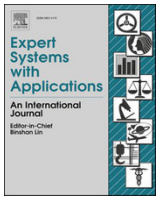
Expert Systems With Applications 273 (2025) 126953



Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



CoTEL-D3X: A chain-of-thought enhanced large language model for drug–drug interaction triplet extraction

Haotian Hu [a,c,e,1], Alex Jie Yang [b,c,2], Sanhong Deng [b,c,3], Dongbo Wang [d,4], Min Song [e,5,*]

[a] Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China
[b] School of Information Management, Nanjing University, Nanjing 210023, China
[c] International Joint Informatics Laboratory, Nanjing University, Nanjing 210023, China
[d] School of Information Management, Nanjing Agricultural University, Nanjing 210095, China
[e] Department of Library and Information Science, Yonsei University, Seoul 03722, Republic of Korea

ARTICLE INFO

Keywords:
Instruction tuning
Large language model
Chain-of-thought
Drug–drug interaction
Triplet extraction
Biomedical information extraction

ABSTRACT

Current state-of-the-art drug–drug interaction (DDI) triplet extraction methods not only fails to exhaustively capture potential overlapping entity relations but also grapples to extract discontinuous drug entities, leading to suboptimal performance in DDI triplet extraction. To address these challenges, we proposed a Chain-of-Thought Enhanced Large Language Model for DDI Triplet Extraction (CoTEL-D3X). Based on the transformer architecture, we designed joint and pipeline methods that can perform end-to-end DDI triplet extraction in a generative manner. Our proposed approach builds upon the novel LLaMA series model as the foundation model and incorporates instruction tuning and Chain-of-Thought techniques to enhance the model's understanding of task requirements and reasoning capabilities. We validated the effectiveness of our methods on the widely-used DDI dataset, which comprises 1025 documents containing 17,805 entity mentions and 4,999 DDIs. Our joint and pipeline methods not only outperformed mainstream generative models, such as ChatGPT, GPT-3, and OPT, on the DDI Extraction 2013 dataset but also improved the current corresponding best F1-score by 9.75% and 5.86%, respectively. Particularly, compared to the currently most advanced few-shot learning methods, our approach achieved more than a two-fold improvement in F1-score. We further validated the method's transferability and generalization performance on the TAC 2018 DDI Extraction and ADR Extraction datasets, and assessed its applicability on real-world data from DrugBank. Performance analysis of the proposed method revealed that the CoT component significantly enhanced the extraction effect. The introduction of generative LLMs allows us to freely define the content and format of inputs and outputs, offering superior usability and flexibility compared to traditional extraction methods based on sequence labeling. Furthermore, as our proposed approach does not rely on external knowledge or manually defined rules, it may lack domain-specific knowledge to some extent. However, it can easily be adapted to other domains.

1. Introduction

When patients take two or more medications concurrently, drug–-drug interaction (DDI) may occur. Such interaction may affect the efficacy of drug treatment, leading to adverse reactions and side effects, and even pose life-threatening risks to patients with severe conditions (Zhang et al., 2020). Although DDI databases, such as DrugBank, currently exist, medical experts and staff find it difficult to manually

* Corresponding author.
E-mail addresses: 20240064@jaas.ac.cn (H. Hu), alexjieyang@smail.nju.edu.cn (A.J. Yang), sanhong@nju.edu.cn (S. Deng), db.wang@njau.edu.cn (D. Wang), min.song@yonsei.ac.kr (M. Song).
[1] ORCID: 0000-0002-4112-8719.
[2] ORCID: 0000-0002-1276-0483.
[3] ORCID: 0000-0002-6910-3935.
[4] ORCID: 0000-0002-9894-9550.
[5] ORCID: 0000-0003-3255-1600.

https://doi.org/10.1016/j.eswa.2025.126953
Received 21 October 2024; Received in revised form 1 February 2025; Accepted 16 February 2025
Available online 17 February 2025
0957-4174/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

update the database owing to the rapid expansion of biomedical literature (Fatehifar & Karshenas, 2021). Therefore, extracting DDI data from vast tracts of literature in an automated way is of great significance for both pharmaceutical research (Asada et al., 2023) and clinical treatment.

Fig. 1 illustrates several typical instances of DDI triplets in biomedical texts. Normally, each entity mentioned in a sentence participates in only one DDI. Overlapping relations occur when a single entity is involved in multiple DDIs. In the provided example, the drug entity "rifampin" interacts with both "corticosteroids" and "corticosteroid," with the types of interaction being "mechanism" and "advice," respectively. A discontinuous entity refers to a case where one entity in the text is split into two or more disjointed fragments. For instance, in the phrase "antiplatelet and anticoagulant medication," the discontinuous entity "antiplatelet medication" exists. Occasionally, these unique cases may co-occur in the text, further increasing the complexity of DDI triplet extraction.

Currently, two main approaches to DDI triplet extraction exist: the pipeline and joint approaches. The pipeline method divides DDI triplet extraction into two independent subtasks: entity recognition and relation classification. This approach bears several advantages. It breaks down the complex task of information extraction into two relatively simpler subtasks, aligning with the human cognitive process. By fine-tuning models for these two stages separately, it can enhance performance in a sequenced manner. However, because the triplet extraction process is artificially divided, it renders it challenging for the second-stage model to effectively utilize entity information and potentially causes error propagation. Consequently, the pipeline method often relies on external information and complex relation extraction rules to mitigate the performance degradation caused by this drawback. This, in turn, reduces the model's transferability when dealing with data from other domains. Moreover, in order to leverage the entity information from the first step, current pipeline methods typically insert special identifiers to locate entity positions or spans. However, these manually defined symbols can, to some extent, disrupt semantic coherence, thereby affecting the accuracy of triplet extraction.

In contrast, the joint method achieves simultaneous extraction of entity and relation by training a single model. This method not only avoids the error propagation issue encountered in the pipeline approach but also leverages entity and interaction information concurrently. Consequently, this method has become the predominant design concept in current DDI triplet extraction research. However, due to limitations in model architecture, most current joint extraction methods only achieve simultaneous modeling of entity and interactions in the encoder or hidden layers of the model. They still employ different decoders to separately predict entity boundaries and relation types during decoding. Although Luo et al. (2020) achieved simultaneous decoding of entities and relations using a set of jointly defined labels through sequence labeling, the complex combination of labels introduced hundreds of decoding types, significantly increasing the computational complexity of the model. Nonetheless, relation matching based on the nearest-neighbor principle makes it susceptible to the oversight of certain entity pairs, especially DDI triplets with a high degree of overlapping entity relations. Besides, the widespread use of transformer encoder architecture (e.g., BERT, BioBERT) and classical neural networks (e.g., RNN, CNN) in triplet extraction is limited by factors such as the number of parameters, pre-training methods, and model architecture (Dong et al., 2021). These limitations render it difficult for joint extraction methods to simultaneously encode and decode DDI triplets.

Furthermore, current methods often rely on enumeration or rule-based matching approaches when dealing with overlapping entity relations. This can lead to the potential omission of triplets. Moreover, current methods struggle to effectively handle a scenario typically encountered during DDI triplet extraction: the extraction of discontinuous entities. For instance, discontinuous entities were detected 38 and nine times in the training and test sets of the DDI Extraction 2013 dataset (Herrero-Zazo et al., 2013), constituting 129 and 28 potential DDI relations, respectively. Although not substantial in quantity, the omission of these relations could lead to the oversight of certain crucial DDIs in the biomedical text, such omissions could potentially cause harm to patients in real-world clinical drug usage. The only known study to address this issue is that by Suárez-Paniagua et al. (2019), which employed an extended BMEWO-V sequence-labeling tag set to label entities using the "V" tag to represent overlapping and discontinuous entities. However, this approach can still lead to erroneous matching during triplet decoding. Therefore, current DDI triplet extraction methods based on sequence labeling, whether employing a pipeline or joint architecture, still struggle to effectively handle the complexities of biomedical information extraction tasks.

To address the aforementioned challenges, we adopted a generative approach toward DDI triplet extraction. Additionally, by drawing inspiration from the step-by-step reasoning process of humans to solve complex problems, we introduced a **C**hain-**o**f-**T**hought **E**nhanced **L**arge Language Model for **D**DI **Triplet** e**X**traction (CoTEL-D3X). The main

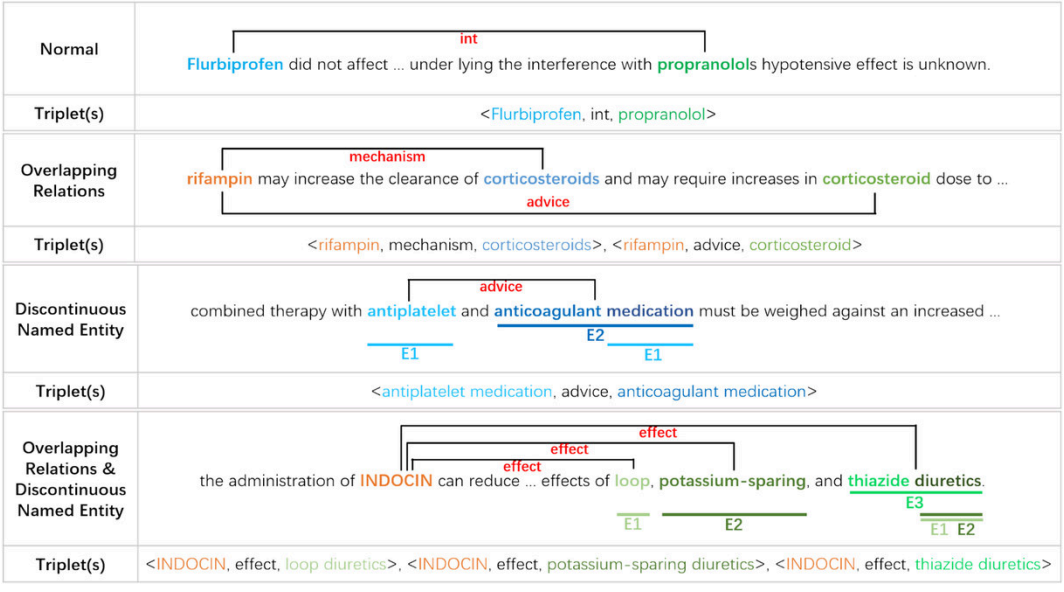


Fig. 1. Examples of DDI Triplets in the DDI Extraction Dataset.

contributions of our study are as follows:

(1) We propose a novel end-to-end generative triplet extraction method that resolves the common challenges of overlapping entity relations and discontinuous entity recognition, which current methods struggle to handle. This approach does not rely on any external domain knowledge, complex features, or manually defined rules, thus offering considerable simplicity and high transferability.

(2) We designed both joint and pipeline methods based on generative large language models (LLMs). The former enables simultaneous modeling of entities and relations throughout both the encoding and decoding processes, while the latter considers all potential entity pairs and their associated interactions. Both methods allow for the flexible definition of output formats, directly generating triplets in the specified format.

(3) We developed a Chain-of-Thought (CoT) enhanced Instruction Tuning method by reconstructing the instruction dataset. This approach decomposes the complex triplet extraction task into a coherent thought process, injecting reasoning abilities into LLMs either implicitly or explicitly. It demonstrates outstanding performance in both controlled environments and real-world engineering scenarios.

In comparative experiments against current advanced LLMs and existing state-of-the-art (SOTA) methods for DDI triplet extraction, both of our proposed methods achieved SOTA results. Specifically, compared with the current leading joint and pipeline methods, our methods displayed an enhancement in F1 scores by 9.75 % and 5.86 %, respectively. To further validate our methods' performance, we conducted experiments on two additional biomedical information extraction datasets: TAC 2018 DDI Extraction and Adverse Drug Reaction (ADR) Extraction. The results demonstrated that our model achieved SOTA performance on these datasets as well. Finally, a case study conducted on real-world data from DrugBank illustrated that the CoTEL-D3X method is capable of clinical application.

