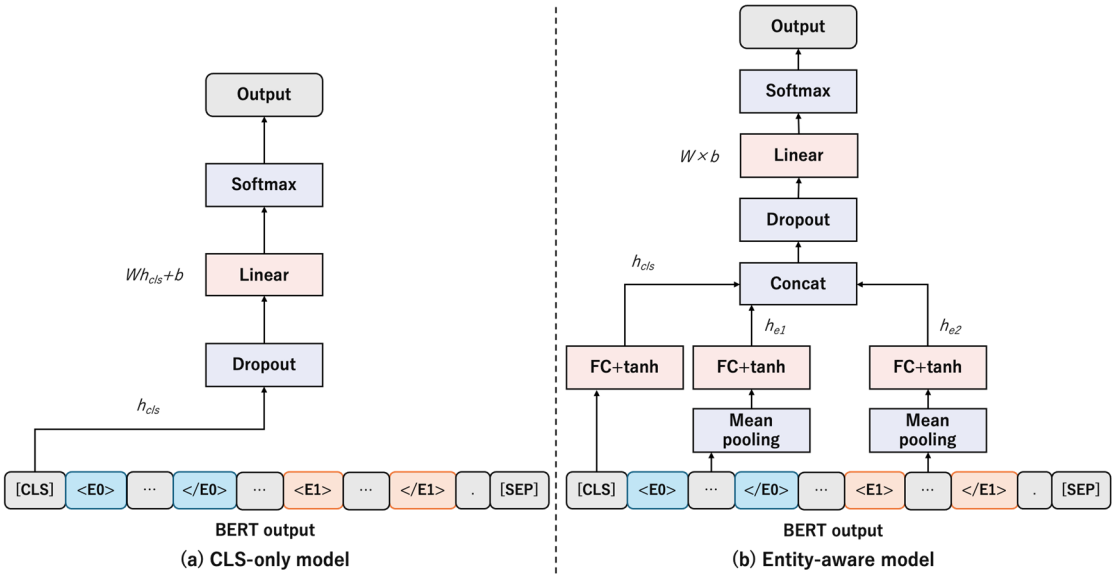

Figure 5. Schematic of classification architectures.
(a) CLS-only model: Performs classification utilizing only the global representation corresponding to the [CLS] token. (b) Entity-aware model: In addition to the [CLS] token, entity-specific representations are extracted via mean pooling of token embeddings 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.

