Applicability Condition Extraction for Therapeutic Drug-Disease Relations

Guanting Luo[1,2] **Noriki Nishida**[2] **Yuji Matsumoto**[2,4] **Yuki Arase**[3,2]
[1]The University of Osaka [2]RIKEN [3]Institute of Science Tokyo [4]Tohoku University
guanting.luo@ist.osaka-u.ac.jp
noriki@norikinishida.com
yuji.matsumoto@riken.jp
arase@c.titech.ac.jp

Abstract

Identifying conditions that a certain drug takes therapeutic effect on a target disease is crucial for clinical decision-making support. However, most existing biomedical information extraction methods have focused on identifying only relations between drugs and diseases, while largely overlooking the context-specific conditions where such relations can apply. To address this problem, we introduce the task of applicability condition extraction for therapeutic drug–disease relations from biomedical research literature. We create the first dataset that has manually annotated triples of drugs, diseases, and applicability conditions on biomedical paper abstracts with 1,119 drug-disease pairs. Using this dataset, we systematically evaluate the performance of a range of existing methods. In addition, we propose a new method that enhances LoRA to consider relations between drugs and diseases. Our method consistently outperforms strong baselines across different evaluation settings. The source code and dataset of this paper can be obtained from: https://github.com/guantingluo98/Drug-ACE

Example of an Annotated Instance

Title: Hydroxyurea in stage D carcinoma of the prostate: a pilot study.
Abstract: There was 13 patients with histologically metastatic **prostatic adenocarcinoma** treated with a single oral dose of 80 mg. per kg. **hydroxyurea** every third day (based on ideal or actual weight, whichever is less) and 12.5 mg. chlorotrianisene per day. Toxicity was mild. The most common manifestations were nausea, occasional vomiting leukopenia. A definite attempt was made to depress the white blood count to approximately 2,000 cells per cu. mm. Hydroxyurea was not discontinued unless the white blood count decreased to less than 2,000 cells per cu. mm., after which a single dose was usually omitted. Omission of a single dose would allow the white blood count to return promptly to more than 2,000 cells per cu. mm. Objective tumor regression was demonstrated in 6 of the 13 patients and all patients had a definite improvement in the quality of life.
Drug–Disease Pair: (Hydroxyurea, prostatic adenocarcinoma)
Applicable Condition: "a single oral dose of 80 mg. per kg. hydroxyurea every third day"
Condition Type: Dosage

Figure 1: An illustrative example of an annotated instance in the Drug-ACE dataset.

1 Introduction

Therapeutic drug–disease relations play a central role in clinical practice and biomedical research, forming the foundation for treatment selection and evidence-based medical decision-making. In real-world clinical settings, however, the applicability of a drug is rarely universal across all patient populations. Whether a drug can be effective to treat a disease often depends on specific patient profiles and contextual factors, reflecting substantial patient heterogeneity. Therefore, it is crucial to identify conditions under which a drug can be effectively and safely applied to treat a target disease. These applicability conditions are critical for translating biomedical evidence into practical clinical decision-making and for accurately interpreting therapeutic claims reported in the biomedical research literature. However, such applicability condition extraction has been little explored despite abundant research efforts on drug-disease relation extraction (Wei et al., 2016; Nguyen and Verspoor, 2018; Bonner et al., 2022; Luo et al., 2022; Wang et al., 2024).

In biomedical research literature, therapeutic evidence is frequently reported in a conditional manner rather than as universally applicable conclusions (Lu, 2011). The effectiveness of a drug is often qualified by specific conditions, such as dosage, patient populations, physiological characteristics, comorbidities, or genetic background. These conditions reflect the inherent diversity of patients and the complexity of disease mechanisms, and are essential for accurately interpreting therapeutic claims (Weinshilboum and Wang, 2017). Unfor-

Findings of the Association for Computational Linguistics: ACL 2026, pages 3135–3148

tunately, such applicability conditions are rarely stated in an explicit manner in a document (Figure 1). Instead, they are often distributed across sentences, embedded within broader experimental or clinical descriptions, and expressed implicitly through contextualized evidence. As a result, understanding therapeutic applicability requires reasoning over lengthy and nuanced textual contexts.

Despite substantial progress in biomedical information extraction, most existing studies have primarily focused on identifying whether a certain relation exists between biomedical entities (Roy and Pan, 2021; Jin et al., 2022; Xiao et al., 2024), or on extracting specific phenomena such as adverse effects (Alimova and Tutubalina, 2019; Henry et al., 2020; D'Oosterlinck et al., 2023; Sahoo et al., 2024). However, the conditions under which the relation is applicable or not have been little explored. As a result, current biomedical information extraction frameworks often provide incomplete representations of therapeutic knowledge, limiting their usefulness for clinical decision support.

To address this problem, we extract conditions for therapeutic drug–disease relations. Specifically, we create the Drug–Disease Applicability Condition Extraction (Drug-ACE) dataset as illustated in Figure 1. The dataset contains $1,119$ instances, each associated with a therapeutic drug–disease pair and corresponding PubMed paper titles and abstracts. Each instance is annotated with the conditions under which the given drug can treat or alleviate the target disease. We also propose a Role-Conditioned LoRA that explicitly incorporates the relation role between the drug and disease into parameter-efficient low-rank adaptation (LoRA) (Hu et al., 2022). Our comprehensive benchmarking study on the Drug-ACE dataset comparing existing biomedical relation extraction methods reveals that our method consistently outperforms strong baselines.

Our contributions are threefold:

- We introduce the task of drug–disease applicability condition extraction and release Drug-ACE, annotating $1,119$ instances.
- We deliver a comprehensive evaluation of conventional biomedical relation extraction methods on Drug-ACE, including span-based models, LoRA-tuning and prompting large language models.
- We propose a method for applicable condition extraction of drug-disease relations, which consistently outperforms strong baselines across different evaluation settings.