2. Research objectives

The primary research objectives and issues to be addressed are as follows:

(1) Can general generative LLMs be utilized for extracting DDI triples in specialized domains? To what extent can the effect be achieved?
(2) Do Instruction Tuning and the Chain-of-Thought (CoT) technique contribute to enhancing the understanding of task requirements and boosting reasoning capabilities in LLMs?
(3) Can the proposed method resolve persistent challenges in DDI triple extraction, such as the recognition of overlapping relations and discontinuous entities?
(4) Compared to existing mainstream joint and pipeline methods, what advantages does our proposed approach offer in terms of performance, availability, and ease of use?

3. Related work

In this section, we provide a detailed report on the current main biomedical entity and relation extraction methods from the perspectives of pipeline and joint approaches, with a particular focus on DDI triplet extraction. We thoroughly analyze the characteristics of each method, and ultimately propose our improvements to address the shortcomings of existing approaches.

DDI triplet extraction refers to the simultaneous recognition of drug entities and their interactions. Currently, most DDI triplet recognition research is based on Bidirectional Encoder Representations from Transformers (BERT)-like models. A common approach is the pipeline method (Zaikis & Vlahavas, 2021), where drug entities are initially identified via sequence labeling and subsequently used as known information for relation classification. While this step-by-step approach is flexible, it ignores inter-task correlation owing to the splitting of entity recognition and relation classification into two independent subtasks (Su et al., 2022).

Suárez-Paniagua et al. (2019) introduced the first pipeline-based DDI extraction method. To enable Named Entity Recognition, they integrated character, word, and sense embeddings to improve the understanding of polysemous words. They used a Bidirectional Long-Short-Term Memory and Conditional Random Field (Bi-LSTM-CRF) model to identify entities and their types in a sequence-labeling manner. To facilitate relation extraction, they initially generated instances for all possible entity pairs in a sentence and subsequently used a convolutional neural network (CNN) to predict relation types. This method, owing to its simple combination of two subtasks and the limited text representation capacity of the embeddings used, resulted in overall suboptimal recognition performance. Zaikis and Vlahavas (2021) proposed an end-to-end two-stage extraction method that employed the BioBERT model, which possesses stronger text representation capabilities, to encode input text. After entity recognition, they applied a set of predefined entity filtering rules to correct mislabeled instances, reducing the possibility of error propagation. For relation extraction, they wrapped entities with special tokens to explicitly indicate their positions in a sentence. This method significantly improved DDI triplet extraction performance; nevertheless, its entity correction rules assumed the preceding tag to be correct, potentially limiting its effectiveness. Yang et al. (2023a) used a span-based approach to recognize drug entities and relations. Unlike that employed by Zaikis and Vlahavas (2021), their method not only introduced entity position information to the input but also included entity category information during relation recognition. Experimental results revealed that this approach effectively enhanced triplet recognition performance. However, it did not address the error propagation issue and was constrained by span length and computational efficiency, hindering it from recognizing entities longer than the maximum span length. Machado et al. (2023) developed an automated system for extracting biomedical triplets, it is capable of identifying drug entities from unstructured biomedical texts and determining the presence of interaction relations. They utilized BioBERT for entity extraction and designed a multi-feature classifier based on machine learning models to find relations, ultimately generating triplets. Although the system automates the extraction and storage of triplets through a pipeline process, it lacks the ability to categorize interaction types with fine granularity. Park et al. (2023) also constructed an automatic system for extracting biomedical entities and relations. For given PubMed identifiers (PMIDs), the system can automatically retrieves paper abstracts and uses the BioBERT to identify diseases and gene/protein entities and relations, presenting the results with labeled documents and entity relationship graphs. Gajendran et al. (2023) employed a BERT-BiLSTM-CRF framework to automatically extract biomedical entities such as diseases, chemicals, and proteins. Using the SciBERT, they identified and classified chemical-disease and chemical-protein relations, and directly combined the outputs of the two models to form biomedical triplets. They further constructed a COVID-19 knowledge graph and conducted representation learning studies. Ai and Kavuluru (2023) introduced an end-to-end model for recognizing chemical-protein interactions, using a span-based method to identify potential entity boundaries and types, and incorporating entity labels when predicting entity relations. This method achieved the best results on the ChemProt dataset, although the introduction of fine-grained labels resulted in reduced precision for longer entity spans. Zhang et al. (2024) proposed a multi-view merged representation model for extracting biomedical chemical-disease relations. During entity extraction, prior knowledge from the CTD database was incorporated, and a combination of U-Net network layers and GCNs was used to identify relations between entities, thereby capturing global semantic information. This approach outperformed the benchmark models in the task of extracting chemical-disease interaction relations. Li et al. (2024) employed ensemble learning techniques and pretrained language models like

PubMedBERT for biomedical entity recognition. After rule-based corrections, they used pretrained language models with prompt templates and a voting mechanism to determine relation types. Their research also found that ChatGPT-3.5 and ChatGPT-4 performed significantly worse than existing methods without fine-tuning.

Joint recognition methods can effectively integrate features between entities and relations, thereby simultaneously improving entity recognition and relation extraction performance. To achieve overlapping entity relation extraction, Luo et al. (2020) improved the method applied by Zheng et al. (2017) by defining a labeling system that simultaneously included DDI entity types and relation categories. They used the ELMo-BiLSTM-CRF-Att model to jointly recognize triplets via sequence labeling. For entities involving overlapping relations, specific label types were annotated and corresponding triplets restored based on overlapping relation extraction rules. However, the labeling system introduced hundreds of combination labels to distinguish overlapping relations, significantly increasing classification difficulty. Additionally, owing to the limitations of the labeling system, it could only match triplets based on the nearest-neighbor principle, rendering it unable to capture distant DDI relations (Sun et al., 2022). Fei et al. (2021) proposed a span-based graph neural model to address overlapping entity relation extraction. They designed a span attention algorithm to identify all potential entity spans and used an entity scorer to determine candidate entity spans. For two spans determined as genuine, they employed relation scorers to identify DDI types. This method effectively solved the predicament of overlapping entity recognition. However, since it used two independent output layers to decode entities and relations, it still suffered from error propagation. Sun et al. (2022) combined BERT and a machine-reading comprehension framework, treating the relation as a Query and the sentence as Context. They concatenated the Query and Context as input, expecting the model to output all entities (Answer) corresponding to that relation. To generate annotated results, they further extracted all DDI triplets using label types and entity combination rules. This method used two separate decoders to determine the initiation and termination positions of spans, impeding it from sharing entity span information. Additionally, this method also relied on the nearest-neighbor rule when combining entity pairs, leading to the omission of certain triplets. Sun et al. (2023) proposed a biomedical triplet extraction method that combines table filling and feature filtering. This method transforms one-dimensional text information into a two-dimensional structured table format, where each cell contains entity and relation information. A 2D-CNN network is utilized to jointly learn the interactions between the subtasks of entity recognition and relation extraction. This approach achieved commendable F1-scores on both the ADE and DDI datasets. However, the feature filtering module of this method cannot ensure the complete removal of redundant information, thus the recognition of drug entities in DDI triple extraction tasks remains inadequate. To address this issue, Wang et al. (2023) introduced a deep purified feature mining model that employs Lightweight Shared Representation Learning (LSRL) and a Hybrid Convolutional Neural Network (HCNN) to learn shared representations. On this basis, they designed a Task-aware Information Bottleneck (TIB) to obtain purified task-specific features, thereby enhancing the performance of both NER and RE subtasks. Al-Sabri et al. (2023) presented an automatic modeling framework for joint extraction of biomedical entities and relations using a Multi-View Graph Neural Network Auto-Modeling System (MVGNAS). This framework utilizes Graph Neural Architecture Search (GNAS) technology to automatically select the optimal GNN configuration. In this system, nodes represent entity span types, and edges represent the types of relations between entities. The integration of global node embeddings from different views completes the joint modeling of entities and relations.

While the aforementioned joint recognition methods enable the sharing of single-model parameters, they still employ mutually independent linear layers or multilayer perceptron classifiers for decoding entities and relations. Moreover, both joint and pipeline methods tend to rely on custom-defined rules for determining relations between entities, which remains insufficient for more complex overlapping entity and relation extraction. Additionally, constrained by the transformer encoder architecture, they cannot handle the extraction of discontinuous entities and classification of their interactions. In contrast, LLMs such as GPT-3 (Brown et al., 2020), which employ transformer decoder architecture, are capable of simultaneously performing sequence labeling and text classification tasks in a sequence-to-sequence manner (Mazumdar et al., 2023). This aligns well with the tasks of entity extraction and DDI classification. Furthermore, since the generative model is not restricted by the output form, it possesses inherent advantages in handling triplet extraction involving overlapping relations and discontinuous entities. Therefore, we designed a generative DDI triplet extraction method based on the transformer decoder architecture. We leveraged LLaMA series models (Touvron et al., 2023a; Touvron et al., 2023b) as foundational models and employed instruction tuning and CoT techniques to further enhance the LLMs' task-understanding and reasoning capabilities. By employing parameter-efficient tuning, we also managed to mitigate training costs.

4. Methods

In this section, we provide a detailed explanation of our proposed CoTEL-D3X method, covering aspects such as the method overview, CoT-enhanced Instruction Tuning, model architecture, and implementation details of the joint and pipeline approaches.

4.1. Overview of the proposed method

We combined Instruction Tuning with CoT Prompting from few-shot learning to design CoT-Enhanced Instruction Tuning. Through training, this approach effectively guided LLMs to align with task requirements, enabling DDI triplet extraction in a text generation format. Additionally, we constructed both joint and pipeline methods to explore the performance of these two mainstream modeling paradigms in the DDI triplet extraction task. The overall architecture of the proposed method is illustrated in Fig. 2.

The implementation process of the proposed method is as follows: For the DDI dataset, the data was preprocessed and then converted into Instruction Tuning and CoT formats for both joint and pipeline methods. During the training phase, instruction text with explicit task requirements and CoT-formatted datasets were used for Instruction Tuning in both joint and pipeline methods. The generative LLMs decomposed the complex task into logically coherent steps in an end-to-end manner, learning each step separately while using the output of preceding steps as known conditions for subsequent tasks.

During inference, the initialization instruction was concatenated before the text to be extracted. The model, leveraging its acquired reasoning abilities, coherently inferred entities and relations and directly outputted the specified DDI triplets in the desired format through text generation. In cases where multiple triplets existed within a sentence, the triplets were outputted sequentially in the order of entity occurrence, separated by newline characters.

4.2. Instruction Tuning

When applying pre-trained models, such as BERT, to downstream tasks, the current standard practice is to perform further Supervised Fine-Tuning (SFT), also known as task tuning, on the task-specific dataset to align the model's outputs with task requirements. Since no explicit instruction is conveyed to the model regarding task requirements, this typically requires a substantial amount of specifically labeled training data. However, Instruction Tuning (Wei et al., 2021) adds natural language instructions before the input sequence, explicitly informing the model of the task requirements, thereby enhancing the model's understanding of the task during training.

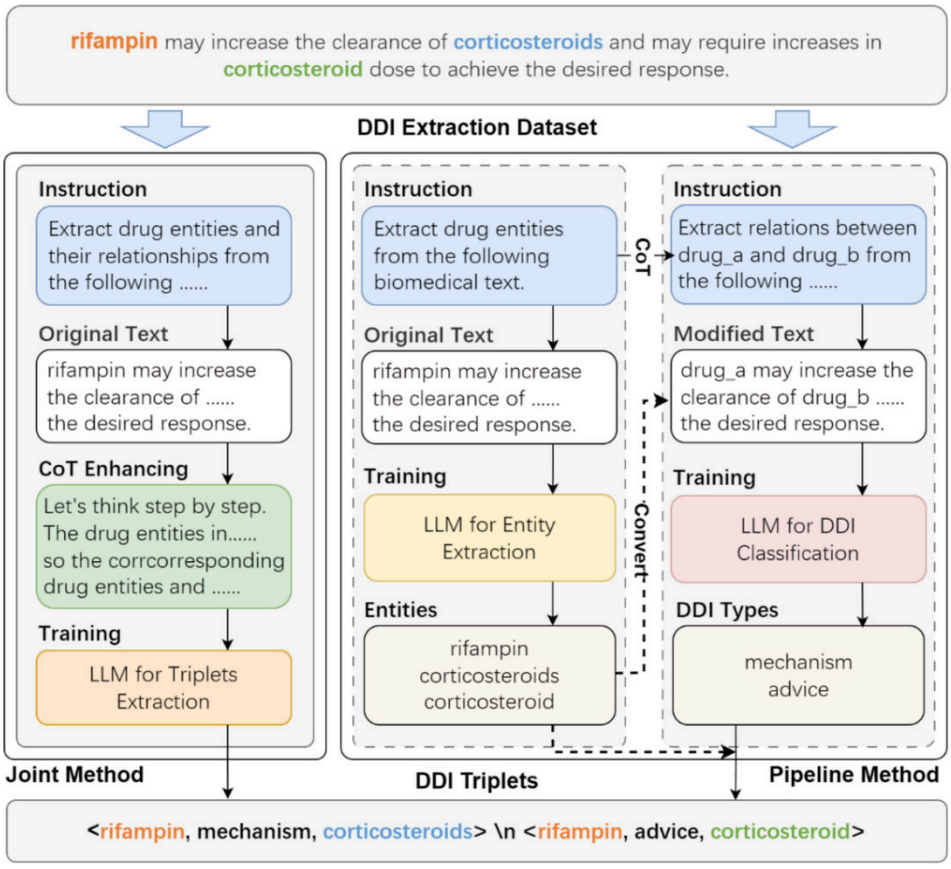


Fig. 2. Overall Architecture of the Proposed Method.