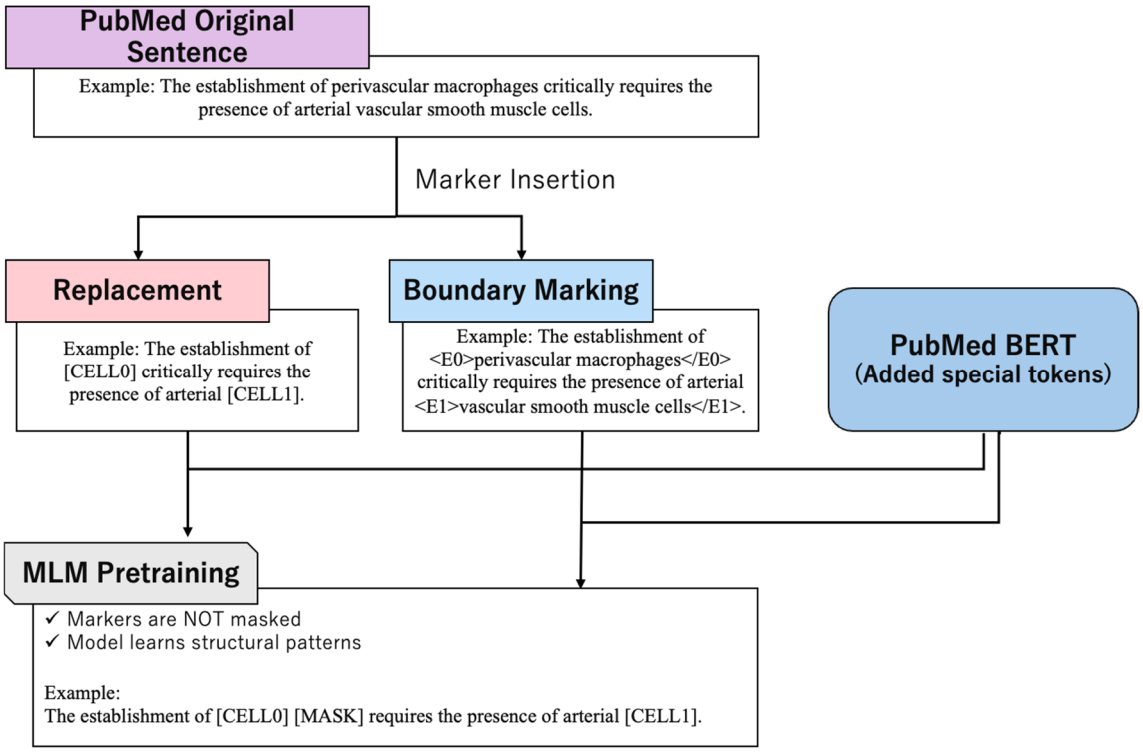

Figure 6. Overview of the Continued Pre-training (CPT) pipeline.
Special tokens are inserted into original PubMed sentences using either the Replacement or Boundary Marking method. Subsequently, Masked Language Modeling (MLM) is performed with a model where these special tokens have been added to the vocabulary. Crucially, the inserted special tokens are explicitly excluded from masking targets and remain visible. This strategy enables the model to effectively learn the structural semantics and patterns associated with the markers during the pre-training phase.