2 Related Work

2.1 Biomedical Relation Extraction

Biomedical relation extraction has been extensively studied as a core task in biomedical natural language processing, with particular attention to identifying relations between chemicals, diseases, and genes from scientific literature. Early benchmark efforts, such as the BioCreative V CDR (Li et al., 2016) task corpus, established standard evaluation settings for chemical–disease relation extraction and facilitated the development of supervised learning approaches. Subsequent work further expanded the scope of biomedical relation extraction by constructing larger and richer datasets, including BioRED (Luo et al., 2022) and ChemDisGene (Zhang et al., 2022), which cover diverse entity types and multiple relation categories. DrugProt (Miranda-Escalada et al., 2023) introduced a large-scale gold standard for granular drug-gene/protein interactions. Sosa et al. (2023) frame the association of cell types and tissues with protein-protein interactions as a classification task, utilizing syntactic and meta-discourse features to enrich literature-derived knowledge graphs.

In parallel, researchers have explored alternative learning paradigms to address data sparsity and annotation cost in biomedical relation extraction. Xiao et al. (2024) extend document-level relation extraction to a federated learning setting for the first time and propose a novel non–independently and identically distributed scenario based on graph structural entropy. Wang et al. (2024) investigate a pipeline that performs sentence-level relation classification prior to entity extraction to alleviate entity ambiguity, and further incorporate structural constraints between entities and relations to guide the model's hypothesis space.

These studies have significantly advanced the modeling of biomedical relations under various practical constraints. However, despite these efforts, existing work has primarily focused on identifying the presence or type of relations between entities, and does not explicitly model the applicability conditions under which therapeutic drug–disease relations hold.

2.2 Fine-grained Information Extraction Benchmarks

Beyond relation extraction, prior work has explored a variety of fine-grained information extraction tasks that aim to identify condition-like or attribute-level information from textual data. In the clinical domain, adverse event extraction has been studied as a representative task, where models are required to extract specific event spans and assign them to predefined categories (D'Oosterlinck et al., 2023; Sahoo et al., 2024; Guellil et al., 2025). Early approaches typically rely on a sequence tagging framework, such as conditional random fields (CRFs), to model token-level dependencies and capture structured output constraints (Guellil et al., 2025). Srivastava et al. (2025) explore instruction-tuning on large language models (LLMs) for event extraction, leveraging textual annotation guideline to guide model predictions. Their results show that prompt- and instruction-based approaches can serve as effective alternatives to traditional supervised models, particularly in low-resource or cross-schema settings. These findings support the use of prompting-based methods as a reasonable approach for fine-grained information extraction tasks.

3 Drug-ACE Dataset

We create the Drug-ACE dataset that manually annotates conditions under which a certain drug takes a therapeutic effect against a disease. Figure 1 presents an example, showing the input and the annotated applicability condition.

3.1 Annotation Data Preparation

Our dataset is constructed on top of the ChemDisGene dataset (Zhang et al., 2022), which consists of PubMed[1] biomedical abstracts annotated with drug, disease, and gene entities, as well as pairwise relations among them.

The original ChemDisGene dataset covers a diverse set of entity types and relation categories. In this work, we restrict our focus to instances involving *therapeutic* drug–disease relations for their practical values and filter out instances corresponding to other relation types. The original ChemDisGene includes relations identified by *in vivo* and *in vitro* experiments. We manually reviewed and further filtered those that do not mention clinical studies or clinical trials, retaining only clinically grounded drug–disease relations for applicability condition annotation. We further manually inspected the therapeutic relation annotations and removed instances which seem to have incorrect or inconsistent relation annotations.

Note that our preprocessing step does not modify the original textual content of the biomedical literature. Rather, we only restrict instances included in Drug-ACE to be ones with therapeutic relations with sufficiently reliable evidence.

3.2 Condition Types

To better understand applicability conditions and facilitate their systematic modeling, we assign type labels to each extracted condition. There has been no consensus on an exhaustive taxonomy of applicability conditions for drug–disease relations. To design a reasonable set of condition types, we reviewed relevant biomedical literature (Wu et al., 2019; Bhatt et al., 2021; Hanlon et al., 2023) and identified commonly discussed conditions that constrain or qualify therapeutic applicability. We then consult with a domain expert to define the annotation scope, and finally selected six condition types after multiple rounds of discussions.

The following six condition types are frequently observed in biomedical literature and are clinically meaningful which can substantially influence treatment effectiveness and clinical decision-making.

- *Dosage*, indicates the dosage or amount of the drug administered, including specific dosage values and ranges required for effective or safe treatment. E.g.: "**3.3 mg/70 kg of M6G**"
- *Age*, specifies the patient's age or age group, including explicit ages or age-related categories that affects the applicability of the treatment. E.g.: "**children** with nephrotic syndrome"
- *Gene*, specifies genetic characteristics of patients, such as the presence of particular genes, that influence drug response or treatment suitability. E.g.: "**HDL-bound paraoxonase-1 (PON1)**"
- *Gender*, indicates the biological sex or gender of the target patient population when treatment applicability differs across genders. E.g.: "Fourteen **male** patients"
- *Comorbidity*, refers to the presence of one or more additional pre-existing diseases, disor-

[1] https://pubmed.ncbi.nlm.nih.gov/

ders, or risk factors other than the index disease under investigation, which may affect the drug's applicability or therapeutic effect. E.g.: "children with **nephrotic syndrome**"

- *Body Type*, describes general body characteristics or physical conditions of patients, including pregnancy, obesity, underweight status, or other body composition–related factors that may influence treatment outcomes. E.g.: "**primigravida** woman"

3.3 Annotation

The annotation has been conducted from Oct. to Dec., 2025. The primary communication tool was Slack; the annotators could ask questions anytime.

Annotator Selection We recruited two graduate students majoring in life science and technology, and pathological biochemistry, respectively. To ensure annotation quality, these annotators were first asked to conduct trial sessions, during which their domain knowledge and understanding of the annotation guideline were carefully evaluated. The annotators were paid about $1 per instance.

Annotation Guideline We provided annotation guidelines to the annotators, which consists of task definition, inclusion and exclusion criteria, taxonomy for the condition types, and annotation scope. The complete version of the guideline is presented in the Appendix A.