For LLMs, the predominant application method is In-context Learning. Here, the model was provided with prompts in a conversational format or supplied with several examples to further guide it, achieving interactive generative responses. For instance, Fig. 3(a) is a typical example using zero-shot learning for DDI triplet extraction. However, even with the most powerful models, such as GPT-4, limitations are encountered when dealing with complex domain-specific information extraction tasks. Therefore, further training of LLMs through Instruction Tuning to adjust model parameters is necessary, enabling them to learn the unique content and format features of DDI triplets.

4.3. CoT enhanced Instruction Tuning

Regardless of their human-defined discrete prompts or automatically identified continuous prompts, the inferential capabilities of current LLMs still fall short. Wei et al. (2022) discovered that introducing intermediate reasoning steps into model inference significantly enhances the model's reasoning abilities. The CoT concept was derived from human intuitive thinking. It involves breaking down complex tasks into several simple intermediate steps, progressively solving them, focusing on the current task at each step, and using the generated results for subsequent tasks. By imparting LLMs with a similar human-like coherent reasoning ability, more accurate answers can ultimately be generated.

The complex task of DDI triplet extraction can be broken down into

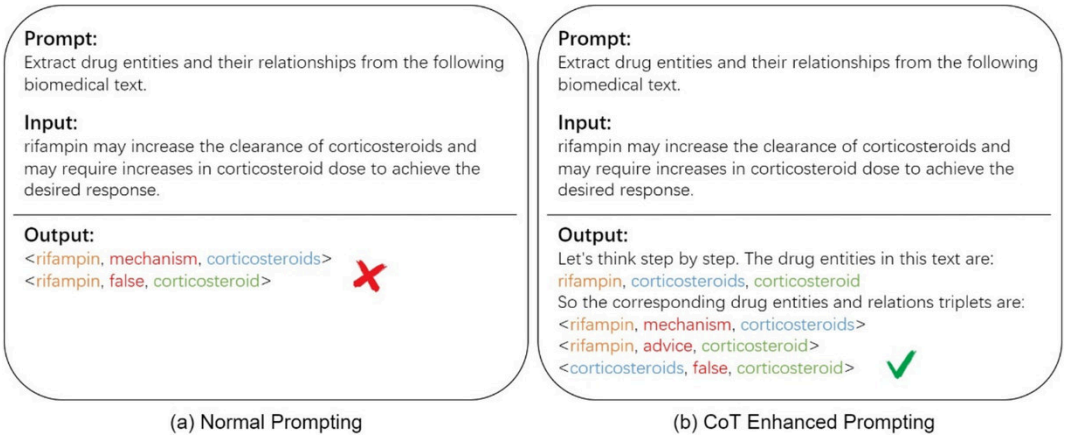


Fig. 3. Comparison of Different DDI Triplet Extraction Prompting Methods.

independent subtasks: (1) entity extraction and (2) classifying relationships between entity pairs. This process can be explicitly split into two separate steps, similar to traditional pipeline methods, or it can be performed implicitly using a joint approach. Fig. 3(b) illustrates the concept of completing triplet extraction using CoT-enhanced prompting. Unlike conventional prompting, which directly generates DDI triplets, this method initially identifies entity mentions within the text to be extracted, combines them into entity pairs, determines the relation type, and finally, generates all DDI triplets in sequence. This human-like thinking process refines the problem-solving workflow, ultimately enhancing extraction performance.

However, CoT Prompting was initially designed for In-context Learning (few-shot learning; prompt learning), which entails providing CoT examples for zero-shot or few-shot prompting without additional training or gradient updates for LLMs. Nevertheless, this approach requires models with hundreds of billions of parameters (Wei et al., 2022; Kim et al., 2023) and does not yield satisfactory results for biomedical information extraction tasks. To further align model outputs with requirements, we extended its application to Instruction Tuning, creating CoT-Enhanced Instruction Tuning.

Specifically, on constructing the training and test dataset, we redefined the LLM's input and output formats, allowing the model to learn reasoning ability more effectively during Instruction Tuning. Through multiple iterations of training rather than In-context Learning, the parameters of the original general-purpose LLMs underwent domain-specific and task-specific adjustments, ultimately solidifying this reasoning process.

Compared to In-context Learning and traditional Instruction Tuning, CoT-Enhanced Instruction Tuning exhibited the following characteristics in DDI extraction:

(1) During training, the model leveraged instructions and chain-of-thought (CoT) reasoning to allocate limited computational resources more effectively across intermediate steps of complex tasks, enabling more reasonable joint modeling and reasoning for entity and relation extraction tasks.
(2) During inference, the model, having undergone Instruction Tuning, inherently possessed CoT reasoning capabilities. As a result, no additional learning examples were required, allowing the model to automatically trigger CoT reasoning, thereby accelerating computation and reducing costs.
(3) The Chain-of-Thought mimicked the human process of incrementally thinking through a problem to arrive at an answer. This enhanced the interpretability of LLMs and the reliability of the results, potentially uncovering issues within the reasoning process.

We elaborate on how we designed data formats tailored to both joint and pipeline methods in Sections 4.5 and 4.6.

4.4. LLMs

An LLM, also known as a Foundation Model, refers to a category of models that undergo self-supervised pre-training on large-scale datasets and can be adapted for various downstream tasks (Bommasani et al., 2021). Based on differences in model architecture, they can be categorized into three main types: (1) Encoder-Only structures, such as BERT and ALBERT; (2) Encoder–Decoder structures like BART, T5, and GLM; and (3) Decoder-Only structures, including GPT and LLaMA, among others. In particular, following the popularity of ChatGPT, generative Decoder-Only LLMs have undergone rapid development. The number of parameters in these models has vastly surpassed those in BERT, typically ranging in the tens, or even hundreds, of billions of parameters. This characteristic allows them to perform downstream tasks without fine-tuning, relying solely on In-Context Learning through provided prompts or a few examples. Additionally, research has demonstrated (Touvron et al., 2023b) that further SFT with a limited quantity of high-quality data can unleash even greater potential in these LLMs.

Therefore, we selected two of the current top-performing open-source models: LLaMA (Touvron et al., 2023a) and LLaMA-2 (Touvron et al., 2023b) as the foundational models. The LLaMA series comprises models that have been pre-trained using an improved Transformer Decoder architecture (similar to GPT-3) (Vaswani et al., 2017) and are augmented with additional training data. Resultantly, the performance of the LLaMA model's 13B version has surpassed that of the GPT-3 model's 175B version on certain benchmark tests. The LLaMA-2 model further extends sequence length and corpus size, modifies attention mechanisms, and achieves comparable results with current closed-source dialogue models, such as ChatGPT and PaLM. Fig. 4 illustrates the architecture of the LLaMA series 13B model.

The input text for LLaMA comprises a combination of CoT-enhanced instructions, sentences to be extracted, and all DDI triplets. First, its tokenizer employs the Byte Pair Encoding algorithm (Sennrich et al., 2015), implemented via SentencePiece, to transform the input text into "subwords" and map it into embeddings of 5120 dimensions. Second, after undergoing RMSNorm normalization (Zhang and Sennrich, 2019), it calculates attention in conjunction with Rotary Positional Embeddings (RoPEs) (Su et al., 2021) and the Causal Mask method, followed by the application of the SwiGLU activation function (Shazeer, 2020) for transformation. Finally, a Softmax classifier is used to compute probabilities, and the output sequence is generated in an autoregressive manner from left to right, with each token in the output sequence decoded into its corresponding text.

RMSNorm. The calculation is based on equation (1), where ϵ is a considerably small constant used to prevent division by zero, and W represents learnable weights. Compared with normalization techniques like LayerNorm used in models such as BERT and GPT, the RMSNorm normalizing function eliminates the need to calculate mean and variance and omits the bias parameter β. This results in improved training speed and stability.

$$RMSNorm : y = \frac{x}{\sqrt{Mean(x^2) + \epsilon}} * W \tag{1}$$

$$Mean(x^2) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 \tag{2}$$

SwiGLU. ReLU has been replaced with the SwiGLU activation function, which is computed using equation (3). In this equation, W, V, b, and c are learnable parameters, and β is a constant. Compared with ReLU, SwiGLU yields smoother computation results, enhancing the network's expressive and generalization capabilities.

$$SwiGLU : y = Swish_\beta(xW + b) \otimes (xV + c) \tag{3}$$

$$Swish_\beta(x) = x * sigmoid(\beta x) \tag{4}$$

RoPE. This is a form of relative positional encoding represented by absolute positional encoding. It not only incorporates absolute position information into Q and K but also introduces relative positional information during attention calculation. RoPEs are applied in every Transformer Decoder layer to achieve this.

4.5. Joint method

The joint DDI triplet extraction method based on Instruction Tuning is illustrated in Fig. 5. Considering that LLMs are highly sensitive to instructions, after exploratory experiments, the instruction provided for model training was "Extract drug entities and their relations from the following biomedical text." This instruction was subsequently concatenated with the sentence from which entities and relations needed to be extracted. The example here is a special case from Fig. 1, which displays overlapping relations and discontinuous entities. The primary objective

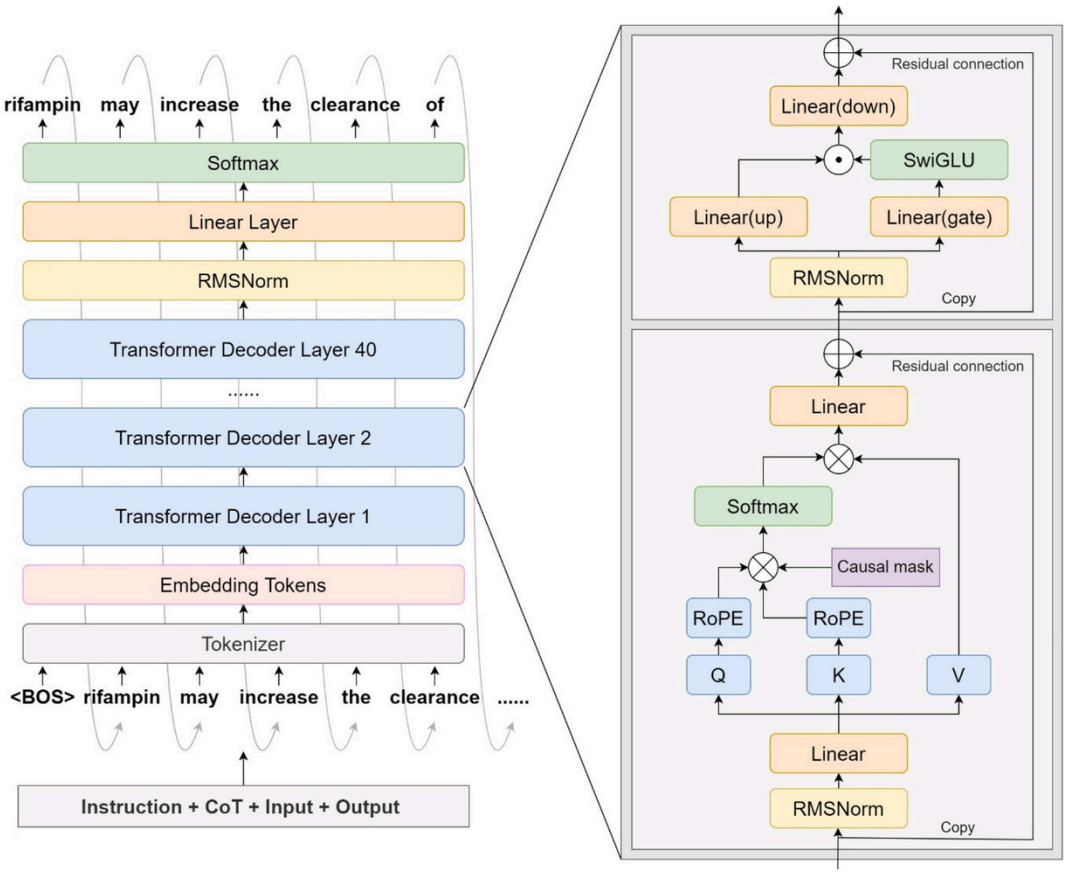


Fig. 4. The Architecture of the LLaMA Series 13B Model for DDI Triplet Extraction.