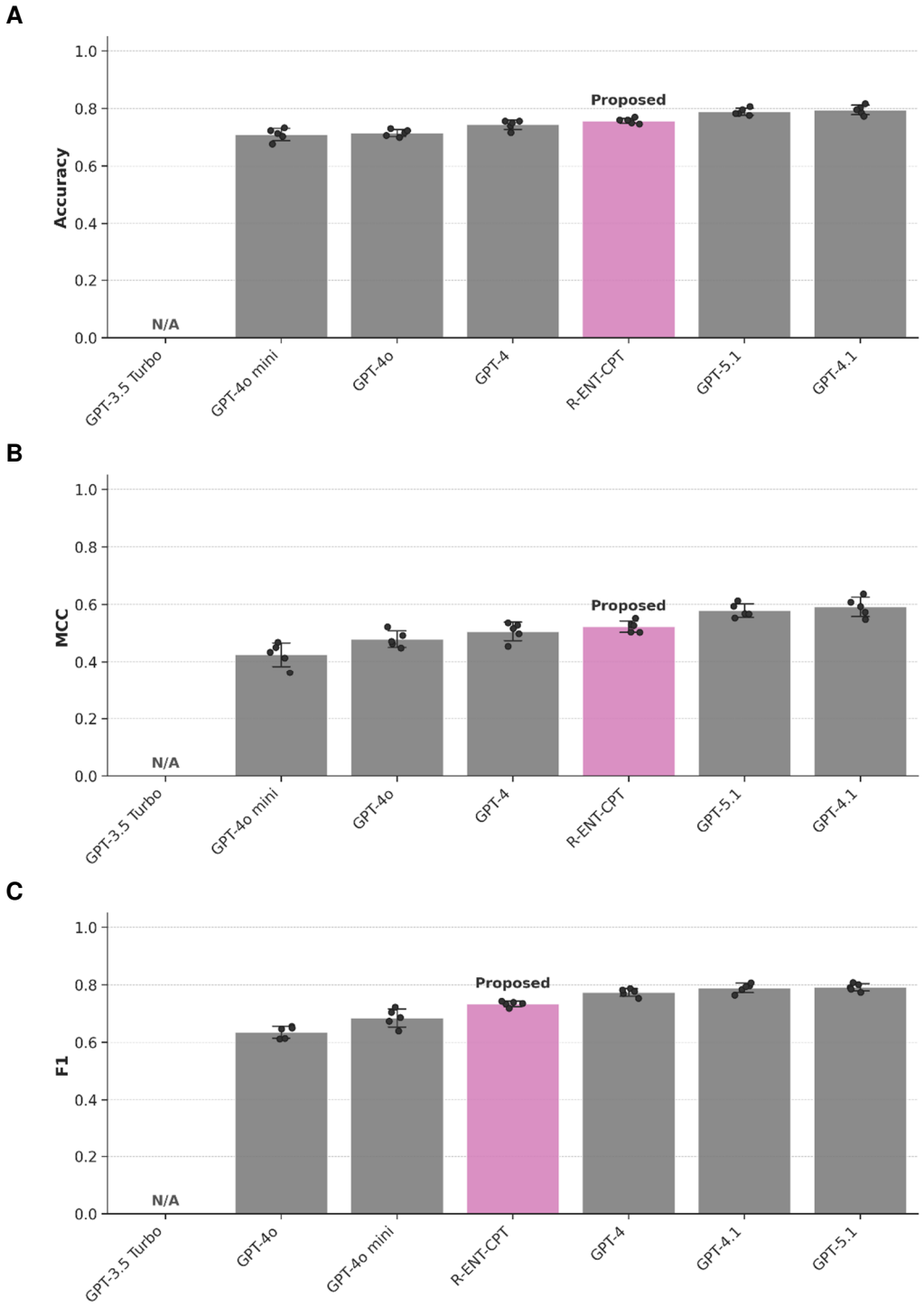

Figure 7. Performance comparison with Large Language Models (LLMs).
Evaluation results of the proposed model (R-ENT-CPT) and various OpenAI GPT models on the PubMed test set (OOD data). The panels display (A) Accuracy, (B) Matthews Correlation Coefficient (MCC), and (C) F1-score. The pink bar represents 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.

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

Table 2. List of model configurations compared in this study.
The table presents the combinations of Entity Marking, Classifier architecture, and the presence of CPT. The symbol ✓ indicates the application of CPT. "Model Name" denotes the abbreviations used in the subsequent experimental results (Figures 7, Table 3-5), following the naming convention of [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

Table 3. Detailed scores of BERT-based models on the PubMed test set (Ablation Study). List of Accuracy, MCC, and F1 scores for all eight models. This table quantitatively demonstrates the contribution of each component.

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 4. Performance evaluation on the SemMed dataset and the generalization gap. The table reports the performance on the in-distribution (ID) SemMed test set and the performance drop (Δ) observed when evaluated on the out-of-distribution (OOD) PubMed test set. Δ represents the generalization gap, calculated as $Score_{SemMed} - Score_{PubMed}$. A smaller Δ indicates higher robustness against domain shifts.

Model	Accuracy	ΔAccuracy	MCC	ΔMCC	F1	ΔF1
R-ENT-CPT	**0.843±0.041**	0.085	**0.690±0.081**	0.167	**0.852±0.037**	0.118
R-ENT-base	0.840±0.053	0.109	0.682±0.106	0.208	0.845±0.052	0.145
R-CLS-CPT	0.816±0.030	0.101	0.637±0.059	0.202	0.825±0.028	0.130
R-CLS-base	0.818±0.030	0.104	0.638±0.059	0.207	0.826±0.028	0.131
B-ENT-CPT	0.825±0.057	0.080	0.655±0.115	0.157	0.836±0.053	0.117
B-ENT-base	0.820±0.034	0.117	0.642±0.068	0.227	0.825±0.036	0.157
B-CLS-CPT	0.829±0.038	0.091	0.662±0.076	0.182	0.839±0.035	0.116
B-CLS-base	0.840±0.023	0.097	0.684±0.046	0.192	0.848±0.021	0.124

Table 5. Performance comparison with Large Language Models (LLMs).
Evaluation results of the proposed model (R-ENT-CPT) and general-purpose LLMs on the PubMed test set (out-of-distribution data). The table reports Accuracy, Matthews Correlation Coefficient (MCC), and F1-score. R-ENT-CPT denotes the proposed model employing entity replacement, an entity-aware architecture, and continued pre-training.

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