Annotation Procedure We provided 200 abstracts as a batch to the annotators to control the annotation quality. Each abstract was independently annotated by the two annotators. To ensure the consistency and reliability of the annotated applicability conditions, one of the authors reviewed all the results. In cases of disagreement, the third annotator (one of the authors) served as a judge and determined the final annotation through adjudication. The initial agreement between the two annotators was 61%. The revised annotations were feedback to the annotators to improve the task understanding and agreement in the next batch. The agreement rate of the final batch improved to 86%.

3.4 Resultant Dataset

As the final outcome of our annotation, our DurgACE dataset in total consists of 1,119 drug-disease pairs from 667 unique Pubmed abstracts associated with conditions. Namely, 2,290 applicability condition spans were identified, with an average of approximately two annotated spans per drug-disease pair. Figure 2 shows the distribution of condition types. The majority type was Dosage, followed by Age, Gene, and Gender.

Category	#Abstracts	#Pairs	DocLen	#Spans
Train	334	558	266.0	2.01
Dev	110	182	270.3	2.38
Test	223	379	283.4	1.94

Table 1: Dataset statistics for the train, development, and test categories.

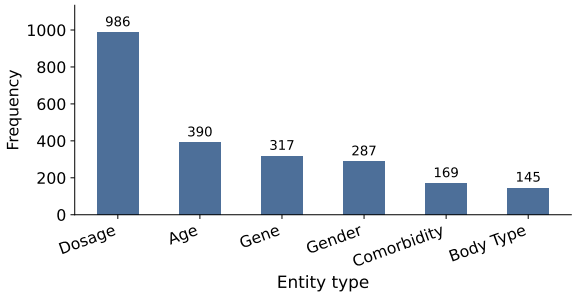

Figure 2: The distribution of applicability conditions.

We split the dataset for training, development, and testing. Table 1 summarizes the dataset statistics for each split: the number of unique Pubmed abstracts (*#Abstracts*), the number of drug-disease pairs (*#Pairs*), the average document length measured in tokens (*DocLen*), and the average number of applicability condition spans per pair (*#Spans*).

4 Applicability Condition Extraction

We propose a method to extract applicability conditions of a given therapeutic drug-disease pair from biomedical paper titles and abstracts. Our method utilizes LLMs' strong capability in understanding long-texts and conduct parameter-efficient tuning with LoRA to predict conditions. To address the complex nature of biomedical literature understanding, we propose to explicitly model *relation roles* between the drug and disease. Following He et al. (2025) who employ tabular structure information for table understanding, we explicitly encode the relation roles in LoRA.

4.1 Task Definition

We first define the task of therapeutic drug–disease applicability condition extraction. Given a pair of drug and disease, the goal is to identify conditions in texts under which the drug takes a therapeutic effect to treat the disease. More specifically, the input and output consist of the following.

Input An input consists of (i) a therapeutic drug–disease pair and (ii) the title and abstract of a biomedical paper where the pair is mentioned. The title and abstract provide broader biomedical context, from which the applicability conditions for the given drug–disease pair can be extracted.

Output The output is a set of applicability conditions of the given drug–disease pair. Each condition specifies a particular circumstance under which the drug is applicable, and is associated with predefined condition types.

4.2 Relation Roles

Drug–disease applicability condition extraction requires identifying conditions distributed across sentences from broader experimental or clinical descriptions, which can be implicitly mentioned. Therefore, it is insufficient to look for possible spans appearing close to the drug-disease pair. Furthermore, multiple drugs and diseases may co-occur, thus multiple conditions co-exist in text. Another hurdle is that biomedical entities tend to be split into subword tokens, which dilutes span-level representations (Balde et al., 2024). These unique characteristics easily confuse a model from understanding "who applies to whom" from context.

To address these challenges, we encode the subject and object of a target condition into LoRA. Given the input consisting of a biomedical research literature $T = \{t_i\}_{i=1}^{|T|}$ (title and abstract) and a queried drug–disease pair (d, s), we assign a relation role to each input token to explicitly encode its participation in the queried relation. Specifically, each token is labeled as one of three roles: OBJ for tokens belonging to the given drug $d = \{d_1, d_2, \ldots, d_n\}$, SUBJ for tokens belonging to the given disease $s = \{s_1, s_2, \ldots, s_m\}$, and NA for all remaining tokens.

Formally, we determine a role sequence $\{y_i\}_{i=1}^{|T|}$:

$$y_i = \begin{cases} \text{OBJ}, & t_i \in d, \\ \text{SUBJ}, & t_i \in s, \\ \text{NA}, & \text{otherwise}. \end{cases} \quad (1)$$

We first locate the character-level spans of d and s in T by exact matching after lemmatization, and then map these spans to token indices using tokenizer offsets. When an abstract contains multiple mentions of the same drug and disease string, we assign the same role to all of their occurrences. This role labelling is model-agnostic and introduces only minimal overhead while providing an explicit subject–object signal for subsequent LoRA tuning.

4.3 Role-Conditioned LoRA

Preliminary: LoRA We briefly review standard LoRA (Hu et al., 2022) before introducing our method. Given a frozen weight matrix $W_0 \in \mathbb{R}^{d\times k}$ of an LLM, LoRA models the parameter update ΔW as a low-rank decomposition.

$$\mathbf{h} = W_0\mathbf{x} + \Delta W\mathbf{x} = W_0\mathbf{x} + BA\mathbf{x}, \quad (2)$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$. By assuming that the parameter update is low rank, i.e., $r \ll \min(d, k)$, LoRA significantly reduces the parameters to tune.

Proposed Method: Role-Conditioned LoRA In the original LoRA, the low-rank update is applied uniformly across all tokens, without distinguishing the relation roles of different spans in the input, which can result in suboptimal performance (Zhao et al., 2023; Liu and Demberg, 2024). We guide the fine-tuning process by injecting the relation role into LoRA, as shown in Figure 3. Specifically:

$$\mathbf{h} = W_0\mathbf{x} + BA\mathbf{x} + B_y\,\mathbf{e}_y, \quad (3)$$

where $y \in \{\text{OBJ}, \text{SUBJ}, \text{NA}\}$ denotes the relation role assigned to the input token, $\mathbf{e}_y \in \mathbb{R}^r$ is a learnable embedding corresponding to the role y, and $B_y \in \mathbb{R}^{d\times r}$ is a role-conditioned low-rank projection matrix. The matrix B_y shares the same shape as the original LoRA matrix B, and is initialized to zero in the same manner.