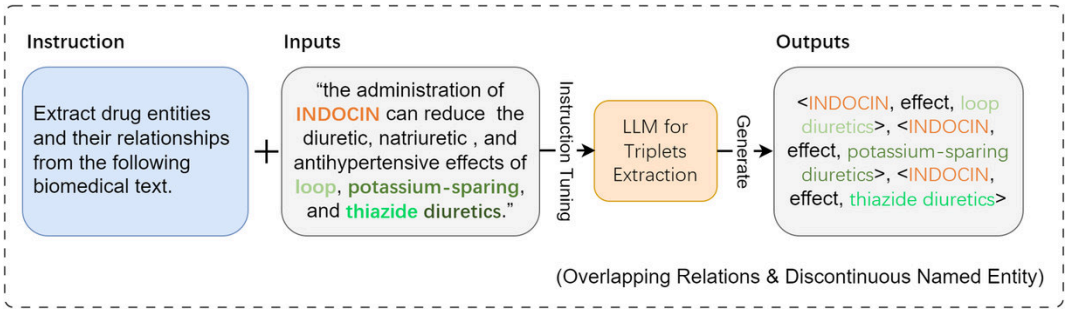


Fig. 5. Joint Extraction of DDI Triplets Based on Instruction Tuning.

of the LLM was triplet extraction, and consequently, the output was structured to include all DDI triplets found in the sentences, presented in sequential order. Each triplet comprised entity pairs and their corresponding relations. Different triplets were separated by newline characters ("\n").

Due to the flexibility of generative LLMs in defining input and output formats, this approach allowed us to directly inform the model of task requirements in natural language. Therefore, the method presented in Fig. 5 effectively addressed the challenges of extracting overlapping entity relations and discontinuous entities. Compared with current mainstream approaches based on BERT-like models, this method not only offers a simpler model architecture but also circumvents the introduction of any manually defined rules or external knowledge. However, since the DDI triples extraction involves two processes: extracting drug entities and relation types, it presents greater complexity. Therefore, we further employed CoT-enhanced Instruction Tuning, as illustrated in Fig. 6, for model training. We named this approach CoTEL-D3X (joint).

The joint method implicitly accomplished CoT reasoning during the Instruction Tuning process, enabling the LLMs to systematically address complex problems, thereby enhancing the performance of Instruction Tuning. The model's instruction and input remained consistent with the previously mentioned format. On constructing the output text, we initially instructed the model using the phrase "Let's think step by step." While this statement is exceedingly concise, it can stimulate the model to engage in unguided CoT reasoning (Kojima et al., 2022). Subsequently, for the DDI triplet extraction task, implicit CoT reasoning was conducted. For ease of explanation, we broke it down into two steps: entity extraction (Step 1) and entity relation inference (Step 2). However, in practice, the model seamlessly completed this process with a single output, without any manual intervention.

Initially, the model was guided to recognize all drug entity names within a sentence. This was achieved using the Step 1 instruction: "The drug entities in this text are: [entity list]," which guided the model to

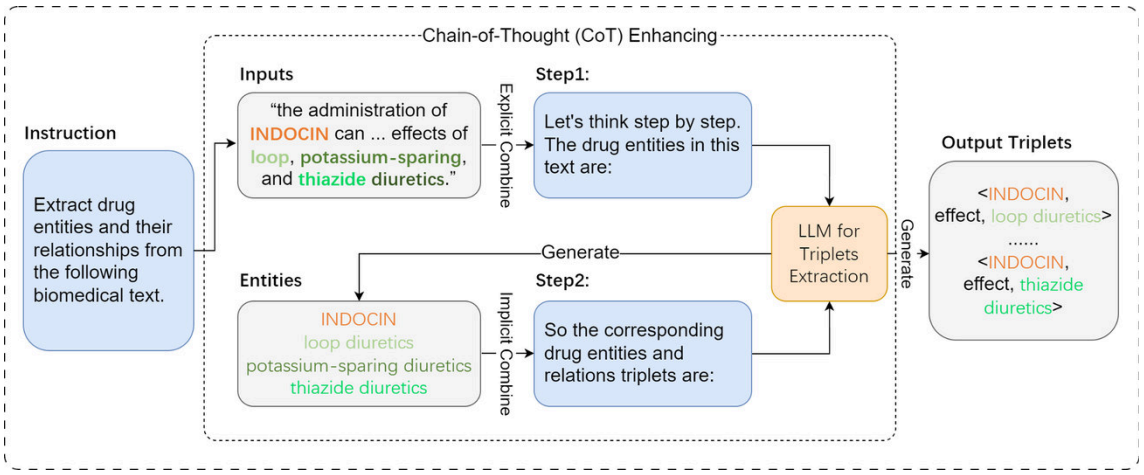


Fig. 6. Joint Method of DDI Triplet Extraction Enhanced with CoT-based Instruction Tuning.

recognize all potential entities. The "entity list" comprised all entities that appeared in the sentence in order. To ensure that the model focused on correctly identifying entity boundaries, entity categories were not considered; only entity names were recognized (Zaikis & Vlahavas, 2021).

Subsequently, the identified entities were concatenated with the Step-1 instruction and included at the end of the initial input. Thereafter, the Step-2 instruction was provided: "So the corresponding drug entities and relations triplets are: [triplet list]." This instructed the model to use this information, along with the sentence content, to infer and identify all DDI triplets within the sentence. This process was not artificially split into two separate outputs; instead, the model implicitly completed entity recognition as an intermediate step during inference and directly generated all DDI triplets. Resultantly, it avoided error propagation issues observed in pipeline methods and effectively modeled entity and relation information throughout the encoding and decoding process.

As generative models are not bound by predefined labels, they can automatically generate all DDI types involving overlapping entities and list all triplets in sequence. In the case of discontinuous entities, the model can also seamlessly combine separate entity fragments within the sentence to form complete entities and generate the corresponding DDI triplets.

4.6. *Pipeline method*

The joint method is based on a single model that implements the CoT technique. While it can directly output triplets, the length of the context sequence can become substantial. In cases where numerous triplets are to be extracted, the model's output may be truncated due to exceeding the sequence length limit, resulting in the inability to generate all DDI triplets. Furthermore, during model training, jointly modeling both tasks can be computationally intensive. Considering the limited scale of the DDI dataset, it may not capture the full spectrum of associations between entities and relation types.

Hence, we also designed a CoT-enhanced, pipeline-style DDI triplet

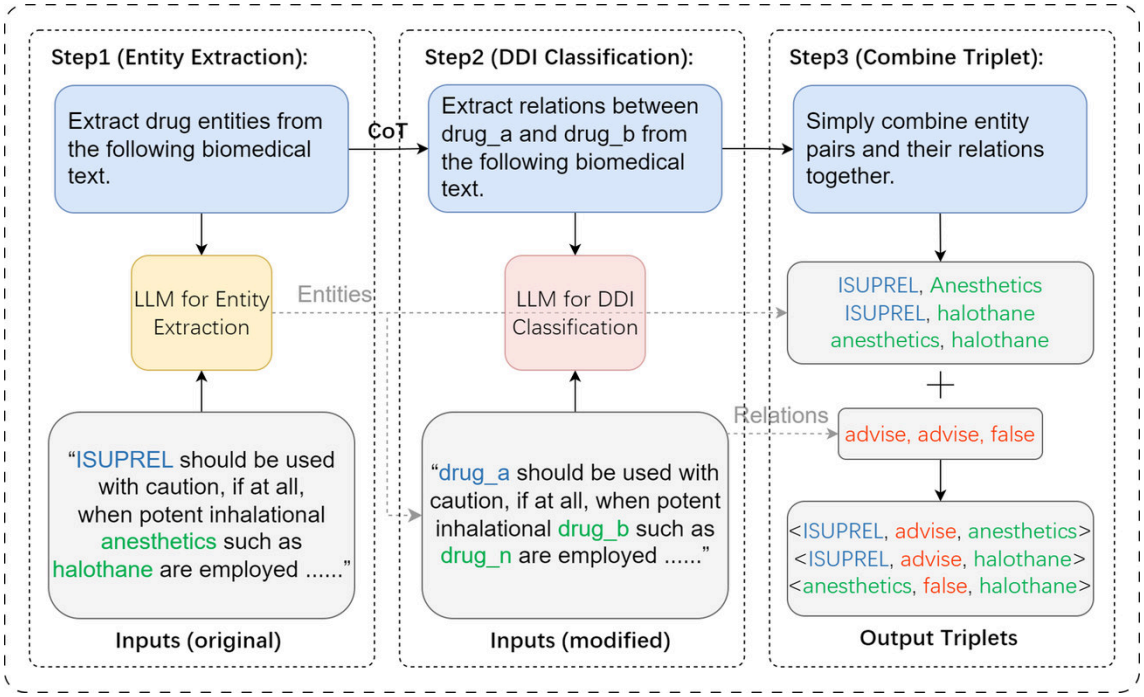


Fig. 7. Pipeline Method for DDI Triplet Extraction Enhanced with CoT.

extraction method. We referred to this method as CoTEL-D3X (pipe), and its implementation is illustrated in Fig. 7. The pipeline method breaks down triplet extraction into two conceptually linked subtasks: entity extraction (Step 1) and DDI classification (Step 2). Each of these tasks was trained separately via Instruction Tuning, explicitly completing the two CoT steps. The output from Step 1′s model was utilized to construct the input for Step 2, and finally, all DDI triplets were generated using a straightforward combination of outputs.

First, based on Instruction Tuning, all entity names within the sentence were generated. After exploratory experiments, the final instruction for Step 1 was as follows: "Extract drug entities from the following biomedical text." For the "N" entities generated by the LLM, all possible pairwise combinations were formed, resulting in "N (N-1)/2" entity pairs to be classified. Since a generative LLM was employed, potential "n" discontinuous entities were automatically reconstructed as continuous entity names in the model's output, ensuring coverage of any form of entity mention. This also reduced the possibility of error propagation, thus ensuring the precision of subsequent tasks.

For the entity pairs generated in Step 1, the corresponding sentence was reproduced "N (N–1)/2" times. The two target entities in each entity pair were anonymized to avoid introducing meaningless segments during serialization and enhance performance, according to the approach proposed by Zaikis and Vlahavas (2021). Specifically, the target entities in the sentence were replaced with "drug_a" and "drug_b," while the remaining entities were replaced with "drug_n." Subsequently, the Step-2 instruction was articulated as follows: "Extract relations between drug_a and drug_b from the following biomedical text." Through Instruction Tuning, the model was used to determine the relation type between the target entities (Suárez-Paniagua et al., 2019). The model's output was the DDI relation category for each entity pair. Finally, the entities and relations were straightforwardly combined, forming all DDI triplets in the order they appeared in the sentence. Since all possible entity combinations were considered, this method ensured the identification of all overlapping entity relations.