Note that the role embedding $\mathbf{e}_y$ is unique for each transformer layer. This design allows different layers to capture role information at varying levels of abstraction. This allows greater flexibility and expressiveness compared to sharing the same embedding across layers, while incurring only minimal additional parameters.

Finally, the LLM outputs a list of extracted applicability conditions, with each item consisting of a textual span and an associated type, formatted as `Span: <span> | Label: <type>`. We employed the standard cross-entropy loss for training.

5 Experiment Setup

Our Drug-ACE is the first dataset for drug–disease applicability condition extraction, thus we benchmark strong baseline methods on this dataset to clarify challenges in applicability condition extraction. We also empirically evaluate the effectiveness of the proposed method.

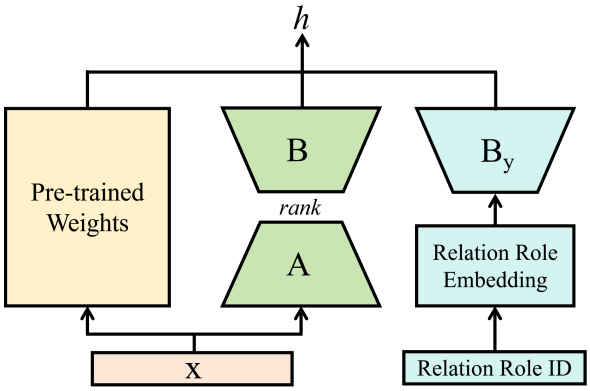

Figure 3: Model architecture: The diagram illustrates the integration of the relation role into the LoRA model. The left side is the standard LoRA, while the right side depicts our method.

5.1 Baselines

We evaluate the following existing methods[2]:

- **SpanMarker**[3] adhering to the Packed Levitated Marker architecture (Ye et al., 2022). Each model leverages a pre-trained language model as the backbone encoder, including RoBERTa (base and large) (Liu et al., 2019), BERT (cased and uncased) (Devlin et al., 2019), BiomedBERT (Gu et al., 2021), BioBERT (Lee et al., 2020), and Bio_ClinicalBERT (Alsentzer et al., 2019). Input sequences are processed using a levitated marker mechanism, in which pairs of trainable marker tokens are inserted into the self-attention layers to aggregate span-specific contextual representations. The resulting span embeddings are then fed into a linear classifier for final applicability condition prediction.

- **Standard LoRA** on widely used LLMs, including Gemma2-9B (Team et al., 2024), Qwen2.5-7B (Yang et al., 2025b), Qwen3-4B (Yang et al., 2025a), Gemma3-4B (Team et al., 2025), and its medical domain-adapted counterpart MedGemma-4B (Sellergren et al., 2025). We apply LoRA to all linear modules in the backbone models. In our main experiments, the LoRA rank is set to 8.

- **2-Shot Prompting** on DeepSeek-R1-70B (DeepSeek-AI et al., 2025), Llama3.3-70B (Grattafiori et al., 2024), and Qwen2.5-72B (Yang et al., 2025b). These models are selected as representative large-scale LLMs that are publicly accessible and have demonstrated strong performance on general reasoning and instruction-following tasks. To construct 2-shot prompts, we randomly sample two drug-disease pairs from the training set. The prompt is presented in the Appendix B

We adopt a linear learning rate schedule with warmup and decay following (Devlin et al., 2019) when training SpanMarker models as well as LoRA models. All experiments were conducted on a single NVIDIA H100 GPU.

5.2 Implementation of Proposed Method

For a fair comparison, our Role-Conditioned LoRA was implemented on the same backbone models as the standard LoRA baselines. We also use the same LoRA rank, setting it to 8 in all main experiments. He et al. (2025) empirically showed that applying task-specific LoRA adaptations to key and value projections is an effective design choice. Following this, we apply our relation-role LoRA to the key and value projection layers, and employ standard LoRA to the remaining linear layers.

5.3 Evaluation Metrics

We evaluated the performance of the models at (a) the span level, where only applicability condition spans are counted, and (b) span and type, where both of the condition span and condition type matter. We employ Hard and Soft matching for evaluating condition spans.

The final evaluation score is the average of F1 scores computed per sample. When an instance has no ground-truth applicability condition, the F1 is defined as 1.0 if the model correctly predicts zero conditions, and 0.0 otherwise.

Hard Matching Under the hard matching criterion, a predicted span is considered correct only if it exactly matches the gold reference.

Soft Matching We adopt the soft matching algorithm proposed by Han et al. (2024) to allow flexibility in evaluation. Soft matching relaxes the strict boundary requirement by considering both span containment and textual similarity between

[2]We also experimented with token-level BERT-BiLSTM-CRF and BERT-CRF baselines (Hochreiter and Schmidhuber, 1997; Lafferty et al., 2001; Akbik et al., 2019). Despite careful tuning under the same training and evaluation protocols, these models ended up predicting no condition span, likely due to insufficient training samples. We therefore omit these baselines from the main comparison.

[3]https://tomaarsen.github.io/SpanMarkerNER/

a predicted span and its gold reference. Specifically, a predicted span is regarded as a soft match if there exists a containment relation between the predicted and the gold reference span and their textual similarity exceeds a predefined threshold. In our experiments, the threshold is set to 0.5.

6 Results and Analysis

6.1 Main Results

We report the average F1 scores with standard deviation across three random seeds for SpanMarker and LoRA fine-tuning, and five seeds for 2-shot prompting. Table 2 reports the main results under both Hard and Soft matching at the span-only and span and condition type predictions across different methods and base LLMs.

The proposed method achieved the best soft and hard F1 scores on Qwen3-4B and Gemma3-4B, respectively. Table 3 presents the statistical significance analysis of RCLoRA across five different LLM backbones. Notably, the proposed method significantly outperforms the standard LoRA fine-tuning ($p < 0.05$) consistently across all primary metrics.