The pipeline-style DDI triplet extraction method allowed for focused training on each respective task, optimizing for the distinct tasks of entity extraction and relationship classification. Since the outputs of the two stages were entity names and relationship types, respectively, the generated sequence lengths were considerably shorter than those in the joint method. This also enabled the pipeline method to achieve faster generation speeds with smaller memory consumption during inference. However, since entities need to be anonymized in Step 2, the anonymization of discontinuous entities potentially leads to erroneous matches. To avoid exacerbating error propagation, the current pipeline method focuses only on continuous entity situations when determining DDI types in Step 2.

5. Experiments and results

In this section, we first briefly introduced the datasets used for performance evaluation, the preprocessing process, and the model hyperparameters settings. For the analysis of the experimental results, we comprehensively compared the performance between the proposed method and current mainstream In-context Learning methods, generative methods, and extractive methods, and conducted a case study on real-world dataset.

5.1. Dataset

5.1.1. The DDI extraction 2013 dataset

The DDI Extraction 2013 dataset (Herrero-Zazo et al., 2013) is a gold standard corpus that has been manually annotated for drug entities and their interactions. This dataset was first introduced at the DDIExtraction 2013 Shared Task challenge (Segura-Bedmar et al., 2013), a subtask of SemEval-2013 Workshop, and has been widely used for evaluation of drug entity and DDI extraction methods. (Fei et al., 2021; Sun et al., 2022; Yang et al., 2023a). This entire corpus comprises 1,025 documents, consisting of 233 MedLine abstracts and 792 DrugBank documents that describe drug–drug interactions. At document level, sentences and their corresponding annotations are stored in XML format, and the corpus is pre-divided into training and test sets, encompassing separate datasets for entity recognition and relation classification. The training set includes 714 documents with a total of 6,976 sentences, while the relation classification test set comprises 191 documents with 1,299 sentences.

There are four entity categories in total, namely, "drug," "brand," "group," and "drug_n." "Drug" refers to the generic name of a medication, "Brand" is the trade or brand name of the medication, "Group" refers to a set of drugs involved in DDI, and "Drug_n" represents drugs not approved for human use. There are four types of DDIs: "mechanism," "effect," "advice," and "int." "Mechanism" refers to pharmacokinetic interactions, which include one drug's influence on the absorption and metabolism, among others, of another. "Effect" encompasses pharmacological effects, changes in symptoms, toxicity, and pharmacodynamic mechanisms. "Advice" refers to the provision of recommendations or guidance when two drugs are used simultaneously. "Int" denotes the presence of DDI without providing information regarding the interaction type. Additionally, a "false" type indicates the absence of interaction between two drugs. Table 1 presents statistics on the DDI corpus. Notably, one instance existed in the DDI dataset where DDI was labeled "true" but lacked a specified type, signifying an erroneous annotation. Therefore, we removed this record (pair id: DDI-DrugBank.d236.s29.p0).

The entire DDI dataset is significantly imbalanced in terms of category distribution. The "Drug" entity type constitutes 63 % of the dataset, while the "Drug_n" type accounts for only 4 %. Among the four DDIs, the "Effect" type accounts for 41 %, whereas the least frequent "Int" type constitutes less than 6 %. Furthermore, annotations that do not contain any DDI relations make up a substantial 85 % of the dataset (Zaikis & Vlahavas, 2021). Additionally, approximately 60 % of DDIs involve overlapping entity relations (Luo et al., 2020). Considering the aforementioned statistical characteristics of the dataset, we followed the data preprocessing steps used in prior studies (Xiong et al., 2019; Zhu et al., 2020; Shi et al., 2022) and performed the following preprocessing steps on the DDI dataset: (1) converting all text to lowercase and (2) filtering out entity pairs that comprised identical entities or cases where one entity was an abbreviation or acronym of another.

5.1.2. TAC 2018 DDI extraction dataset

Due to the fact that the benchmark dataset DDI Extraction 2013 originates from MedLine abstracts and DrugBank documents, it may overlook DDIs present in other textual forms. To explore the generalizability of our proposed method, we further validated it on another DDI dataset from a different source, TAC 2018 DDI Extraction (Demner-Fushman et al., 2018), in order to examine potential data biases.

The TAC 2018 DDI Extraction dataset originates from the U.S. Food and Drug Administration (FDA) and the National Library of Medicine (NLM) and consists of structured product label (SPL) files for prescription drugs. Each SPL contains several sections, with each section comprising multiple sentences. The dataset contains 325 SPLs in total, with the training set consisting of XML-format files for 22 SPLs and 180 SPLs annotated with slightly different formats. Two test sets are provided, containing 57 and 66 SPLs, respectively. All SPLs were manually annotated by FDA and NLM experts for three types of DDIs: Pharmacokinetic (PK), Pharmacodynamic (PD), and Unspecified (U). In addition to common text data, the source also includes tables and other types of DDI data, which can be used to assess our method's performance on diverse data sources.

5.1.3. Adverse drug Reaction (ADR) extraction dataset

Models that perform well on the DDI extraction task may suffer from limitations in transferability and generalization ability. Therefore, we

Table 1
Statistics of Entities and Relations in the DDI Extraction 2013 Dataset.

	Type	Train			Test		
		DrugBank	MedLine	Overall	DrugBank	MedLine	Overall
Entity	Drug	8,197	1,228	9,425	1,518	346	1,864
	Brand	1,423	14	1,437	347	22	369
	Group	3,206	193	3,399	626	41	667
	Drug_n	103	401	504	21	119	140
	Total	12,929	1,836	14,765	2,512	528	3,040
DDIs	Mechanism	1,257	62	1,319	278	24	302
	Effect	1,535	152	1,687	298	62	360
	Advice	818	8	826	214	7	221
	Int	178	10	188	94	2	96
	False	22,216	1,555	23,771	4,381	356	4,737
	Total	26,004	1,787	27,791	5,265	451	5,716

also introduced another biomedical information extraction task: Adverse Drug Reaction (ADR) Extraction (Roberts et al., 2017). ADR refers to unexpected events or adverse reactions associated with medication, which may lead to prolonged hospital stays, disability, or even death (Modi et al., 2024). The goal of this task is to use Natural Language Processing (NLP) techniques to extract ADRs from unstructured biomedical and clinical texts, thereby accelerating the review of event reports in the FDA Adverse Event Reporting System (FAERS).

The ADR Extraction dataset also originates from the SPLs published by the FDA, and the construction of the gold-standard dataset was manually performed by FDA and NLM experts. The training set consists of 101 SPLs, and the test set contains 99 SPLs. The dataset includes annotations for three types of ADR relations: Negated, Hypothetical, and Effect. The ADR Extraction dataset is organized and annotated at the document level, and during data preprocessing, we only considered ADRs that appeared within the same sentence. By examining the performance of the proposed method on the ADR Extraction dataset, we aimed to verify the method's transferability and generalization ability.

5.2. Hyperparameters

During model training, for the joint method, the hyperparameters were set as follows: a learning rate of $2e^{-4}$, 20 epochs, maximum sequence length of 1,024, batch size of 16, micro-batch size of 8, and warmup ratio of 0.1. For the pipeline method, the entity extraction model had a maximum sequence length of 512 and 20 epochs, while the relation classification model had a maximum sequence length of 32 and 15 epochs. Other hyperparameters for the pipeline method were consistent with those of the joint method.

During inference, we primarily adhered to the original LLaMA-2 hyperparameters (Touvron et al., 2023b), with a temperature of 0.1, top-p value of 0.75, repetition penalty of 1.2, top-k value of 40, number of beams set to 4, sample set to "false," and maximum new tokens aligned with the maximum sequence length.

Both methods utilized the 13B version of LLaMA and LLaMA-2 official weights for initialization, featuring a hidden 5120 dimension, 40 self-attention heads, and 40 layers. The AdamW optimizer was employed. We adopted PEFT's QLoRA (Dettmers et al., 2023) for both model training and inference, inserting adapters into all fully connected linear layers within each transformer layer. The hyperparameters were as follows: r value of 64, lora alpha of 16, and lora dropout of 0.05.

5.3. Evaluation metrics

We assessed the performance of our model's DDI triplet extraction using three key metrics: precision, recall, and F1-score. Considering the concurrent extraction of entities and relationships, we adhered to the same evaluation criteria as those of previous studies, such as those by Luo et al. (2020), Fei et al. (2021), and Sun et al. (2022), where an extraction was only considered correct if both entity boundaries and the DDI category had been accurately predicted.

5.4. Performance of the proposed method

To evaluate the performance of the proposed CoTEL-D3X method, we conducted a comparative analysis with conventional SFT methods with the aim of understanding the impact of Instruction Tuning and CoT on enhancing model performance. For the joint extraction method, we compared our model with LLaMA-2 using SFT and Instruction Tuning, albeit without CoT. For the pipeline extraction method, as the two designed steps could explicitly accomplish CoT, we only compared our model with the SFT model. Additionally, to investigate the effectiveness of LLaMA-2, we trained models based on LLaMA while keeping other settings constant. The experimental results are presented in Table 2, where the "SFT" prefix indicates classic SFT, "Instruct" prefix indicates Instruction Tuning, and "CoT" prefix indicates CoT-enhanced Instruction Tuning.

On comparing our model with LLaMA /LLaMA-2 under different strategies, the incorporation of both Instruction Tuning and CoT evidently enhanced the capability of DDI triple extraction. Following Instruction Tuning, the F1 scores for the joint and pipeline methods improved by 0.95 % and 3.17 %, respectively. Notably, the former witnessed a remarkable 3.34 % increase in the F1 score after the introduction of CoT. These findings demonstrate the effectiveness of the proposed approach. Instruction Tuning explicitly informs the model of its task requirements using natural language, rendering it more goal-oriented than the classical SFT approach. In contrast, CoT inclusion decomposes the original complex problem into subtasks, allowing the model to engage in reasoning while possessing reflective capabilities. This, in turn, enables the model to better capture associations between entities and relations, thereby stimulating its capacity for distributed reasoning. Importantly, this reasoning process is conducted within the same modeling space, leading to a significant enhancement in performance.

The pipeline approach using SFT alone has already demonstrated remarkable performance; notwithstanding, it remains inferior to the joint method. On incorporating Instruction Tuning, the model gained an enhanced understanding of the specific task requirements for each of the two subtasks, leading to significant performance improvement that

Table 2
Performance of CoTEL-D3X.

Method	Precision (%)	Recall (%)	F1 (%)
SFT + LLaMA-2 (joint)	83.07	79.33	81.16
Instruct + LLaMA-2 (joint)	81.77	82.44	82.11
CoT + LLaMA (joint)	81.22	80.96	81.09
CoT + LLaMA-2 (joint)	**84.52**	**86.41**	**85.45**
SFT + LLaMA-2 (pipe)	85.42	84.77	85.09
CoT + LLaMA (pipe)	86.85	84.53	85.67
CoT + LLaMA-2 (pipe)	**88.23**	**88.30**	**88.26**

surpassed the already well-performing joint method. Notably, the pipeline approach not only performed well but also exhibited a slight advantage over the LLaMA-2 model within the joint method category. Previous studies (Wang et al., 2018; Wei et al., 2020; Sun et al., 2022) have considered joint methods to be typically superior to pipeline methods. However, our results diverge from this consensus. One potential explanation could be the presence of a substantial number of entity and relation overlaps within the DDI dataset, rendering joint recognition challenging. The joint method we designed does not incorporate any external knowledge or manually defined rules, adding to the complexity of tackling this task. In contrast, the pipeline approach, with its division into two relatively simpler subtasks and explicit instructions for each, allows the model to focus on task-specific requirements.

We ultimately selected the two methods initialized with LLaMA-2 weights as the proposed CoTEL-D3X for subsequent experimental comparisons.

5.5. Comparison with in-context learning methods

5.5.1. Baseline methods

One of the most common applications of LLMs is In-Context Learning, also known as few-shot learning. In this approach, task instructions are directly provided to the model, guiding its output in an interactive and conversation-like manner. Therefore, we compared our proposed method with ChatGPT, a widely recognized and high-performing general question-answering model, to evaluate its performance. ChatGPT is available in two versions: GPT-3.5 and GPT-4, both of which exhibit strong one-shot and few-shot learning capabilities across various tasks.

Following previous practices (Brown et al., 2020; Rae et al., 2021), we evaluated the triplet extraction capabilities under one-shot, five-shot, and 20-shot scenarios. We finalized the instruction template as follows:

> Extract drug entities and their relations from the following biomedical text. Note, only five types of relations are available for selection: mechanism, effect, advise, int, and false.
> Input: example sentence #1
> Output: triplet(s) extracted from example sentence #1
> Input: example sentence #2
> Output: triplet(s) extracted from example sentence #2
> ……

We conducted predictions using the OpenAI API with the identical test sets. The hyperparameter settings for generation were as follows: temperature set to 1, max tokens set to 1,024, top P set to 1, and both frequency and presence penalties set to 0.

5.5.2. Performance comparison results

The performance comparison between the two proposed CoTEL-D3X methods and two versions of ChatGPT in DDI triplet extraction is presented in Table 3.

While ChatGPT excels in a wide range of general natural language understanding and generation tasks, its performance significantly lags behind that of our proposed CoTEL-D3X method with respect to the extraction of complex DDI triplets within the biomedical domain. Although ChatGPT's few-shot learning capability improves with an increase in the number of learning samples, owing to the necessity of domain knowledge and a complex output structure, even the current leading question-answering model, GPT-4, struggles with this task. Moreover, we observed that the precision of the ChatGPT model is typically approximately twice its recall, indicating a significant deficiency in terms of answer coverage, rendering it less suitable for DDI triplet extraction tasks that involve numerous overlapping entity relations.

Table 3
Performance Comparison between CoTEL-D3X and ChatGPT.

Method	Precision (%)	Recall (%)	F1 (%)
GPT-3.5 (1-shot)	15.54	7.55	10.16
GPT-3.5 (5-shot)	32.46	12.88	18.44
GPT-3.5 (20-shot)	40.54	14.64	21.51
GPT-4 (1-shot)	32.93	14.36	20.00
GPT-4 (5-shot)	51.98	26.35	34.98
GPT-4 (20-shot)	57.76	29.15	38.75
CoTEL-D3X (joint)	84.52	86.41	85.45
CoTEL-D3X (pipe)	88.23	88.30	**88.26**

In contrast, our proposed approach not only outperformed ChatGPT by a wide margin but also demonstrated excellent and balanced performance across all three metrics. This highlights the fact that for domain-specific information extraction tasks, especially those involving complex tasks with multiple subtasks, current LLMs, which do not update model gradients and rely solely on few-shot learning, still fail to achieve a practical level of effectiveness. Our methods exceled with a parameter count much smaller than that of the ChatGPT series models. This not only demonstrated that domain fine-tuning and domain adaptation training remain among the most effective approaches, but also highlighted that the performance of our designed CoT-Enhanced Instruction Tuning surpasses that of CoT Prompting in In-context Learning scenarios.

5.6. Comparison with existing SOTA methods

5.6.1. Generative baseline methods

We selected the following current SOTA generative models for comparison:

(1) REBEL_{pt} (Cabot & Navigli, 2021): REBEL_{pt}, short for Relation Extraction By End-to-end Language generation, is a model constructed on the BART (Encoder–Decoder architecture) framework. It further pre-trains on relation triplet corpora constructed from Wikipedia. This model transforms relation triplet extraction into a seq2seq text generation task and has demonstrated outstanding performance in general relation extraction benchmark tasks.
(2) BioGPT (Luo et al., 2022): BioGPT is a representative model for text generation tasks in the biomedical domain. It continues pre-training on GPT-2 weights using biomedical PubMed abstract texts and has achieved SOTA performance on multiple biomedical generative tasks.

Considering that the performance of generative models is influenced by the number of parameters (Brown et al., 2020), for a fair comparison, we also included currently popular LLMs and evaluated them on the same test dataset:

(3) GPT-3 (Brown et al., 2020): GPT-3 is a representative model in the field of text generation. It is pre-trained in an autoregressive manner based on the transformer decoder architecture. GPT-3 is known for its ability to achieve SOTA performance in generative tasks with zero-shot or few-shot learning, without fine-tuning or gradient updates. While OpenAI does not release open-source models, it provides official weights for API fine-tuning. In this study, we initialized the GPT-3 model with curie (13B) weights, thus matching the parameter magnitude of our method.
(4) OPT (Zhang et al., 2022): OPT employs the same decoder model architecture as GPT-3 and achieves performance comparable to GPT-3. The model weights are open-sourced, and we selected the 13B version of the weights for fine-tuning.
(5) LLaMA (Touvron et al., 2023): LLaMA builds upon the GPT-3 architecture with improvements in embedding. It is one of the foundational models widely used for few-shot learning and domain fine-tuning applications, giving rise to numerous

excellent general conversation models, such as Alpaca and Vicuna. These models are all fine-tuned in a supervised manner on the DDI dataset, extracting both entities and relation triplets end-to-end. In cases with overlapping entity relations, all triplets are output in the order of their occurrence.

5.6.2. Performance comparison with generative approaches

As CoTEL-D3X is fundamentally a generative LLM, we initially compared its performance in DDI triplet extraction with that of the generative baseline approaches listed in the previous section. The experimental results are presented in Table 4.

Compared with the best baseline model, GPT-3, both CoTEL-D3X versions exhibited significant improvements in precision, recall, and F1 scores. The pipeline approach notably enhanced the F1 score by nearly 10 %. BioGPT, which currently performs well in biomedical generative tasks, has only achieved an F1 score of 40.76 %, which is even weaker than that of the classical LSTM neural network model (Table 5). This suggests that for generative models with a smaller parameter size, satisfactory information extraction results are challenging to achieve, even with domain-specific fine-tuning. Compared to BioGPT, the foundational model parameters of our proposed method were significantly increased. Both the general knowledge it encompasses and the domain-specific knowledge acquired during training far exceed that of BioGPT, resulting in a direct performance boost of more than double.

Large-scale generative LLMs, such as OPT, LLaMA, and GPT-3, which possess parameters on the scale of tens of billions, exhibit a $>$ 30 % improvement in performance compared with BioGPT owing to their ability to model world knowledge from massive real-world language data. This underscores the significant influence of parameter size in downstream tasks for generative models. However, because these models rely solely on task tuning without providing task instructions using natural language or specifically promoting their reasoning abilities, they remain inferior to the current SOTA extractive methods (Table 5). Compared with those of generative LLMs of similar parameter magnitude, the F1 scores of the two CoTEL-D3X versions increased by 6.59 % and 16.82 %. This once again emphasizes that the Instruction Tuning approach combined with CoT effectively stimulates the model's task understanding and reasoning abilities. By providing instruction during training, it explicitly guides the model to focus on key information, while CoT-based reasoning encourages the model to generate more accurate DDI triplets, both in terms of content and format.

5.6.3. Extractive baseline methods

The current extractive DDI triple extraction methods can be categorized into two approaches: joint and pipeline methods. Among these, the joint method, which enables joint modeling of entity and relation information, is currently the mainstream application. In contrast, while the pipeline method is prone to error propagation issues, it decomposes the complex task into two relatively simpler subtasks, rendering it more suitable for model parameter adjustments.

For joint methods, the following baseline models were used for comparison:

Table 4
Performance Comparison between CoTEL-D3X and Generative Approaches.

Method	Precision (%)	Recall (%)	F1(%)
$REBEL_{pt}$ (2021)	46.59	39.60	40.56
BioGPT (2022)	41.70	44.75	40.76
OPT (2022)	73.09	69.87	71.44
LLaMA (2023)	69.13	76.01	72.41
GPT-3 (2020)	80.01	77.73	78.86
CoTEL-D3X (joint)	84.52	86.41	85.45
CoTEL-D3X (pipe)	**88.23**	**88.30**	**88.26**

Table 5
Performance Comparison Between CoTEL-D3X and Extractive Methods.

Method	Precision (%)	Recall (%)	F1(%)
LSTM-LSTM-Bias (2017)	64.40	38.80	48.40
Graph Tagging (2018)	58.70	57.10	57.90
CasRel (2020)	66.90	64.80	65.80
Att-BiLSTM-CRF + ELMo (2020)	75.00	75.20	75.10
Span-Graph 2021	—	—	68.02
MRC4BioER (2022)	76.90	74.60	75.70
CoTEL-D3X (joint)	**84.52**	**86.41**	**85.45**
Bi-LSTM-CRF + CNN (2019)	71.26	62.82	66.78
TP-DDI (2021)	86.40	78.80	82.40
SPBERE (2023)	78.50	79.70	79.20
CoTEL-D3X (pipe)	**88.23**	**88.30**	**88.26**

(1) LSTM-LSTM-Bias: This is an end-to-end entity and relation joint extraction method proposed by Zheng et al. (2017). It defines a set of labels containing entity and relation information and performs DDI triple joint extraction using Bi-LSTM in a sequence-labeling manner.
(2) Graph Tagging: Wang et al. (2018) introduced this joint extraction method based on directed graph parsing. It uses a transition-based approach to simultaneously decode entities and relations while identifying overlapping entity relations.
(3) CasRel: Wei et al. (2020) proposed a cascade-style entity and relation triple extraction method. It initially identifies all candidate subject entities and subsequently, through an object tagger, simultaneously labels all potential relations and their corresponding object entities. PubmedBERT is used in DDI triple extraction experiments to replace the original BERT embeddings.
(4) Att-BiLSTM-CRF + ELMo: Luo et al. (2020) presented a joint sequence labeling-based extraction method. It introduces entity categories and overlapping relation information by improving the labeling rules proposed by Zheng et al. (2017) and employs ELMo as an encoder.
(5) Span-Graph: Fei et al. (2021) proposed a joint method that extracts DDI triplets after identifying all possible entity spans. This is achieved through an entity scorer and integrated relation scorers.
(6) MRC4BioER: Sun et al. (2022) introduced a method that transforms triplet extraction into a machine-reading comprehension task. It uses the relation type as a query and provides all entity spans as the corresponding answer.

The baseline pipeline methods used for comparison include:

(7) Bi-LSTM-CRF + CNN, a two-stage extractive approach proposed by Suárez-Paniagua et al. (2019), utilizes Bi-LSTM-CRF to identify all entities and subsequently employs a CNN to determine relation categories.
(8) TP-DDI, an end-to-end pipeline extraction method introduced by Zaikis and Vlahavas (2021), identifies entities and relations based on both sequence labeling and classification principles and utilizes BioBERT embeddings for DDI triplet extraction experiments.
(9) SPBERE, a span-based method presented by Yang et al. (2023a), deals with identified entity spans and incorporates additional entity type markers when inputting into the relation classification model. PubmedBERT is employed as the encoder in both stages.

5.6.4. Performance comparison with extractive methods

Table 5 presents the comparative results between the proposed CoTEL-D3X and SOTA methods for DDI triplet extraction. The first six baseline models are joint methods, while the latter three are pipeline methods.

Compared with that of other joint methods, the F1 score of our CoTEL-D3X (joint) achieved a substantial improvement of 9.75 % over

the current SOTA methods. This indicates that our approach is better at comprehensively sharing joint information between entities and relations to perform complex triplet extraction. The introduction of Instruction Tuning and Chain-of-Thought (CoT) technologies enables the CoTEL-D3X (joint) model to perform entity recognition and classification of relation types in a stepwise manner. Since both sub-tasks are modeled simultaneously from input to output, it essentially still trains a single extraction model, which avoids the impact of error propagation, thereby ensuring the performance of the CoTEL-D3X (joint) method.

Compared with other pipeline methods, our CoTEL-D3X (pipe) demonstrated a notable improvement of 5.86 % over the current strongest model. Compared with the best-performing joint methods, the CoTEL-D3X (pipe) achieved an improvement of 12.56 % in the F1-score. By dividing the complex task of triplet extraction into two relatively simpler subtasks, entity recognition and relation extraction, and explicitly informing the model of task requirements through instructions, the model is able to focus on enhancing the performance of these subtasks, thereby ultimately achieving an excellent F1-score. Unlike extractive methods, which often introduce manually defined indicators in the second stage of relation extraction to mark entity positions, our method does not use any artificial symbols throughout the process, thereby ensuring semantic coherence.