SpanMarker. Among SpanMarker baselines, models initialized with domain-specific pretraining generally achieve stronger performance. In particular, BiomedBERT-base performs strongly despite its smaller model size for both span-only and span and type predicions. This confirms the effectiveness of BiomedBERT's extensive pretraining on abstracts from PubMed.[4]

LLMs with LoRA fine-tuning. Different from SpanMarker models, MedGemma-4B was inferior to its general-domain counterpart, Gemma3-4B. This suggests that domain adaptation alone is not always sufficient for LLMs.

Few-shot prompting. Despite significantly larger model sizes, 2-shot prompting exhibits substantially lower performance compared with fine-tuned models, demonstrating the difficulty of the task and the necessity of supervised adaptation. These results further indicate that prompt-based inference alone is insufficient to conduct complex reasoning required by applicability condition extraction.

[4]Note that the PubMed abstracts used in Drug-ACE are newer than those for BiomedBERT, thus there should be no data leakage.

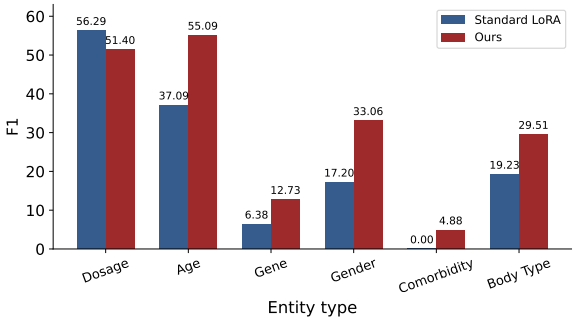

Figure 4: Performance comparison between standard LoRA and our method across different condition types

6.2 Performance per Condition Types

Figure 4 illustrates the performance comparison between standard LoRA and our method across different condition types on Gemma2-9B. Our method outperforms the standard LoRA on all entity types: Age, Body Type, Comorbidity, Gender, and Gene, except Dosage. While the standard LoRA performs well on the most common type, i.e., Dosage, it struggles to identify other types. In particular, it yielded near-zero performance on Comorbidity-related conditions, whereas our method is able to capture meaningful signals for this type. We conjecture that condition types such as Gene and Comorbidity are challenging not only for their sparsity but also the highly specialized natures. Improvements on such types constitutes our future work.

6.3 Ablation Study

The previous section demonstrated the effectiveness of encoding relation roles into LoRA. To further understand the effectiveness of our design, we compare to the following methods.

Marker: For each drug-disease pair, we mark all occurrences of the corresponding entities in the input text using special tokens: `[Drug] entity [/Drug]` and `[Disease] entity [/Disease]`. We couple this marked input with a standard LoRA setup, keeping the model architecture untouched to isolate the effect of input-level markers.

Role-Specific Vectors: We directly add role-specific vectors at each layer instead of using low-rank matrices:

$$W_0\mathbf{x} + BA\mathbf{x} + \mathbf{e}_y. \tag{4}$$

Single-B Matrix: We replace the two-matrix formulation of Eq. (3) with a single matrix, sharing

Model	Trainable Params	Span		Span and Type	
		Hard	Soft	Hard	Soft
SpanMarker					
RoBERTa-base	124.67 M	34.86 ± 2.12	48.00 ± 1.14	34.81 ± 2.19	47.45 ± 1.45
RoBERTa-large	355.39 M	37.81 ± 2.11	52.16 ± 1.50	37.81 ± 2.11	51.90 ± 1.51
BERT-base-cased	108.33 M	30.73 ± 2.19	45.39 ± 3.15	30.69 ± 2.23	44.93 ± 3.19
BERT-base-uncased	109.51 M	33.07 ± 0.52	45.20 ± 1.64	32.96 ± 0.32	44.80 ± 1.70
BiomedBERT-base	108.26 M	39.35 ± 1.06	53.23 ± 1.68	39.35 ± 1.06	52.91 ± 1.82
BioBERT-base-cased	108.33 M	33.57 ± 1.25	45.49 ± 0.78	33.38 ± 1.56	45.08 ± 0.61
Bio_ClinicalBERT	108.33 M	32.96 ± 2.97	46.40 ± 1.90	32.94 ± 3.01	46.03 ± 2.01
LLMs with LoRA Fine-tuning					
Gemma2-9B$_{\text{LoRA}}$	27.01 M	48.03 ± 2.77	59.09 ± 2.38	47.60 ± 2.68	58.30 ± 1.99
Gemma2-9B$_{\text{RCLoRA}}$	28.39 M	49.89 ± 3.39	59.57 ± 3.01	49.57 ± 3.46	58.83 ± 3.02
Qwen2.5-7B$_{\text{LoRA}}$	20.19 M	46.53 ± 2.43	56.12 ± 1.75	46.42 ± 2.47	55.51 ± 1.69
Qwen2.5-7B$_{\text{RCLoRA}}$	20.42 M	46.82 ± 1.53	56.99 ± 0.65	46.52 ± 1.61	56.23 ± 0.56
Qwen3-4B$_{\text{LoRA}}$	16.52 M	49.37 ± 2.99	59.84 ± 3.12	49.19 ± 3.11	59.18 ± 3.30
Qwen3-4B$_{\text{RCLoRA}}$	17.11 M	51.18 ± 0.46	$\mathbf{60.62 \pm 0.61}$	50.91 ± 0.45	$\mathbf{59.78 \pm 0.56}$
Gemma3-4B$_{\text{LoRA}}$	14.90 M	49.17 ± 1.75	56.53 ± 0.99	49.01 ± 1.89	56.10 ± 1.19
Gemma3-4B$_{\text{RCLoRA}}$	15.46 M	$\mathbf{51.43 \pm 4.23}$	59.74 ± 3.74	$\mathbf{51.20 \pm 4.17}$	59.11 ± 3.59
MedGemma-4B$_{\text{LoRA}}$	14.90 M	46.62 ± 1.12	55.35 ± 1.02	46.39 ± 1.22	54.51 ± 0.90
MedGemma-4B$_{\text{RCLoRA}}$	15.46 M	48.63 ± 3.01	57.40 ± 2.60	48.40 ± 3.11	56.56 ± 2.63
LLMs with 2-Shot Prompting					
DeepSeek-R1-70B	-	13.91 ± 1.83	35.31 ± 2.42	13.73 ± 1.80	35.00 ± 2.46
Llama3.3-70B	-	23.75 ± 2.70	32.37 ± 0.73	23.71 ± 2.75	32.12 ± 0.77
Qwen2.5-72B	-	23.04 ± 1.44	31.35 ± 0.61	22.82 ± 1.37	30.92 ± 0.63

Table 2: Performance comparison under Hard and Soft matching of span-only and span and type prediction: the best scores are indicated by **bold** fonts, and the underlines indicate that the proposed method (**RCLoRA**) outperforms corresponding standard LoRA counterparts.