5.7. Performance on other biomedical information extraction datasets

5.7.1. Performance on TAC 2018 DDI extraction dataset

Unlike DDI Extraction 2013, which focuses on extracting both entities and relations, we concentrated on extracting DDI relations in order to facilitate comparison with existing methods. Therefore, during the model training and inference process, we directly used gold-standard entities as known information. For the CoTEL-D3X (pipe) method, the target entities in the sentence were replaced with “entity_a” and “entity_b”. We also provided the joint extraction results based on CoTEL-D3X (joint) as a reference, named CoTEL-D3X (joint w/o gold). Tables 6 and 7 present the comparative experimental results on Test Set #1 and Test Set #2, respectively.

The baseline methods used for comparison are:

(1) KLncLSTMsentClf: Proposed by Baruah and Kolla (2018), this method uses biomedical domain text-trained word embeddings and employs a neural network model with a 2-layer bidirectional LSTM to predict whether a sentence contains interactions and classify the interaction types.
(2) DDI-MuG: Proposed by Yang et al. (2023b), this method uses PubMedBERT to represent text and a multi-aspect graph to capture information from both sentence and corpus levels using two GCNs. There are three variations of this model: DDI-MuG-WGraph, which uses a Word co-occurrence-based graph; DDI-MuG-DGraph, which uses a Dependency-based graph; and DDI-MuG-all, which incorporates both types of graphs.

The results indicate that both of the methods we proposed achieved the best DDI relation classification performance on the two test sets of the TAC 2018 DDI Extraction dataset. Notably, on Test Set #2, the CoTEL-D3X (joint w/o gold) method, which does not use gold-standard entities (i.e., using the entities predicted in the first step as input for the second step), surpassed the method using gold-standard entities in terms of F1-score. This further underscores the significant impact of CoT on enhancing the model’s reasoning capabilities. The validation on the TAC 2018 DDI Extraction dataset demonstrated that our model performs well when dealing with DDI corpora from different data sources and relatively new datasets.

Table 6
Performance Comparison on TAC 2018 DDI Extraction Test Set #1.

Method	Precision (%)	Recall (%)	F1(%)
KLncLSTMsentClf (2018)	47.00	62.00	53.00
DDI-MuG-WGraph (2023)	71.70	71.20	71.50
DDI-MuG-DGraph (2023)	68.80	71.80	70.30
DDI-MuG-all (2023)	72.10	72.80	72.30
CoTEL-D3X (joint w/o gold)	67.28	73.83	70.40
CoTEL-D3X (joint)	**83.11**	**82.24**	**82.44**
CoTEL-D3X (pipe)	72.28	74.95	72.49

Table 7
Performance Comparison on TAC 2018 DDI Extraction Test Set #2.

Method	Precision (%)	Recall (%)	F1(%)
KLncLSTMsentClf (2018)	49.00	67.00	56.70
DDI-MuG-WGraph (2023)	71.00	72.60	71.80
DDI-MuG-DGraph (2023)	71.30	73.00	72.10
DDI-MuG-all (2023)	71.70	74.30	72.90
CoTEL-D3X (joint w/o gold)	74.93	71.39	73.12
CoTEL-D3X (joint)	**88.35**	**87.91**	**87.32**
CoTEL-D3X (pipe)	74.93	74.99	74.89

5.7.2. Performance on ADR extraction dataset

For ADR Extraction, in order to facilitate comparison with existing methods, we also focused on the recognition and classification of ADR relations, using gold-standard entities during relation extraction. For the CoTEL-D3X (pipe) method, the target entities in the sentence were similarly replaced with “entity_a” and “entity_b”. Table 8 presents the comparison results.

The baseline models used for comparison are:

(1) MC-UC3M: Proposed by (Martınez et al., 2017), this method uses a multi-feature SVM for relation classification.
(2) Attention-BiLSTM: Proposed by Dandala et al. (2017), this method is based on a BiLSTM model with an Attention mechanism, where locator tags are inserted before and after the entities in the input sequence.
(3) ADERel: Proposed by El-Allaly et al. (2022), this model formulates relation extraction as an N-level sequence labeling task, combining BioBERT text representation with a weighted GCN neural network and an attention mechanism to identify relation types.

In the ADR relation extraction experiments, both our joint and pipeline extraction methods outperformed the current state-of-the-art methods. We observed that in the DDI triplet joint extraction experiments, our pipeline method performed better than the joint method. However, the situation reversed in the TAC 2018 and ADR relation extraction experiments. Additionally, in the TAC 2018 experiments, the joint method with gold-standard entities showed a significant improvement in F1-score compared to when gold-standard entities were not used. This phenomenon may suggest that the overall triplet extraction performance of the CoTEL-D3X (joint) method is more sensitive to the accuracy of entity recognition.

5.8. Case study on DrugBank

To validate the applicability of our proposed method in real-world engineering environments, particularly in clinical settings, we

Table 8
Performance Comparison on ADR Extraction Dataset.

Method	Precision (%)	Recall (%)	F1(%)
MC-UC3M (2017)	46.70	49.97	47.32
Attention-BiLSTM (2017)	84.09	81.74	82.90
ADERel (2022)	72.96	70.14	71.52
CoTEL-D3X (joint)	**90.24**	**89.76**	**89.97**
CoTEL-D3X (pipe)	89.15	84.17	85.90

conducted usability verification experiments on the latest DrugBank online database. The DrugBank database contains information on drug names, drug targets, and drug–drug interactions (DDIs) for both FDA-approved and experimental drugs, making it one of the most widely used reference drug resources (Knox et al., 2024). We randomly selected 30 DDI descriptions for 10 drugs from the DrugBank database to create a real-world dataset, and applied our proposed CoTEL-D3X (joint) method to extract DDI triplets from these descriptions. To ensure the reliability of the results, we used the official evaluation tool, Drug Interaction Checker (DIC), provided by DrugBank to verify the extraction results. Table 9 presents five representative extraction results.

In Table 9, the bolded words in the "Description" column represent the correct drug names, the "Triplet" column contains the extracted DDI triplets, and the "DIC Severity" column provides the DDI severity rating given by the Drug Interaction Checker (DIC). The "DDI found" column indicates whether CoTEL-D3X successfully and accurately extracted the DDI triplets. As can be seen, the first four examples were correctly extracted, while in the fifth example, there was an error in recognizing the drug name boundary for "Edetate calcium disodium anhydrous." This case study demonstrates that our proposed method also shows good applicability in real-world, authentic texts and can handle most DDI extraction scenarios. However, it may encounter boundary identification issues when dealing with more complex drug names.

6. Discussion and implications

In this section, we focused on discussing the differences between the proposed method and existing state-of-the-art approaches in terms of model architecture design and performance. We analyzed the advantages of the proposed method and the reasons behind its performance improvements. Additionally, we elaborated on the potential theoretical and practical implications of our research and pointed out the existing limitations.

6.1. Discussion on the comparison with SOTA approaches

Current SOTA joint methods primarily rely on sequence-labeling paradigms, outputting manually defined tag sequences. Further programming is required to extract entity positions and relation categories and subsequently convert them into triplets. This procedure not only tends to overlook overlapping entity relations but also struggles to identify discontinuous entities. For instance, the method proposed by Luo et al. (2020) extracts overlapping entity relations by constructing joint labels and rules for entities and relations. However, because of the nearest-neighbor principle followed during triplet combination, it still overlooks relations between distant entities. Our proposed CoTEL-D3X (joint) method can directly generate triplets in the < entity1, relation, entity2 > format via text generation. This approach requires no further conversion in terms of both content and format, and effectively addresses the challenges of overlapping relations and discontinuous entities. Moreover, since entity recognition and relation extraction are implicitly reasoned through CoT, there is no need to define any manual rules for combining potential relations.

Furthermore, current methods have shortcomings in jointly modeling entities and relations. For example, while Fei et al. (2021) simultaneously encoded entities and relations in the hidden layers, they still had to identify entities first and then determine relations during prediction. The method by Sun et al. (2022) determines entity positions based on the input relation type and sentence but employs two independent classifiers to identify the initiation and termination positions of entities. Our proposed CoTEL-D3X (joint) method jointly models entities and relations at both the hidden and output layers, with a single decoder simultaneously producing both entities and relation triplets. This approach not only allows for full information sharing between the two steps but also avoids issues like incorrect combinations that can arise from span-based methods.

Current SOTA pipeline methods based on BERT-like models typically employ supervised fine-tuning for entity recognition and relation classification. After completing entity recognition, manual rules are required to correct potential errors in identified entities. For instance, Zaikis and Vlahavas (2021) defined a set of entity label conversion rules to rectify potential errors in entity recognition results. However, these rules simply replaced the label types of subsequent tokens with those of the initial token, potentially leading to incorrect tagging and subsequently affecting downstream relation extraction. Our proposed CoTEL-D3X (pipe) method explicitly guides the model's task requirements through instruction tuning. In the entity recognition task, it directly generates all entities in the sentence, including discontinuous entities. As a result, it eliminates the need for post-processing steps to correct entity recognition results, as required by sequence labeling methods, and does not require the introduction of any entity combination rules.

Moreover, the introduction of artificial tags, which do not exist in the pre-training stage, in the second subtask input sequence disrupts natural language modeling and may result in mismatches in downstream tasks, thereby impairing classification performance. Yang et al. (2023a), on one hand, implicitly informed the model to focus on entity information by introducing special initiation and termination markers for entities. On the other hand, they used a span prediction strategy to identify potential entities. However, this approach was limited by span length and computational complexity, especially for longer entities, where omissions could occur. Our proposed CoTEL-D3X (pipe) method conveys instructions to the LLM in natural language at both stages. This approach not only aligns with the data distribution characteristics during the pre-training phase but also ensures that no unseen labels are introduced throughout the entire process (Luo et al., 2022). Therefore, after completing entity extraction, even without deploying any correction strategies, our approach considerably enhances DDI-triple-extraction capability.

In real-world applications, especially in clinical settings, the interpretability of deep learning models directly impacts whether medical institutions and clinicians are willing to trust the model's predictions. Our proposed CoTEL-D3X method has certain advantages in terms of interpretability. Specifically, the CoT module we designed endows the model with problem-solving capabilities similar to human reasoning

Table 9
DDI Extraction Results of CoTEL-D3X on the DrugBank Database.

Description	Triplet	DDI found	DIC Severity
The risk or severity of CNS depression can be increased when **Acetazolamide** is combined with **Lormetazepam**.	<Acetazolamide, effect, Lormetazepam>	✓	moderate
Pralatrexate may decrease the excretion rate of **Iobenguane sulfate I-123** which could result in a higher serum level.	<Pralatrexate, mechanism, Iobenguane sulfate I-123>	✓	minor
The therapeutic efficacy of **Nomegestrol acetate** can be decreased when used in combination with **Phenoxymethylpenicillin**.	<Nomegestrol acetate, effect, Phenoxymethylpenicillin>	✓	minor
Dabigatran etexilate may decrease the excretion rate of **Pentostatin** which could result in a higher serum level.	<Dabigatran etexilate, mechanism, Pentostatin>	✓	moderate
Edetate calcium disodium anhydrous may decrease effectiveness of **Technetium Tc-99 m oxidronate** as a diagnostic agent.	<Edetate calcium disodium, effect, technetium tc-99 m oxidronate>	×	major

(Wei et al., 2022), allowing for the visualization of the model's reasoning path and uncovering the causal relationships between inputs and outputs. The instruction tuning module fine-tunes and solidifies the model's parameters, aligning the output with the DDI extraction task's content and structure requirements. Moreover, in Transformer-based generative LLMs, the core mask attention mechanism has strong representational and interpretative properties (Shi et al., 2022). During computation, the model mimics the human behavior of focusing on key terms when reading text, enabling it to better attend to words related to DDI in the text and learn the probability distribution of text generation based on this.

6.2. Theoretical and practical implications

Our proposed methods displayed improvements of 3.05 % and 5.86 % over the current best TP-DDI model. They do not require manual features or the definition of extraction rules. In particular, our joint method enables the direct extraction of triplet outputs in a single training step. Furthermore, our method extends the generated sequence length to 1,024, which is twice that of prevailing SOTA methods, enabling it to cover nearly all entities.

Regarding our pipeline method, although it needs to train separate models for entity recognition and relation classification, its end-to-end design allows the results of the first step to be directly applied to the second step without any manual intervention. This design also confers domain-agnostic characteristics to the method, rendering it highly transferable and easily extensible to triplet extraction tasks in other languages and domains.

Regarding the broader implications of our research, from a societal perspective, our method could help predict more potential DDIs from literature and clinical texts, providing valuable insights for the biomedical information extraction community. From a clinical perspective, our approach has the potential to reduce the risk of DDIs during patient treatment, thus helping to prevent unnecessary harm to health. Since the model can be deployed on a local computer for both training and inference, it mitigates the risk of patient data and model leakage compared to closed-source models.

We also observed that not only did our proposed pipeline method outperform our joint method but that pipeline methods in baseline models generally outperform joint methods. In validation studies by Yang et al. (2023a) and Zhong and Chen (2021), using separate encoders was found to yield better results than simply sharing a single encoder. This finding suggests that designing appropriate pipeline methods and leveraging the ability of LLMs to focus on solving subtasks can compensate for performance loss due to error propagation and may even achieve better results.

6.3. Limitations

Manually defined extraction rules are typically crafted by domain experts, offering high reliability and interpretability. By summarizing and refining core vocabulary, syntactic, and grammatical features of DDIs, certain relationships that may not be detected by LLMs could be identified. In the current phase of our research, we did not incorporate manually defined rules and patterns, which may lead the model to generate hallucinated outputs in specific texts. The integration of human-expert-defined rules with the powerful feature-learning capability of LLMs might enhance DDI extraction abilities. Therefore, in the future, we plan to explore how to design a rational framework for incorporating human-defined rules and investigate their impact on model performance.

Introducing external domain knowledge is one of the main strategies to address the lack of vertical domain knowledge in general LLMs. Most existing studies use external knowledge sources like domain-specific knowledge bases and knowledge graphs to facilitate the transfer of general LLMs to specialized fields like biomedicine. In our current research, we focused on model training and inference solely based on task-specific datasets, considering model transferability and flexibility. However, introducing biomedical external knowledge during instruction tuning and text generation could improve the model's ability to handle domain-specific texts. In the future, we will also explore how to integrate external knowledge while maintaining model transferability, to further enhance domain knowledge extraction.

Moreover, deep learning models, especially large language models, are often criticized for their "black-box" nature due to their high model complexity, and may produce factually incorrect outputs in certain cases. Our research mitigates this issue by incorporating CoT and restructuring the Instruction Tuning dataset. However, model interpretability still requires improvement. Investigating how to reveal the relationships between the model's internal parameters and features will be an important direction for our future research.

7. Conclusion

In this study, we introduced a Chain-of-Thought Enhanced Large Language Model for DDI Triplet eXtraction (CoTEL-D3X). To address the limitations of current mainstream modeling approaches, we developed both joint and pipeline methods and investigated their performance in DDI triplet extraction. For the joint method, we implemented CoT-enhanced Instruction Tuning, incorporating an intermediate step of entity extraction. Generative LLMs implicitly follow a logical reasoning process during inference, simplifying the task by initially extracting entities and subsequently generating the interactions between them. Furthermore, full-process joint modeling of entities and relations from encoding to decoding effectively resolves the issues of overlapping relations and discontinuous entities. The pipeline method decomposes the original task into two consecutive subtasks: entity extraction and entity pair relation classification, thereby explicitly employing CoT reasoning. The instructions tailored for different subtasks significantly enhance the task-comprehension capabilities of generative LLMs, thus improving DDI triplet extraction performance without the introduction of external knowledge or manual rules. Compared with current mainstream generative and extraction methods, both CoTEL-D3X approaches achieved SOTA results, underscoring their superiority in DDI triplet extraction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2B5B02002359), South Korea; China Scholarship Council (CSC) (No. 202206190144), China. The authors would like to thank the reviewers for their insightful comments and useful suggestions.

Data availability

The DDI corpus is available at: https://doi.org/10.1016/j.jbi.2013.07.011; The TAC 2018 corpus is available at: https://bionlp.nlm.nih.gov/tac2018druginteractions/; The ADR corpus is available at: https://bionlp.nlm.nih.gov/tac2017adversereactions/.

References

Ai, X., & Kavuluru, R. (2023). End-to-End models for chemical–protein interaction extraction: better tokenization and span-based pipeline strategies. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)* (pp. 610–618). IEEE.

Al-Sabri, R., Gao, J., Chen, J., Oloulade, B. M., & Lyu, T. (2023). Multi-view graph neural architecture search for biomedical entity and relation extraction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20*(02), 1221–1233.

Asada, M., Miwa, M., & Sasaki, Y. (2023). Integrating heterogeneous knowledge graphs into drug–drug interaction extraction from the literature. *Bioinformatics, 39*(1), Article btac754.

Baruah, G., & Kolla, M. (2018). KlickLabs at the TAC 2018 Drug-drug Interaction Extraction from Drug Labels Track. In *TAC*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv*, *2108*, 07258.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Cabot, P. L. H., & Navigli, R. (2021). REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2370–2381). Association for Computational Linguistics.

Dandala, B., Mahajan, D., & Devarakonda, M. V. (2017). IBM Research System at TAC 2017: Adverse Drug Reactions Extraction from Drug Labels. In *TAC*.

Demner-Fushman, D., Fung, K. W., Do, P., Boyce, R. D., & Goodwin, T. R. (2018). Overview of the TAC 2018 Drug-Drug Interaction Extraction from Drug Labels Track. In *TAC*.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv, 2305*, 14314.

Dong, Y., Cordonnier, J. B., & Loukas, A. (2021). Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning* (pp. 2793–2803). Proceedings of Machine Learning Research.

El-Allaly, E. D., Sarrouti, M., En-Nahnahi, N., & El Alaoui, S. O. (2022). An attentive joint model with transformer-based weighted graph convolutional network for extracting adverse drug event relation. *Journal of Biomedical Informatics, 125*, Article 103968.

Fatehifar, M., & Karshenas, H. (2021). Drug-drug interaction extraction using a position and similarity fusion-based attention mechanism. *Journal of Biomedical Informatics, 115*, Article 103707.

Fei, H., Zhang, Y., Ren, Y., & Ji, D. (2021). A span-graph neural model for overlapping entity relation extraction in biomedical texts. *Bioinformatics, 37*(11), 1581–1589.

Gajendran, S., Manjula, D., Sugumaran, V., & Hema, R. (2023). Extraction of knowledge graph of Covid-19 through mining of unstructured biomedical corpora. *Computational Biology and Chemistry, 102*, Article 107808.

Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., & Declerck, T. (2013). The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of Biomedical Informatics, 46*(5), 914–920.

Kim, S., Joo, S. J., Kim, D., Jang, J., Ye, S., Shin, J., & Seo, M. (2023). The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv*, *2305*, 14045.

Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., et al. (2024). DrugBank 6.0: The DrugBank knowledgebase for 2024. *Nucleic Acids Research, 52* (D1), D1265–D1275.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35*, 22199–22213.

Li, Z., Wei, Q., Huang, L. C., Li, J., Hu, Y., Chuang, Y. S., et al. (2024). Ensemble pretrained language models to extract biomedical knowledge from literature. *Journal of the American Medical Informatics Association, ocae061*.

Luo, L., Yang, Z., Cao, M., Wang, L., Zhang, Y., & Lin, H. (2020). A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of Biomedical Informatics, 103*, Article 103384.

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., et al. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics, 23*(6), Article bbac409.

Machado, J., Rodrigues, C., Sousa, R., & Gomes, L. M. (2023). Drug–drug interaction extraction-based system: An natural language processing approach. *Expert Systems, e13303*.

Martınez, J. L., Martınez, P., Segura-Bedmar, I., Carruana, A., Naderi, A., & Polo, C. (2017). MC-UC3M participation at TAC 2017 adverse drug reaction extraction from drug labels. In *TAC*.

Mazumdar, H., Chakraborty, C., Sathvik, M. S. V. P. J., & Panigrahi, P. K. (2023). GPTFX: A Novel GPT-3 based framework for mental health detection and explanations. *IEEE Journal of Biomedical and Health Informatics*. https://doi.org/10.1109/JBHI.2023.3328350

Modi, S., Kasmiran, K. A., Sharef, N. M., & Sharum, M. Y. (2024). Extracting adverse drug events from clinical Notes: A systematic review of approaches used. *Journal of Biomedical Informatics, 104603*.

Park, Y. J., Lee, M. A., Yang, G. J., Park, S. J., & Sohn, C. B. (2023). Web Interface of NER and RE with BERT for Biomedical Text Mining. *Applied Sciences, 13*(8), 5163.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., et al. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv*, *2112*, 11446.

Roberts, K., Demner-Fushman, D., & Tonning, J. M. (2017). Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. In *TAC*.

Segura-Bedmar, I., Martínez Fernández, P., & Herrero Zazo, M. (2013). Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 341–350). Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv*, *1508*, 07909.

Shazeer, N. (2020). Glu variants improve transformer. *arXiv preprint arXiv*, *2002*, 05202.

Shi, Y., Quan, P., Zhang, T., & Niu, L. (2022). DREAM: Drug-drug interaction extraction with enhanced dependency graph and attention mechanism. *Methods, 203*, 152–159.

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2021). Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv*, *2104*, 09864.

Suárez-Paniagua, V., Zavala, R. M. R., Segura-Bedmar, I., & Martínez, P. (2019). A two-stage deep learning approach for extracting entities and relations from medical texts. *Journal of Biomedical Informatics, 99*, Article 103285.

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2022). Mrc4bioer: Joint extraction of biomedical entities and relations in the machine reading comprehension framework. *Journal of Biomedical Informatics, 125*, Article 103956.

Sun, Z., Xing, L., Zhang, L., Cai, H., & Guo, M. (2023). Joint biomedical entity and relation extraction based on feature filter table labeling. *IEEE Access, 11*, 127422–127430.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv*, *2302*, 13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv*, *2307*, 09288.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, S., Zhang, Y., Che, W., & Liu, T. (2018). Joint extraction of entities and relations based on a novel graph scheme. In *IJCAI* (pp. 4461–4467). AAAI Press.

Wang, Y., Wang, Y., Sun, Z., Li, Y., Hu, S., & Ye, Y. (2023). Deep purified feature mining model for joint named entity recognition and relation extraction. *Information Processing & Management, 60*(6), Article 103511.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., et al. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv*, *2109*, 01652.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Wei, Z., Su, J., Wang, Y., Tian, Y., & Chang, Y. (2020). A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1476–1488). Association for Computational Linguistics.

Xiong, W., Li, F., Yu, H., & Ji, D. (2019). Extracting drug-drug interactions with a dependency-based graph convolution neural network. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 755–759). Institute of Electrical and Electronics Engineers.

Yang, C., Deng, J., Chen, X., & An, Y. (2023). SPBERE: Boosting span-based pipeline biomedical entity and relation extraction via entity information. *Journal of Biomedical Informatics, 145*, Article 104456.

Yang, J., Ding, Y., Long, S., Poon, J., & Han, S. C. (2023). DDI-MuG: Multi-aspect graphs for drug-drug interaction extraction. *Frontiers in Digital Health, 5*, Article 1154133.

Zaikis, D., & Vlahavas, I. (2021). TP-DDI: Transformer-based pipeline for the extraction of drug-drug Interactions. *Artificial Intelligence in Medicine, 119*, Article 102153.

Zhang, B., & Sennrich, R. (2019). Root mean square layer normalization. *Advances in Neural Information Processing Systems, 32*.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv*, *2205*, 01068.

Zhang, T., Leng, J., & Liu, Y. (2020). Deep learning for drug–drug interaction extraction from the literature: A review. *Briefings in Bioinformatics, 21*(5), 1609–1627.

Zhang, Y., Peng, J., Cheng, B., Liu, Y., & Jiang, C. (2024). MMR: A Multi-view Merge Representation model for Chemical-Disease relation extraction. *Computational Biology and Chemistry, 110*, Article 108063.

Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., & Xu, B. (2017). Joint extraction of entities and relations based on a novel tagging scheme. *arXiv preprint arXiv*, *1706*, 05075.

Zhong, Z., & Chen, D. (2021). 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* (pp. 50–61). Association for Computational Linguistics.

Zhu, Y., Li, L., Lu, H., Zhou, A., & Qin, X. (2020). Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *Journal of Biomedical informatics, 106*, Article 103451.