Method	Span		Span & Type	
	Hard	Soft	Hard	Soft
LoRA (Avg.)	47.94	57.39	47.72	56.72
RCLoRA (Avg.)	49.59	58.86	49.32	58.10
Δ	**+1.65**	**+1.47**	**+1.60**	**+1.38**
p-value	0.013	0.018	0.015	0.022

Table 3: Statistical significance test results of RCLoRA across 5 backbones on LLMs with LoRA fine-tuning (3 seeds x 5 models = 30 independent runs). Note: p-values are calculated using a paired t-test across all 15 experimental pairs.

Method	Span		Span & Type	
	Hard	Soft	Hard	Soft
Standard LoRA	49.37	59.84	49.19	59.18
Marker	49.26	58.37	48.99	57.53
Role-Specific Vectors	43.93	52.13	43.72	51.58
Single B-Matrix	49.10	58.57	48.59	57.64
Random Role	47.43	56.85	47.16	56.27
RCLoRA (Proposed)	**51.18**	**60.62**	**50.91**	**59.78**

Table 4: Ablation study results on Qwen3-4B

the same B matrix with LoRA:

$$W_0\mathbf{x} + BA\mathbf{x} + B\mathbf{e}_y. \quad (5)$$

This setting removes the separation between the transformation applied to the input vector and that applied to the relation role embedding.

Random Roles: We input random ids of 0, 1, and 2 to the relation embeddings, to distinguish the effect of using roles from simple increase in parameter sizes.

Results We employed Qwen3-4B as the backbone LLM for the ablation study for computational efficiency. For each setting, scores were averaged over three runs with different random seeds. Table 4 presents the results. Compared with the standard LoRA, all variants lead to noticeable performance drops. In particular, the lower performance of the Marker setting suggests that explicitly highlighting drug and disease entities might inevitably cause the LLM to over-focus on the marked entities. This potentially leads the model to overlook critical contextual expressions in the biomedical context, which are significant for identifying the applicability conditions. Furthermore, marking every occurrence of these entities throughout the text may introduce redundant signals that confuse LLMs' reasoning. These results indicate that relation role requires distinctive LoRA matrix, and relation roles are crucial for the performance gain of the proposed method.

6.4 Effects of LoRA Ranks

Figure 5 shows the effects of LoRA ranks when using Gemma2-9B as the backbone LLM. For each rank, Soft F1 scores of span and condition type pre-

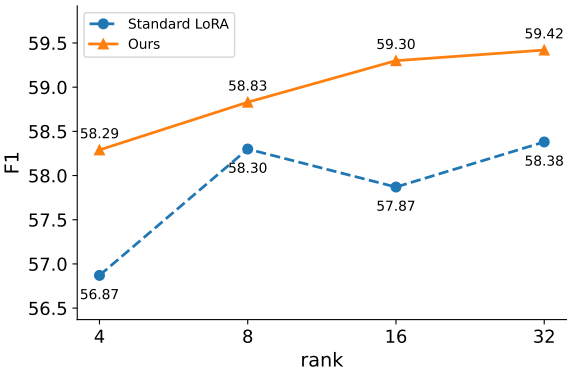

Figure 5: Effects of LoRA ranks on the Gemma2-9B

diction were averaged over three runs with different random seeds. Across different rank settings, our method consistently outperforms standard LoRA. Furthermore, the performance of our method enhance with larger rank, which aligns with the findings of Zhang et al. (2023).

Table 5 reports the corresponding number of trainable parameters under different rank settings for both standard LoRA and our method. While our method slightly increases the number of trainable parameters at the same rank, the gaps are modest and scale consistently with the rank.

6.5 Effects of Soft Matching Threshold

To investigate the impact of the similarity threshold used in soft matching, we conducted a sensitivity analysis comparing our primary model, Gemma3-4B-RCLoRA, against the standard Gemma3-4B-LoRA baseline on soft F1 scores of span and condition type. As shown in Table 6, while performance predictably decreases as the matching criterion becomes more stringent, RCLoRA consistently outperforms standard LoRA across all thresholds.[5] We selected 0.5 because hard matching (1.0) is overly restrictive for complex clinical spans. For example, if the target is "titrated to 50 mg once daily" and the model predicts "50 mg once daily" , exact matching would treat this as a complete failure. This would unfairly discard the model's success in capturing the core dosage information. A threshold of 0.5 provides a reasonable trade-off that recognizes such clinically useful partial extractions.

[5]Increasing the similarity threshold makes the matching criterion more stringent, which reduces the number of True Positives. Given that the total counts of predictions and ground-truth are fixed, both Precision and Recall decrease monotonically.

Method	Trainable Parameters			
	r=4	r=8	r=16	r=32
Standard LoRA	13.5 M	27.0 M	54.0 M	108.0 M
Ours	14.2 M	28.4 M	56.8 M	113.5 M

Table 5: Comparison of trainable parameters under different LoRA ranks on Gemma-9B

Method	Similarity Threshold				
	0.1	0.3	0.5	0.7	0.9
LoRA	64.27	61.67	57.45	55.07	51.39
RCLoRA	**68.88**	**67.26**	**63.16**	**60.49**	**56.26**

Table 6: Performance comparison between LoRA and RCLoRA across different similarity thresholds

7 Conclusion

In this study, we introduced a novel task of drug–disease applicability condition extraction. We created the annotation dataset, named Drug-ACE, and proposed a method for applicability condition extraction that enhances LoRA by explicitly encoding relation roles between drugs and diseases. Our benchmarking results of conventional biomedical relation extraction methods on Drug-ACE highlight the challenges posed by applicability condition extraction and demonstrate the effectiveness of our method.

Our future work includes expanding the scale of the dataset, which may further improve model robustness. Improvement on challenging condition types, such as Gene and Comorbidity is also crucial. In addition, while this study focuses on therapeutic drug–disease relations, it should be worthy to extend to cover other biomedical relations, such as gene–disease interactions.

Limitations

Despite the contributions of this study, several limitations remain. First, the overall size of the proposed dataset is modest. Although the dataset is carefully annotated, scaling up the data size could further improve model robustness and enable more comprehensive empirical analysis.

In addition, this study focuses exclusively on therapeutic drug–disease relations. While this scope allows for a focused investigation of applicability conditions, other biologically meaningful relations, such as gene–disease interactions, also exhibit conditional applicability and warrant further exploration. Extending the task and dataset to cover a broader range of biomedical relations is an important direction for future research.

Potential Risks Our Drug-ACE dataset is intended for research purposes only. The annotated conditions reflect text-level reporting in biomedical research, which may include exploratory or experimental findings. Therefore, Drug-ACE should be viewed as an auxiliary tool for literature synthesis rather than a source of clinically validated medical truth. Inappropriate use without expert validation may lead to misconduct in clinical environments.

AI Assistant Use We used AI assistants for improving writing; they were used exclusively for enhancing readability and correcting grammar. They did not contribute to the scientific content of the manuscript.

Acknowledgment

We thank the domain expert, Narumi Tokunaga, and our annotators, Leonardo Ken Okumura and Miko Oikawa, for their significant contributions in Drug-ACE creation. This work was supported by Cross-ministerial Strategic Innovation Promotion Program (SIP) on "Integrated Health Care System" Grant Number JPJ012425.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilseyar Alimova and Elena Tutubalina. 2019. Detecting adverse drug reactions from biomedical texts with neural networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 415–421, Florence, Italy. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Gunjan Balde, Soumyadeep Roy, Mainack Mondal, and Niloy Ganguly. 2024. Medvoc: Vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. In *IJCAI*, pages 6180–6188.

Arjun Bhatt, Ruth Roberts, XI Chen, Ting Li, Skylar Connor, Qais Hatim, Mike Mikailov, Weida Tong, and Zhichao Liu. 2021. Dice: A drug indication classification and encyclopedia for ai-based indication extraction. *Frontiers in Artificial Intelligence*, 4:711467.

Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L Hamilton. 2022. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, 23(6):bbac404.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Karel D'Oosterlinck, François Remy, Johannes Deleu, Thomas Demeester, Chris Develder, Klim Zaporojets, Aneiss Ghodsi, Simon Ellershaw, Jack Collins, and Christopher Potts. 2023. BioDEX: Large-scale biomedical adverse drug event extraction for real-world pharmacovigilance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13425–13454, Singapore. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Imane Guellil, Salomé Andres, Atul Anand, Bruce Guthrie, Huayu Zhang, Abul Hasan, Honghan Wu, and Beatrice Alex. 2025. Adverse event extraction from discharge summaries: A new dataset, annotation scheme, and initial findings. In *Proceedings of the 63rd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 28532–28562, Vienna, Austria. Association for Computational Linguistics.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. 2024. An empirical study on information extraction using large language models. *Preprint*, arXiv:2305.14450.

Peter Hanlon, Elaine W Butterly, Anoop SV Shah, Laurie J Hannigan, Jim Lewsey, Frances S Mair, David M Kent, Bruce Guthrie, Sarah H Wild, Nicky J Welton, and 1 others. 2023. Treatment effect modification due to comorbidity: Individual participant data meta-analyses of 120 randomised controlled trials. *Plos medicine*, 20(6):e1004176.

Xinyi He, Yihao Liu, Mengyu Zhou, Yeye He, Haoyu Dong, Shi Han, Zejian Yuan, and Dongmei Zhang. 2025. TableLoRA: Low-rank adaptation on table structure understanding for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22376–22391, Vienna, Austria. Association for Computational Linguistics.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yifan Jin, Jiangmeng Li, Zheng Lian, Chengbo Jiao, and Xiaohui Hu. 2022. Supporting medical relation extraction via causality-pruned semantic dependency forest. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2450–2460, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, volume 1, pages 282–289.

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

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

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

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

Zhiyong Lu. 2011. Pubmed and beyond: a survey of web tools for searching biomedical literature. *Database*, 2011:baq036.

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

Antonio Miranda-Escalada, Farrokh Mehryary, Jouni Luoma, Darryl Estrada-Zavala, Luis Gasco, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. 2023. Overview of drugprot task at biocreative vii: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database*, 2023:baad080.

Dat Quoc Nguyen and Karin Verspoor. 2018. Convolutional neural networks for chemical-disease relation extraction are improved with character-based word embeddings. In *Proceedings of the BioNLP 2018 workshop*, pages 129–136, Melbourne, Australia. Association for Computational Linguistics.

Arpita Roy and Shimei Pan. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5357–5366, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranab Sahoo, Ayush Singh, Sriparna Saha, Aman Chadha, and Samrat Mondal. 2024. Enhancing adverse drug event detection with multimodal dataset: Corpus creation and model development. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11214–11226, Bangkok, Thailand. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Daniel N. Sosa, Rogier Hintzen, Betty Xiong, Alex de Giorgio, Julien Fauqueur, Mark Davies, Jake Lever, and Russ B. Altman. 2023. Associating biological context with protein-protein interactions through text mining at pubmed scale. *Journal of Biomedical Informatics*, 145:104474.

Saurabh Srivastava, Sweta Pati, and Ziyu Yao. 2025. Instruction-tuning LLMs for event extraction with annotation guidelines. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13055–13071, Vienna, Austria. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the biocreative v chemical-disease relation (cdr) task. *Database*, 2016:baw032.

Richard M Weinshilboum and Liewei Wang. 2017. Pharmacogenomics: precision medicine and drug response. In *Mayo Clinic Proceedings*, volume 92, pages 1711–1722. Elsevier.

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

Yan Xiao, Yaochu Jin, and Kuangrong Hao. 2024. Federated document-level biomedical relation extraction with localized context contrast. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7163–7173, Torino, Italia. ELRA and ICCL.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

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

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing hierarchical guidance into prompt tuning: A parameter-efficient framework for multi-level implicit discourse relation recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6477–6492, Toronto, Canada. Association for Computational Linguistics.

A Annotation Guideline

A.1 Introduction

This task is an applicability condition extraction for therapeutic drug-disease relations. The goal of the task is to extract applicability conditions (i.e., under what conditions a certain drug can treat a certain disease) for the mentioned drug from the given biomedical research literature.

A.2 Task Definition

Given a biomedical research literature (including a title and abstract) and a therapeutic drug–disease relation, annotators are required to identify the applicability condition(s) mentioned in the text that are relevant to the given drug–disease pair and assign a predefined label. Each annotation instance contains only one drug–disease pair. However, a single biomedical document may include multiple therapeutic drug–disease relations. In such cases, please perform the annotation separately for each specific drug–disease pair provided.

A.3 Inclusion and Exclusion Criteria

The corpus may contain biomedical literature related to cell or animal experiments rather than human clinical studies. Annotators are required to verify whether the given biomedical literature pertains to a human clinical study. If the given literature is not about a human clinical study, the entire literature must be skipped. Additionally, the provided drug–disease relation may not be therapeutic. If the relation is determined to be non-therapeutic, the specific annotation instance should be skipped.

A.4 Condition Types

We assign condition type labels to each extracted drug-disease condition. The condition types are as follows:

- *Dosage*, indicates the dosage or amount of the drug administered, including specific dosage values, ranges that are required for effective or safe treatment.
- *Age*, specifies the patient age or age group, including explicit ages or age-related categories.
- *Gene*, specifies genetic characteristics of patients, such as the presence of particular genes, that influence drug response or treatment suitability.
- *Gender*, indicates the biological sex or gender of the target patient population, such as male or female, when treatment applicability differs across genders.
- *Comorbidity*, refers to the presence of one or more additional pre-existing diseases, disorders, or risk factors other than the index disease under investigation, which may affect the drug's applicability or therapeutic effect.
- *Body Type* describes general body characteristics or physical conditions of patients, including pregnancy, obesity, underweight status, or other body composition–related factors that may influence treatment outcomes.

A.5 Annotation Scope

- An applicable condition is defined as any contextual factor (e.g., patient attribute) that is explicitly stated or strongly implied to influence the effectiveness or safety of the drug–disease relation.

  (e.g., "..... Long-term oral warfarin is recommended in pediatric Kawasaki disease patients with large coronary artery aneurysms . ..... **Drug–Disease Pair**: warfarin-Kawasaki disease" **Annotated Conditions**: (1) *"pediatric"* → Age (2) *"large coronary artery aneurysms"* → Comorbidity).

- Not every given biomedical research literature contains applicability conditions for the given therapeutic drug – disease relation.

  (e.g., "..... Moreover, lisinopril and nifedipine appear to be capable of reducing bcl-2 concentrations, with potentially beneficial effects on vascular modifications in patients with hypertension. ..... **Drug–Disease Pair**: nifedipine - hypertension" **Annotated Conditions**: No applicability condition.

- When the applicability conditions present, annotators are instructed to extract them even if they are mentioned as general factors without specific values.

  (e.g., "..... Patient height is the main factor determining warfarin dosage, while genotype effects on warfarin dosage vary among studies. ..... **Drug–Disease Pair**: warfarin-Kawasaki disease" **Annotated Conditions**: *"height"* → Body Type).

- Applicability conditions described quantitatively, such as demographic characteristics presented with percentages, are also considered within the scope of valid extractions.

  (e.g., "..... PATIENTS: These patients were mostly men (57%) older than 30 years (56%) with pulmonary obstruction, ... **Drug–Disease Pair**: zafirlukast - asthma" **Annotated Conditions**: (1) *"men (57%)"* → Gender (2) *"older than 30 years (56%)"* → Age).

B Prompt for 2-shot Prompting

The example B.1 shows the prompt template we use for extracting applicability conditions in 2-shot prompting.

Example B.1 *You are a skilled biomedical text annotator. Given the title, abstract, and a therapeutic drug–disease relation, extract all applicability condition spans mentioned in the text that mention under what conditions the drug is used to treat the disease.*

Applicability conditions include:

- Dosage: indicates the dosage or amount of the drug administered, including specific dosage values, ranges that are required for effective or safe treatment.

- Age: specifies the patient age or age group, including explicit ages or age-related categories.

- Gender: indicates the biological sex or gender of the target patient population,

- Comorbidity: refers to the presence of one or more additional pre-existing diseases, disorders, or risk factors other than the index disease under investigation, which may affect the drug's applicability or therapeutic effect.

- Body type: describes general body characteristics or physical conditions of patients, including pregnancy, obesity, underweight status, or other body composition–related factors that may influence treatment outcomes.

- Gene: specifies genetic characteristics of patients, such as the presence of particular genes, that influence drug response or treatment suitability.

For each identified span, you need to specify the corresponding label, please return it in the format: "['Span: | Label: <type>', 'Span: | Label: <type>', ...]"

Please strictly follow the format I gave you. Only include spans that are explicitly mentioned in the context. Do not infer conditions beyond what is supported by the text.

If no applicability condition is mentioned, return an empty list: []

An example:

TITLE: {Example 1 title}

ABSTRACT: {Example 1 abstract}

DRUG - DISEASE RELATION: {Example 1 drug} – {Example 1 disease}

APPLICABILITY CONDITIONS:

{List of span – type pairs}

Another example:

TITLE: {Example 2 title}

ABSTRACT: {Example 2 abstract}

DRUG - DISEASE RELATION: {Example 2 drug} – {Example 2 disease}

APPLICABILITY CONDITIONS:

{List of span – type pairs}

Now, your turn:

TITLE: {Input title}

ABSTRACT: {Input abstract}

DRUG - DISEASE RELATION: {Input drug} – {Input disease}

APPLICABILITY CONDITIONS: