End-to-end Learning of Logical Rules for Enhancing Document-level Relation Extraction

Anonymous ACL submission

Abstract

Document-level relation extraction (DocRE) aims to extract relations between entities in a whole document. One of the pivotal challenges of DocRE is to capture the intricate interdependencies between relations of entity pairs. Previous methods have shown that logical rules are able to explicitly help capture such interdependencies. These methods either learn logical rules to refine the output of a trained DocRE model, or first learn logical rules from annotated data and then inject the learnt rules to a DocRE model using auxiliary training objective. In this paper, we argue that these learning pipelines may suffer from the issue of error propagation. To mitigate this issue, we propose *Joint Modeling Relation extraction and Logical rules* or *JMRL* for short, a novel rule-based framework that jointly learns both a DocRE model and logical rules in an end-to-end fashion. Specifically, we parameterize a rule reasoning module in JMRL to simulate the inference of logical rules, thereby explicitly modeling the reasoning process. We also introduce an auxiliary loss and a residual connection mechanism in JMRL to better reconcile the DocRE model and the rule reasoning module. Experimental results on four benchmark datasets demonstrate that the proposed JMRL framework is consistently superior to existing rule-based frameworks on all datasets, improving five baseline models for DocRE by a significant margin.

1 Introduction

Relation extraction (RE) plays a vital role in *information extraction* (IE). It aims at identifying relations between two entities in a given text. Early efforts focus mainly on sentence-level RE. In recent years, *document-level relation extraction* (DocRE) has received increasing attention. It aims at identifying relations of all entity pairs in a document. Nowadays, DocRE has been widely applied in downstream applications such as question answering (QA) (Sorokin and Gurevych, 2017), knowledge graph construction (Luan et al., 2018), etc. Compared to sentence-level RE, DocRE imposes a greater challenge for modeling longer contexts and capturing the more complex interdependencies between entity pairs.

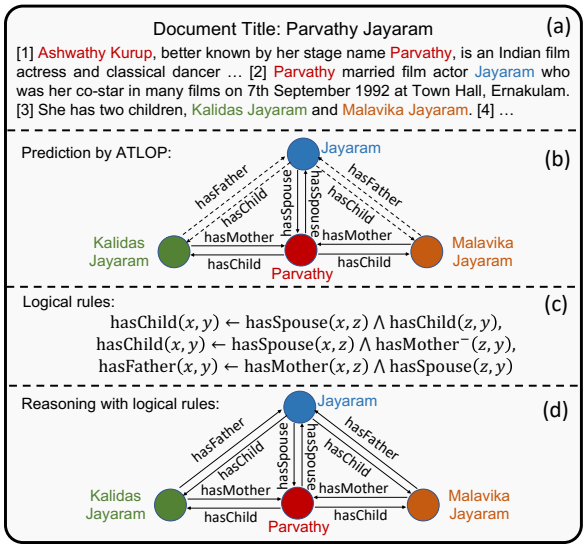

Figure 1: Examples in the DocRED dataset, where solid arrows denote the correct predictions, dotted arrows the missing predictions and r^- the inverse relation of r.

Most previous methods for DocRE focus on capturing interdependencies between entity pairs by learning powerful representations through neural models, such as pre-trained language models (Xu et al., 2021; Zhou et al., 2021a), or graph neural networks (Peng et al., 2017; Sahu et al., 2019; Zeng et al., 2020). However, these methods are usually prone to lose the reasoning ability. Figure 1 illustrates such an example, where sub-figure (a) in Figure 1 shows an example of a document in the DocRED dataset, and sub-figure (b) shows the corresponding predictions yielded by ATLOP, a state-of-the-art (SOTA) method for DocRE. We can observe that ATLOP (Zhou et al., 2021a) only extracts apparent facts such as "$(\text{Parvathy}, \text{hasSpouse}, \text{Jayaram})$"

and "(Parvathy, hasChild, KalidasJayaram)", but fails to identify potential facts such as "(KalidasJayaram, hasFather, Jayaram)" and "(Jayaram, hasChild, MalavikaJayaram)" since they are not explicitly mentioned in the document.

In general, logical rules can be used to improve the reasoning ability for DocRE by inferring missing facts from existing ones. Sub-figure (c) in Figure 1 illustrates three logical rules, and sub-figure (d) shows their ability in inferring missing facts. To enhance existing DocRE models with logical rules, two rule-based frameworks have been proposed, namely LogicRE (Ru et al., 2021) and MILR (Fan et al., 2022). In more details, LogicRE first learns logical rules based on the output logits of a trained neural model and then refines its predicted relations by reasoning with the learnt rules, whereas MILR first learns logical rules from annotated data and then trains a neural model penalized by an auxiliary loss for reflecting the violation of learnt rules. Although both LogicRE and MILR have shown promising results in enhancing performance for DocRE, they still suffer from the error propagation issue due to their pipeline natures.

In this paper, we target jointly learning a neural module for DocRE and a neural module for approximating logical rules in an end-to-end fashion to avoid error propagation. To this end, we propose a novel framework named *Joint Modeling Relation extraction and Logical rules* or *JMRL* for short, as illustrated in Figure 2. The intuition of JMRL is to reduce the rule learning problem in discrete space to a parameter learning problem in continuous space, yielding a neural module for approximating logical rules (called a rule reasoning module) and then integrating it into an existing DocRE model. The parameters of the rule reasoning module is tuned along with the parameters of the backbone DocRE model so that the whole model can be trained in an end-to-end fashion. Furthermore, we introduce an auxiliary loss and a residual connection mechanism in JMRL to better incorporate the backbone DocRE model and the rule reasoning module, so as to further improve the performance.

We impose JMRL to enhance five baseline models for DocRE, including LSTM (Yao et al., 2019), Bi-LSTM (Yao et al., 2019), GAIN (Zeng et al., 2020), ATLOP (Zhou et al., 2021a) and DREEAM (Ma et al., 2023a). Experimental results on four benchmark datasets DWIE (Zaporojets et al., 2021), DocRED (Yao et al., 2019), Re-DocRED (Tan et al., 2022b), and DocGNRE (Li et al., 2023) demonstrate that the proposed JMRL framework is superior to all SOTA rule-based framework for DocRE, improving the baseline models by a significant margin on all datasets. Our analysis and case study further clarify why JMRL is able to improve the performance.

The main contributions of this work include:

(1) We propose a novel framework named JMRL to integrate a neural module for approximating logical rules into a baseline DocRE model, so that the enhanced DocRE model can be trained in an end-to-end fashion. As far as we know, this is the first end-to-end approach for imposing logical rules upon DocRE models.

(2) We theoretically analyze the faithfulness between the rule reasoning module and logical rules.

(3) We conduct extensive experiments on four benchmark datasets, demonstrating that the proposed JMRL framework pushes forward five baseline SOTA DocRE models by a significant margin. In particular, up to the submission date (2024/02/15), the JMRL-enhanced DREEAM model (submissions under the username jmrl) ranks the first in the public DocRED evaluation[1].

2 Preliminaries

Problem formulation for DocRE. Given a document d involving a set of named entities $\mathcal{E}_d = \{e_i\}_{1 \le i \le n_e}$, the task of DocRE aims at predicting the relations among all entity pairs $\{(e_h, e_t) \mid e_h, e_t \in \mathcal{E}_d, e_h \neq e_t\}$. The set of predictable relations is defined as $\mathcal{R}_+ = \mathcal{R} \cup \{\perp\}$, where $\mathcal{R}$ is a pre-defined relation set and $\perp$ the "no relation".

Atoms and facts. An *atom* is of the form $r(x, y)$, where $r \in \mathcal{R}$ is a *predicate*, x and y are entity variables or entity constants. An atom is *ground* if it does not contain any variable. A *fact* is a ground atom of the form $r(a, b)$, which is also expressed as a triple (a, r, b) throughout the paper.

Logical rules. We focus on chain-like logical rules (CRs). A CR is a datalog rule (Abiteboul et al., 1995) where all atoms are binary and every body atom shares variables with the previous atom and the next atom. A CR is called an L-CR if it has L body atoms. An L-CR R is of the form:

$$H(x, y) \leftarrow B_1(x, z_1) \wedge B_2(z_1, z_2) \wedge ... \wedge B_L(z_{L-1}, y)$$

where x the head entity, y the tail entity, and z_1, ..., z_{L-1} variables. The part at the left (resp. right)

[1] https://codalab.lisn.upsaclay.fr/competitions/365

side of $\leftarrow$ is called the *head* (resp. *body*) of R. The rule R is called r-specific if $H = r$. By H_R and B_R we denote the atom in the head of R and the set of atoms in the body of R, respectively. A rule is ground if it does not contain any variable. A rule R is a fact if B_R is empty and H_R is ground. To uniformly represent CRs with fixed-length bodies, we introduce the *identity relation* (denoted by I) to rule bodies. For example, the 1-CR $r(x, y) \leftarrow p(x, y)$ can be converted into a 2-CR $r(x, y) \leftarrow p(x, z) \wedge I(z, y)$.

Given a set of facts $\mathcal{G} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, we denote by $\mathcal{G} \models H_R(a, b)$ if there exists a ground instance R_g of logical rule R such that $H_R(a, b) = H_{R_g}$ and $B_{R_g} \subseteq \mathcal{G} \cup \mathcal{G}^- \cup \{I(e, e) \mid e \in \mathcal{E}\}$, where $\mathcal{G}^- = \{(e_t, r^-, e_h) \mid (e_h, r, e_t) \in \mathcal{G}\}$ and r^- denotes the inverse relation of r. Let Σ be a set of r-specific CRs and $(a, r, b) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ an arbitrary fact. We denote by $\mathcal{G} \models_\Sigma (a, r, b)$ if there exits a logical rule $R \in \Sigma$ such that $\mathcal{G} \models H_R(a, b)$.

3 Related Work

Document-level relation extraction. Early efforts for DocRE focus on better contextualized representations of relations by employing various technologies such as attention mechanisms (Yao et al., 2019; Zhou et al., 2021a), pre-trained language models (Tang et al., 2020; Xu et al., 2021), and knowledge distillation (Tan et al., 2022a; Ma et al., 2023a). To capture more complex interdependencies between entity pairs, recent studies aim at enhancing DocRE models with external modules such as graph neural networks (GNNs) (Christopoulou et al., 2019; Zhang et al., 2020; Zeng et al., 2020) or rule-based frameworks (Ru et al., 2021; Fan et al., 2022). Specifically, LogicRE (Ru et al., 2021) and MILR (Fan et al., 2022) are two SOTA rule-based frameworks for enhancing DocRE. LogicRE first learns logical rules based on the output logits of a trained neural model and then refines the predicted relations of the neural model by learnt rules. MILR first learns logical rules from annotated data and then trains a neural model penalized by an auxiliary loss for reflecting the violation of learnt rules. However, the above two frameworks suffer from the error propagation issue due to their pipeline natures. In contrast, our proposed JMRL framework integrates a neural module for rule reasoning into a backbone DocRE model, enabling the whole model to be trained end-to-end and thus mitigating the error propagation issue.

End-to-end rule learning. In recent years, there is an emerging interest in exploiting neural-based methods (Yang et al., 2017; Sadeghian et al., 2019; Yang and Song, 2020; Xu et al., 2022) for end-to-end rule learning. Inspired by their promising results, we also design a neural-based rule reasoning module in JMRL to approximate logical rules for DocRE. Different from previous methods, our approach can handle the training objective of relation extraction, whereas previous methods are only designed for specific tasks in knowledge graph completion such as link prediction (Bordes et al., 2013) and triple classification (Lin et al., 2015). Furthermore, our approach can deal with the reasoning scenario where existing facts in the background knowledge are all uncertain (i.e., the existing facts are predicted by a DocRE model with continuous values for their truth degrees).

Rule injection in neural models. There exist approaches focusing on injecting logical rules into neural models in different tasks of natural language processing (NLP), including knowledge base construction (Demeester et al., 2016; Ding et al., 2018), natural language inference (Li and Srikumar, 2019), sentiment analysis (Deng and Wiebe, 2015), knowledge graph validation (Du et al., 2019) and information extraction (Wang and Pan, 2020; Zhou et al., 2021b). These approaches require well-prepared hand-crafted rules as input for the enhancement, which may prevent them from being practically used. In contrast, our proposed JMRL framework does not require hand-crafted rules as input.

4 The JMRL Framework

To impose logical rules upon a DocRE model, we propose a novel rule-based framework named *Joint Modeling Relation extraction and Logical rules* or *JMRL* for short, as illustrated in Figure 2. By and large, JMRL first employs a DocRE model to calculate output logits for all potential facts in a document, and then feeds them into a rule reasoning module to produce the rule-enhanced logits. The ultimately predicted logits are calculated by the residual connection of the original DocRE logits and the rule-enhanced logits. Then the entire model is trained by minimizing a weighted sum of classification losses calculated from the original DocRE logits and the ultimately predicted logits. Furthermore, we can extract logical rules from the parameter assignment of the rule reasoning module to compose explanations for the predictions.

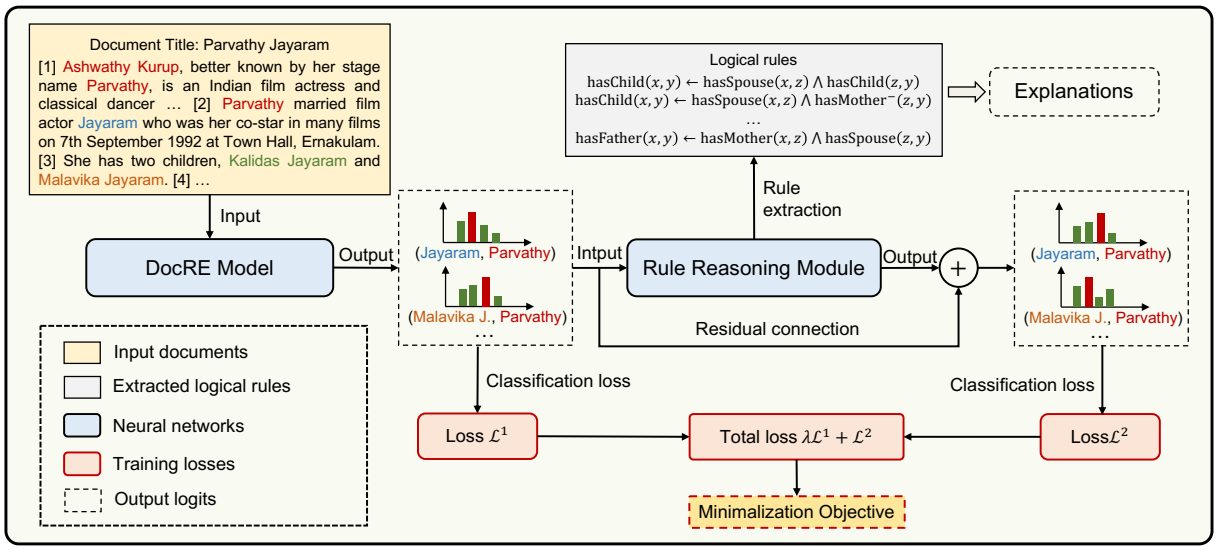

Figure 2: The overview of the proposed JMRL framework.

4.1 Document-level Relation Extraction

Given a document d involving a set of named entities $\mathcal{E}_d = \{e_i\}_{1 \le i \le n_e}$, a typical DocRE model $\mathcal{F}$ calculates a logit $\mathcal{F}(e_h, e_t, d) \in \mathbb{R}^{n+1}$ for each entity pair in $\{(e_h, e_t) \mid e_h, e_t \in \mathcal{E}_d, e_h \neq e_t\}$, where $n = |\mathcal{R}|$, $[\mathcal{F}(e_h, e_t, d)]_i$ denotes the logit for a normal relation for all $1 \le i \le n$, and $[\mathcal{F}(e_h, e_t, d)]_{n+1}$ denotes the logit for $\perp$.

A DocRE model is usually trained by minimizing the binary cross-entropy (BCE) loss (Yao et al., 2019; Zeng et al., 2020) or the adaptive thresholding (AT) loss (Zhou et al., 2021a), a variant of cross-entropy. In the inference phase, the set of predicted facts $\{(e_h, r, e_t) \mid [\sigma(\mathcal{F}(e_h, e_t, d))]_r > \epsilon\}$ are obtained by thresholding the predicted probabilities of each entity pair, where ϵ is a given threshold, σ is an activation function such as the sigmoid function or the softmax function.

4.2 The Rule Reasoning Module

The rule reasoning module is a neural module parameterized to simulate the inference of logical rules, approximating outputs as a rule system does. This module is trained along with the DocRE model to optimize a certain training objective.

Let N be the maximum number of rules to be learnt, L the maximum number of atoms in each rule and $\mathcal{R}_* = \mathcal{R} \cup \mathcal{R}^- \cup \{I\}$. Suppose $\mathcal{R} = \{r_i\}_{1 \le i \le n}$, its corresponding set of inverse relations $\mathcal{R}^- = \{r_i\}_{n+1 \le i \le 2n}$, and $I = r_{2n+1}$. We define an extended logit $\mathcal{F}_+(x, y, d) \in \mathbb{R}^{2n+1}$, where $[\mathcal{F}_+(x, y, d)]_i = [\sigma(\mathcal{F}(x, y, d))]_i$ for all $1 \le i \le n$, $[\mathcal{F}_+(x, y, d)]_{i+n} = [\sigma(\mathcal{F}(y, x, d))]_i$ for all $1 \le i \le n$, and $[\mathcal{F}_+(x, y, d)]_{2n+1} = 1$ if $x = y$ or 0 otherwise. The goal of our rule reasoning module is to estimate a truth degree $s^{(N,L)}_{r,x,y,d}$ for every fact $(x, r, y) \in \mathcal{E}_d \times \mathcal{R}_* \times \mathcal{E}_d$ in every document d, where the estimated truth degree $s^{(N,L)}_{r,x,y,d}$ reflects the degree of whether the fact (x, r, y) can be inferred by N L-CRs. For every normal relation $r \in \mathcal{R}, 1 \le k \le N, 1 \le l \le L$, the intermediate estimated truth degree $s^{(k,l)}_{r,x,y,d}$ for the l^{th} atom in the k^{th} rule is defined as:

$$s^{(k,l)}_{r,x,y,d} = \begin{cases} \sum_{i=1}^{2n+1} w_i^{(r,k,l)} [\mathcal{F}_+(x, y, d)]_i, & l = 1 \\ \sum_{i=1}^{2n+1} w_i^{(r,k,l)} \sum_{\substack{(z, r_i, y) \in \\ \mathcal{E}_d \times \mathcal{R}_* \times \mathcal{E}_d}} s^{(k,l-1)}_{r,x,z,d} [\mathcal{F}_+(z, y, d)]_i, & l > 1 \end{cases} \tag{1}$$

where $w^{(r,k,l)} \in [0, 1]^{2n+1}$ denotes the trainable weights on predicate selection for the l^{th} body atom of the k^{th} rule whose head atom is on r. $w^{(r,k,l)}$ is confined to $[0, 1]$ by a softmax layer. Intuitively, $w_i^{(r,k,l)} = 1$ indicates that the i^{th} relation r_i is selected as the predicate of the l^{th} body atom.

Different from normal relations in $\mathcal{R}$, for the head relation $\perp$, we allow $\perp$ and its reverse relation to appear in predicates of body atoms. To this end, we alter Equation (1) for $r = \perp$ by looping i from 1 to $2n + 3$, redefining $w^{(r,k,l)} \in [0, 1]^{2n+3}$, $\mathcal{R}_* = \mathcal{R} \cup \{\perp\} \cup \mathcal{R}^- \cup \{\perp^-, I\}$, $[\mathcal{F}_+(x, y, d)]_i = [\sigma(\mathcal{F}(x, y, d))]_i$ for all $1 \le i \le n + 1$, $[\mathcal{F}_+(x, y, d)]_{i+n+1} = [\sigma(\mathcal{F}(y, x, d))]_i$ for all $1 \le i \le n + 1$, and $[\mathcal{F}_+(x, y, d)]_{2n+3} = 1$ if $x = y$ or 0 otherwise.

The ultimate truth degree is calculated by aggre-

gating the intermediate degrees of N rules:

$$s^{(N,L)}_{r,x,y,d} = \sum_{k=1}^{N} \alpha^{(k)}_r s^{(k,L)}_{r,x,y,d} \quad (2)$$

where $\alpha^{(k)}_r \in [-1, 1]$ is a trainable weight for the k^{th} rule for the head relation r, which is confined to $[-1, 1]$ by a tanh layer. Intuitively, $\alpha^{(k)}_r$ denotes the confidence score of the k^{th} rule for r.

By introducing the following notion of induced parameter assignment, we show in Theorem 1 that the formalization of the proposed rule reasoning module is faithful to a certain set of CRs.

Definition 1. *Given a set of r-specific L-CRs $\Sigma = \{R_k\}_{1\le k\le N}$ for R_k of the form $r(x, y) \leftarrow p_{k,1}(x, z_1) \wedge \cdots \wedge p_{k,L}(z_{L-1}, y)$, where $p_{k,l} \in \mathcal{R} \cup \{\perp\} \cup \mathcal{R}^- \cup \{\perp^-, I\}$ if $r = \perp$, or $p_{k,l} \in \mathcal{R} \cup \mathcal{R}^- \cup \{I\}$ otherwise, we call a parameter assignment of the rule reasoning module $\theta^{(N,L)}_r = \{w^{(r,k,l)}_i\}_{1\le k\le N, 1\le l\le L, 1\le i\le m} \cup \{\alpha^{(k)}_r\}_{1\le k\le N}$ Σ-induced if it satisfies the following conditions:*

(1) $\forall 1 \le k \le N, 1 \le l \le L, 1 \le i \le m$: $w^{(r,k,l)}_i = 1$ if $p_{k,l} = r_i$ or $w^{(r,k,l)}_i = 0$ otherwise, where $m = 2n + 3$ if $r = \perp$ or $m = 2n + 1$ otherwise.

(2) $\forall 1 \le k \le N, 1 \le l \le L : \alpha^{(k)}_r = 1$.

Theorem 1. *Suppose $[\sigma(\mathcal{F}(x, y, d))]_r = 1$ if the fact (x, r, y) is predicted to be true in document d, or $[\sigma(\mathcal{F}(x, y, d))]_r = 0$ otherwise. Let $\mathcal{R}_\dagger = \mathcal{R}_+$ if $r = \perp$ or $\mathcal{R}_\dagger = \mathcal{R}$ otherwise, $\mathcal{G}_d = \{(x, r, y) \in \mathcal{E}_d \times \mathcal{R}_\dagger \times \mathcal{E}_d \mid [\sigma(\mathcal{F}(x, y, d))]_r = 1\}$ be the set of predicted true facts for d, $\Sigma = \{R_k\}_{1\le k\le N}$ a set of r-specific L-CRs and $\theta^{(N,L)}_r$ the Σ-induced parameter assignment of the rule reasoning module. Then for any fact $(a, r, b) \in \mathcal{E}_d \times \mathcal{R}_\dagger \times \mathcal{E}_d$, $s^{(N,L)}_{r,a,b,d} \ge 1$ if and only if $\mathcal{G}_d \models_\Sigma (a, r, b)$.*

The proof of Theorem 1 is provided in Appendix A. Theorem 1 enables us to extract explainable logical rules from the parameter assignment of the learnt neural module. The rule extraction algorithm is shown in Appendix B.

Residual connection. Considering that there exist DocRE scenarios where logical reasoning is useless, we introduce the well-known residual connection mechanism to incorporate the output logits from the original DocRE model and the estimated truth degrees from the rule reasoning module. The ultimately predicted logit is calculated by:

$$\phi^{(x,y,d)}_r = [\mathcal{F}(x, y, d)]_r + s^{(N,L)}_{r,x,y,d} \quad (3)$$

Dataset	Split	#Doc.	#Rel.	#Ent.	#Facts.
DWIE	train	602	65	16,494	14,403
	dev	98		2,785	2,624
	test	99		2,623	2,495
DocRED	train	3,053	96	59,493	38,180
	dev	998		19,578	12,323
	test	1,000		19,539	-
	test†	500		9,779	17,448

Table 1: Statistics on datasets, where Doc. (resp. Rel or Ent) abbreviates documents (resp. relations or entities).

4.3 Training Objective

JMRL is trained by minimizing a classification loss (BCE or AT, inherited from the backbone DocRE model) calculated by $\phi^{(x,y,d)}_r$. The formal definitions of BCE and AT are given in Appendix C.

In practice, it is hard to accurately train the rule reasoning module at the early stage of training, as the facts predicted by the backbone DocRE model are inaccurate at the early stage. To tackle this issue, we introduce an auxiliary loss in JMRL to improve the efficiency of the entire training process. The classification loss on the output logits $\mathcal{F}(x, y, d)$ of the backbone DocRE model is treated as the auxiliary loss. By $\mathcal{L}^1_\Delta$ and $\mathcal{L}^2_\Delta$ we denote the auxiliary loss and the original loss, respectively, the entire JMRL-enhanced model is trained by minimizing $\lambda\mathcal{L}^1_\Delta + \mathcal{L}^2_\Delta$, where $\Delta \in \{\text{BCE}, \text{AT}\}$ and λ is a hyper-parameter to trade-off the two losses.

5 Evaluation

5.1 Experimental Setup

Datasets and metrics. We used the DWIE, DocRED, Re-DocRED, and DocGNRE datasets for evaluation, where the results on Re-DocRED and DocGNRE are moved to appendix. To fairly compare with MILR on DocRED, we used the same relabeled test set as Huang et al. (2022). Statistics for DWIE and DocRED are reported in Table 1, where test† denotes the relabeled test set. Following Yao et al. (2019), we used F1-score and Ign F1-score as evaluation metrics, where Ign F1-score extends F1-score by omitting facts appearing in the intersection of the training set and the dev (resp. test) set for evaluation on the dev (resp. test) set.

Baselines. To compare JMRL with the SOTA rule-based frameworks LogicRE (Ru et al., 2021) and MILR (Fan et al., 2022), we enhanced four baseline models, including LSTM (Yao et al., 2019), Bi-LSTM (Yao et al., 2019), GAIN (Zeng et al., 2020) and ATLOP (Zhou et al., 2021a). For a more

Method	PLM	Dev		Test		p-value
		Ign F1 (%)	F1 (%)	Ign F1 (%)	F1 (%)	
ChatGPT (5-shot) (Han et al., 2023)	ChatGPT	-	-	-	26.72	-
LSTM (Yao et al., 2019)	GloVe	31.71	38.35	31.65	41.42	2.5e-2
LogicRE-LSTM (Ru et al., 2021)	GloVe	32.02 (+0.31)	38.48 (+0.13)	32.58 (+0.93)	42.03 (+0.61)	2.2e-2
MILR-LSTM (Fan et al., 2022)	GloVe	33.12 (+1.41)	39.95 (+1.60)	33.75 (+2.10)	43.35 (+1.93)	3.9e-2
JMRL-LSTM (this work)	GloVe	36.11 (+5.40)	42.87 (+4.52)	43.16 (+11.51)	50.34 (+8.92)	-
BiLSTM (Yao et al., 2019)	GloVe	32.14	39.66	33.88	43.54	8.0e-3
LogicRE-BiLSTM (Ru et al., 2021)	GloVe	32.39 (+0.25)	40.32 (+0.66)	34.21 (+0.33)	43.95 (+0.45)	1.1e-2
MILR-BiLSTM (Fan et al., 2022)	GloVe	34.05 (+1.91)	41.22 (+1.56)	35.09 (+1.21)	44.65 (+1.11)	2.2e-2
JMRL-BiLSTM (this work)	GloVe	37.88 (+5.74)	43.68 (+4.02)	42.68 (+8.80)	50.70 (+7.16)	-
GAIN (Zeng et al., 2020)	$\text{BERT}_{\text{base}}$	58.89	63.81	61.36	67.45	1.8e-3
LogicRE-GAIN (Ru et al., 2021)	$\text{BERT}_{\text{base}}$	58.98 (+0.09)	64.90 (+1.09)	61.58 (+0.22)	68.71 (+1.26)	3.4e-2
MILR-GAIN (Fan et al., 2022)	$\text{BERT}_{\text{base}}$	61.22 (+2.33)	65.85 (+2.04)	62.77 (+1.41)	69.23 (+1.78)	1.5e-1
JMRL-GAIN (this work)	$\text{BERT}_{\text{base}}$	61.62 (+2.73)	66.03 (+2.22)	64.59 (+3.23)	69.66 (+2.21)	-
ATLOP (Zhou et al., 2021a)	$\text{BERT}_{\text{base}}$	63.37	69.87	67.29	75.13	4.0e-3
LogicRE-ATLOP (Ru et al., 2021)	$\text{BERT}_{\text{base}}$	64.54 (+1.17)	70.66 (+0.79)	68.13 (+0.84)	75.67 (+0.54)	3.5e-3
MILR-ATLOP (Fan et al., 2022)	$\text{BERT}_{\text{base}}$	67.18 (+3.81)	72.05 (+2.97)	69.84 (+2.55)	76.51 (+1.38)	3.9e-3
JMRL-ATLOP (this work)	$\text{BERT}_{\text{base}}$	**68.41 (+5.04)**	**73.91 (+4.04)**	**70.92 (+3.63)**	**77.85 (+2.72)**	-

Table 2: Comparison results on the DWIE dataset.

comprehensive comparison, we also applied JMRL to enhance the SOTA model DREEAM (Ma et al., 2023a) and compared with other SOTA methods SSAN (Xu et al., 2021) and KD-DocRE (Tan et al., 2022a). Note that these baseline models adopt different loss functions, where the BCE loss is used by LSTM, Bi-LSTM and GAIN, and the AT loss is used by ATLOP and DREEAM. We also compared JMRL with large language models (LLMs) such as ChatGPT (Han et al., 2023), GPT-4 (Peng et al., 2023) and FLAN-UL2 (Peng et al., 2023).

Implementation details. We implemented all JMRL-enhanced models by Pytorch 2.0.0 on an NVIDIA A100 GPU[2]. We utilized the public repositories of backbone models such as LSTM and Bi-LSTM[3], GAIN[4], ATLOP[5], and DREEAM[6] to implement our experiments. The hyper-parameter λ for JMRL is set to 1 in all experiments. We provide detailed hyper-parameter settings in Appendix E, where all hyper-parameters were tuned to maximize the Ign F1-score on the dev set.

5.2 Main Results

We use JMRL-X (resp. LogicRE-X or MILR-X) to denote the enhanced models, where X denotes an original DocRE model. Table 2 (resp. Table 3) reports the comparison results on the DWIE (resp. DocRED) dataset. where the results of baselines in Table 3 are sourced from (Fan et al., 2022). Results show that the proposed JMRL framework improves all original DocRE models by a significant margin in both F1-scores and Ign F1-scores with p-values < 0.05 by two-tailed t-tests. These results demonstrate a ubiquitous effectiveness of JMRL across a variety of backbone models which use different kinds of word embedding, language models and loss functions. Furthermore, we can observe that JMRL consistently outperforms both the SOTA rule-based frameworks LogicRE and MILR. Specifically, JMRL-ALTOP outperforms MILR-ATOP by a significant margin of 1.08% (resp. 3.02%) in Ign F1-score on the DWIE (resp. DocRED) dataset. This is in line with our expectation that a joint training framework (e.g. JMRL) is better than a pipeline framework (e.g. LogicRE and MILR) due to the mitigation of error propagation. From the results reported in Table 4 for comparing with the SOTA DocRE model DREEAM, we see that JMRL-DREEAM achieves new SOTA performance on DocRED, namely 67.91% (resp. 65.69%) in F1-score (resp. Ign F1-score). This improvement beyond

Method	Test (using test†)	
	Ign F1 (%)	F1 (%)
ChatGPT (5-shot)	-	28.89
GAIN	41.26	41.68
LogicRE-GAIN	41.53 (+0.27)	41.89 (+0.21)
MILR-GAIN	42.89 (+1.63)	43.17 (+1.49)
JMRL-GAIN	**47.85 (+6.59)**	**49.58 (+7.90)**
ATLOP	41.67	41.95
LogicRE-ATLOP	42.47 (+0.80)	42.73 (+0.78)
MILR-ATLOP	44.30 (+2.63)	44.72 (+2.77)
JMRL-ATLOP	47.32 (+5.65)	47.54(+5.59)

Table 3: Comparison results on the DocRED dataset.

[2]Code and data about our implementations are available at: [link removed during double blind reviewing]

[3]https://github.com/thunlp/DocRED

[4]https://github.com/DreamInvoker/GAIN

[5]https://github.com/wzhouad/ATLOP

[6]https://github.com/YoumiMa/dreeam

Method	PLM	Dev		Test (using test)		p-value
		Ign F1 (%)	F1 (%)	Ign F1 (%)	F1 (%)	
ChatGPT (5-shot) (Han et al., 2023)	ChatGPT	-	32.21	-	-	-
GPT-4 (2-shot) (Peng et al., 2023)	GPT-4	-	-	-	27.90	-
FLAN-UL2 (FT) (Peng et al., 2023)	FLAN-UL2 (20B)	-	-	-	54.50	-
SSAN (Xu et al., 2021)	RoBERTa$_{\text{large}}$	63.76	65.69	63.78	65.92	3.9e-6
KD-DocRE (Tan et al., 2022a)	RoBERTa$_{\text{large}}$	65.27	67.12	65.24	67.28	3.0e-3
DREEAM (Ma et al., 2023a)	RoBERTa$_{\text{large}}$	65.52	67.41	65.47	67.53	2.4e-2
JMRL-DREEAM (this work)	RoBERTa$_{\text{large}}$	**65.64**	**67.61**	**65.69**	**67.91**	-

Table 4: Comparison results on the original DocRED dataset.

Method	DWIE		DocRED		p-val.
	IF1	F1	IF1	F1	
JMRL-ATLOP	**70.92**	**77.85**	**47.32**	**47.54**	-
- residual connection	66.04	73.75	43.70	43.88	8.1e-4
- auxiliary loss	69.62	76.73	44.55	44.75	2.2e-2
Using NeuralLP	68.66	76.60	44.30	44.45	1.1e-2
Using DRUM	69.90	77.08	44.70	44.85	4.0e-2

Table 5: Ablation study on the DWIE test set and original DocRED test set, where p-val. abbreviates p-value.

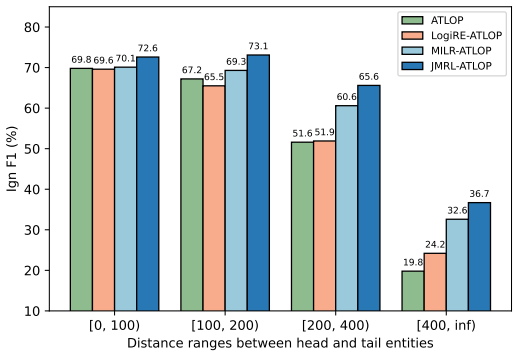

Figure 3: Comparison results for different distances.

SOTA is also statistically significant with a p-value < 0.05. This confirms that JMRL is able to further enhance SOTA DocRE models.

Besides, we also compared JMRL with LLMs, including ChatGPT, GPT-4 and FLAN-UL2 (FT). The comparison results reported in Table 2,3,4 show that LLMs achieve relatively lower performance on both DWIE and DocRED, even though they were fine-tuned on the training data. The reasons are two-fold. On one hand, LLMs like ChatGPT and GPT-4 can hardly make full use of the training data for adapting to a new task. On the other hand, LLMs are generative models that are too general to fit the DocRE task, which is a classification task, when compared with JMRL-enhanced models that are discriminative models. We provide more detailed discussions on LLMs in Appendix F.

5.3 Analysis

Ablation study. Table 5 reports our results for ablation study. In the first variant model, we omitted the residual connection mechanism in JMRL. Results show that the performance of this variant significantly drops compared to JMRL-ATLOP with a low p-value=8.1e-4 by a two-tailed t-test. In the second variant model, we omitted the auxiliary loss in JMRL. Results show that the use of auxiliary loss results in a significant performance gain with a p-value=2.2e-2. These results demonstrate the effectiveness of key components in JMRL. For the third and the fourth variant models, we respectively altered the rule reasoning module by the well-known end-to-end rule learning models NeuralLP (Yang et al., 2017) and DRUM (Sadeghian et al., 2019). Results show that JMRL-ATOP significantly outperforms these two variants with p-values < 0.05. The reason why our proposed rule reasoning module outperforms both NeuralLP and DRUM may lie in the fact that both NeuralLP and DRUM introduce an extra LSTM network to express the relevance of weights for predicate selection in adjacent body atom, while this extra component introduces more parameters that can hardly be optimized by noisy facts output from the backbone DocRE model.

Analysis on long-range dependencies. To verify whether logical rules are benefit for capturing long-range dependencies between entity mentions, we separate the set of entity pairs into four groups according to the distances between entity pairs, where the distance between two entities is measured by the minimum number of tokens between the mentions of these two entities in a document. Figure 3 shows the comparison results on the dev set of DWIE. We can see that JMRL-ATLOP consistently outperforms the baselines in all four groups. Moreover, the performance generally decreases with increasing distances. However, JMRL-ATLOP achieves better performance in the range $[100, 200)$ than in the range $[0, 100)$. These results imply that JMRL is more effective in capturing

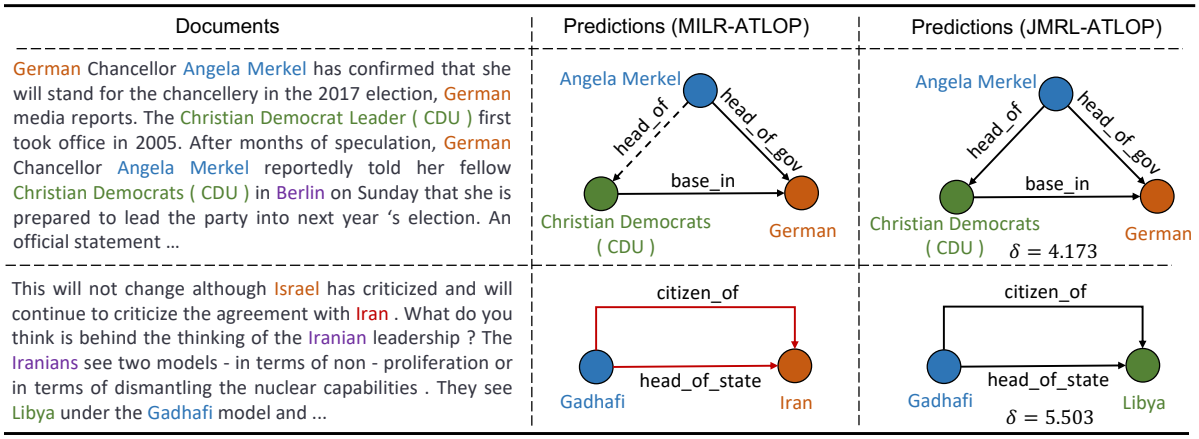

Figure 4: Case study for MILR-ATLOP and JMRL-ATLOP on the DWIE test set, where black solid lines denote true predictions, red lines denote false predictions, and dashed lines denote missing predictions.

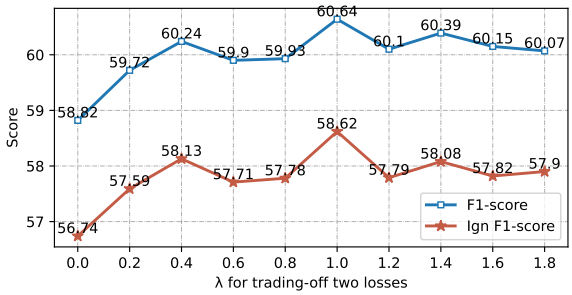

Figure 5: Analysis on the hyper-parameter λ.

long-range dependencies betwen entity mentions.

Analysis on the hyper-parameter λ. We conducted analysis on the hyper-parameter λ, where the experiments were conducted on the dev set of DocRED, based on JMRL-ATLOP. Figure 5 illustrates the comparison results. It can be observed that both F1-score and Ign F1-score only moderately fluctuate when λ ranges from 0 to 1.8, and that both of them reach the maximum when $\lambda = 1.0$. Therefore, we set $\lambda = 1.0$ in all our experiments.

Case study. We conducted case study for comparing MILR-ATLOP with JMRL-ATLOP on the DWIE test set, as shown in Figure 4. We first introduce a metric δ to estimate, in the residual connection, the ratio of the degree that the rule-enhance logit dominates the ultimately predicted logit to the degree that the DocRE logit dominates the ultimately predicted logit; formally, $\delta = \mathrm{dis}(v_{\mathrm{ori}}, v_{\mathrm{ori}} + v_{\mathrm{rule}})/\mathrm{dis}(v_{\mathrm{rule}}, v_{\mathrm{ori}} + v_{\mathrm{rule}})$, where v_{ori} and v_{rule} denote the DocRE logit and the rule-enhanced logit, respectively, and dis is the Euclidean distance function. In the first case, MILP-ATLOP fails to predict the true relation "head_of" between "Angela Merkel" and "Christian Democrats (CDU)", whereas JMRL-ATLOP predicts this true relation. The correct prediction of JMRL-ATLOP can be explained by a rule "$\mathrm{head\_of}(x, y) \leftarrow \mathrm{head\_of\_gov}(x, z) \wedge \mathrm{base\_in}^{-}(z, y)$" extracted from the parameter assignment of the rule reasoning module, while MILP-ATLOP fails to discover this rule. In the second case, MILP-ATLOP predicts two false relations between "Gadhafi" and "Iran", whereas JMRL-ATLOP predicts true relations between "Gadhafi" and "Libya". Although both MILP-ATLOP and JMRL-ATLOP may discover the rule "$\mathrm{citizen\_of}(x, y) \leftarrow \mathrm{head\_of\_state}(x, y)$", MILP-ATLOP propagates the false relation "head_of_state" between "Gadhafi" and "Iran" to final predictions, while JMRL-ATLOP can avoid error propagation by its end-to-end nature. Besides, JMRL-ALTOP has $\delta > 4$ in both cases, implying that it is the rule reasoning module that dominates the ultimate prediction.

6 Conclusion and Future work

In this paper we have proposed an end-to-end learning framework named JMRL to empower existing DocRE models with stronger reasoning abilities. Notably, we have proposed a novel rule reasoning module in JMRL to simulate the inference of logical rules, thereby enhancing the reasoning ability. Furthermore, we have shown theoretically that the parameterization of this module is faithful to the formalization of logical rules. Experimental results on four benchmark datasets verify the effectiveness of JMRL. Future work will extend JMRL to jointly learn named entity recognition (NER), DocRE and more expressive rules in an end-to-end fashion.

7 Limitations

The main limitations of JMRL are two-fold. On one hand, the rule reasoning module in JMRL simulates the inference of chain-like logical rules. However, chain-like logical rules may not be sufficiently expressive in some complex reasoning scenarios, e.g., they cannot express type constraints (Wu et al., 2022) on individual entities. The limited expressivity of chain-like logical rules may impair the reasoning ability of JMRL. On the other hand, JMRL is a rule-based framework for enhancing the DocRE task, whereas the task of DocRE requires a set of entities involved in the given document as input. Therefore, applying JMRL to the real-world scenarios requires a preprocess of named entity recognition (NER). Errors coming from an imperfect NER model may propagate to JMRL, resulting in performance degradation. We will make up for the above deficiencies in future work, by extending JMRL to learn more expressive logical rules and extending JMRL to jointly train an NER module.

8 Ethics Statement

JMRL is a SOTA solution for the DocRE task with high effectiveness and interpretability. Therefore, JMRL may be used to extract private information among different users. To mitigate this concern, we only use public benchmark datasets for evaluation. These datasets do not involve users' private information. Moreover, the proposed JMRL framework should not be used to extract and analyze any private information without user authorization.

References

Serge Abiteboul, Richard Hull, and Victor Vianu. 1995. *Foundations of Databases*. Addison-Wesley.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 2023. CHEER: centrality-aware high-order event reasoning network for document-level event causality identification. In *ACL*, pages 10804–10816.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *EMNLP*, pages 4924–4935.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted rule injection for relation embeddings. In *EMNLP*, pages 1389–1399.

Lingjia Deng and Janyce Wiebe. 2015. Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *EMNLP*, pages 179–189.

Boyang Ding, Quan Wang, Bin Wang, and Li Guo. 2018. Improving knowledge graph embedding using simple constraints. In *ACL*, pages 110–121.

Jianfeng Du, Jeff Z. Pan, Sylvia Wang, Kunxun Qi, Yuming Shen, and Yu Deng. 2019. Validation of growing knowledge graphs by abductive text evidences. In *AAAI*, pages 2784–2791.

Shengda Fan, Shasha Mo, and Jianwei Niu. 2022. Boosting document-level relation extraction by mining and injecting logical rules. In *EMNLP*, pages 10311–10323.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *CoRR*, abs/2305.14450.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in docred. In *ACL*, pages 6241–6252.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *EMNLP*, pages 5495–5505.

Tao Li and Vivek Srikumar. 2019. Augmenting neural networks with first-order logic. In *ACL*, pages 292–302.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187.

Jian Liu, Chen Liang, Jinan Xu, Haoyan Liu, and Zhe Zhao. 2023. Document-level event argument extraction with a chain reasoning paradigm. In *ACL*, pages 9570–9583.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*, pages 3219–3232.

Linhao Luo, Jiaxin Ju, Bo Xiong, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Chatrule: Mining logical rules with large language models for knowledge graph reasoning. *CoRR*, abs/2309.01538.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023a. DREEAM: guiding attention with evidence for improving document-level relation extraction. In *EACL*, pages 1963–1975.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of EMNLP*, pages 10572–10601.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? A study on specification-heavy tasks. *CoRR*, abs/2311.08993.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Trans. Assoc. Comput. Linguistics (TACL)*, 5:101–115.

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning logic rules for document-level relation extraction. In *EMNLP*, pages 1239–1250.

Ali Sadeghian, Mohammadreza Armandpour, Patrick Ding, and Daisy Zhe Wang. 2019. DRUM: end-to-end differentiable rule mining on knowledge graphs. In *NeurIPS*, pages 15321–15331.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *ACL*, pages 4309–4316.

Daniil Sorokin and Iryna Gurevych. 2017. Context-aware representations for knowledge base relation extraction. In *EMNLP*, pages 1784–1789.

Giuseppe Spillo, Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2022. Knowledge-aware recommendations based on neuro-symbolic graph embeddings and first-order logical rules. In *RecSys*, pages 616–621.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of ACL*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred - addressing the false negative problem in relation extraction. In *EMNLP*, pages 8472–8487.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. HIN: hierarchical inference network for document-level relation extraction. In *PAKDD*, volume 12084, pages 197–209.

Wenya Wang and Sinno Jialin Pan. 2020. Integrating deep learning with logic fusion for information extraction. In *AAAI*, pages 9225–9232.

Hong Wu, Zhe Wang, Kewen Wang, and Yi-Dong Shen. 2022. Learning typed rules over knowledge graphs. In *KR*.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *AAAI*, pages 14149–14157.

Zezhong Xu, Peng Ye, Hui Chen, Meng Zhao, Huajun Chen, and Wen Zhang. 2022. Ruleformer: Context-aware rule mining over knowledge graph. In *COLING*, pages 2551–2560.

Fan Yang, Zhilin Yang, and William W. Cohen. 2017. Differentiable learning of logical rules for knowledge base reasoning. In *NIPS*, pages 2319–2328.

Yuan Yang and Le Song. 2020. Learn to explain efficiently via neural logic inductive learning. In *ICLR*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *ACL*, pages 764–777.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag. (IPM)*, 58(4):102563.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *EMNLP*, pages 1630–1640.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Yubin Wang, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *COLING*, pages 1630–1641.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021a. Document-level relation extraction with adaptive thresholding and localized context pooling. In *AAAI*, pages 14612–14620.

Yichao Zhou, Yu Yan, Rujun Han, J. Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021b. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *AAAI*, pages 14647–14655.

A Proof

A.1 Proof of Lemma 1

To prove Theorem 1, we first introduce Lemma 1.

Lemma 1. *Suppose* $[\sigma(\mathcal{F}(x,y,d))]_r = 1$ *if the fact* (x,r,y) *is predicted to be true in document* d*, or* $[\sigma(\mathcal{F}(x,y,d))]_r = 0$ *otherwise. Let* $\mathcal{R}_\dagger = \mathcal{R}_+$ *if* $r = \perp$ *or* $\mathcal{R}_\dagger = \mathcal{R}$ *otherwise,* $\mathcal{G}_d = \{(x,r,y) \in \mathcal{E}_d \times \mathcal{R}_\dagger \times \mathcal{E}_d \mid [\sigma(\mathcal{F}(x,y,d))]_r = 1\}$ *be the set*

of predicted true facts for d, R an r-specific L-CR of the form $r(x,y) \leftarrow r_1(x,z_1) \wedge r_2(z_1,z_2) \wedge ... \wedge r_L(z_{L-1},y)$, and $\theta_r^{(1,L)}$ the $\{R\}$-induced parameter assignment of the rule reasoning module in JMRL. Then for any fact $(a,r,b) \in \mathcal{E}_d \times \mathcal{R}_\dagger \times \mathcal{E}_d$, we have: (1) $s_{r,a,b,d}^{(1,L)} \geq 1$ if $\mathcal{G}_d \models H_R(a,b)$, and (2) $s_{r,a,b,d}^{(1,L)} = 0$ if $\mathcal{G}_d \not\models H_R(a,b)$.

Proof. Let $\mathcal{K}_d = \mathcal{G}_d \cup \mathcal{G}_d^- \cup \{(e,I,e) \mid e \in \mathcal{E}_d\}$, where $\mathcal{E}_d$ is the set of entities appearing in $\mathcal{G}_d$.

(I) Consider the case where $\mathcal{G}_d \models H_R(a,b)$. There exists at least one ground instance R_g of R such that $H_R(a,b) = H_{R_g}$ and $B_{R_g} \subseteq \mathcal{K}_d$. There will be a sequence of entities $c_1, \ldots, c_{L-1}$ and a sequence of relations $r_1, \ldots, r_L$ such that $(a,r_1,c_1), (c_1,r_2,c_2) \ldots, (c_{L-1},r_L,b) \in \mathcal{K}_d$. Suppose r_1 is the k^{th} relation in $\mathcal{R}_\dagger \cup \mathcal{R}_\dagger^- \cup \{I\}$, then by Condition 1 in Definition 1, we have $w_k^{(r,1,1)} = 1$ for some k. By Equation (1), we further have $s_{r,a,c_1,d}^{(1,1)} \geq 1$. Likewise, suppose r_2 is the k^{th} relation in $\mathcal{R}_\dagger \cup \mathcal{R}_\dagger^- \cup \{I\}$, then by Condition 1 in Definition 1, we have $w_k^{(r,1,2)} = 1$. By Equation (1), we further have $s_{r,a,c_2,d}^{(1,2)} \geq 1$. In the same way, we can show that $s_{r,a,c_3,d}^{(1,3)} \geq 1, \ldots, s_{r,a,c_{L-1},d}^{(1,L-1)} \geq 1$ and $s_{r,a,b,d}^{(1,L)} \geq 1$ in turn. Therefore, we have $s_{r,a,b,d}^{(1,L)} \geq 1$ if $\mathcal{G}_d \models H_R(a,b)$.

(II) Consider the case where $\mathcal{G}_d \not\models H_R(a,b)$. Suppose $s_{r,a,b,d}^{(1,L)} \geq 1$, then by Equation (1), there must be some $k \in \{1,\ldots,m\}$ such that $w_k^{(r,1,1)} = 1$, where $m = 2n+3$ if $r = \perp$ or $2n+1$ otherwise, there exists $(a,r_k,c_1) \in \mathcal{K}_d$ fulfilling $s_{r,a,c_1,d}^{(1,1)} \geq 1$. Since $s_{r,a,c_1,d}^{(1,1)} \geq 1$, by Equation (1), there must be also some $k \in \{1,\ldots,m\}$ such that $w_k^{(r,1,2)} = 1$, where $m = 2n+3$ if $r = \perp$ or $2n+1$ otherwise, there exists $(c_1,r_k,c_2) \in \mathcal{K}_d$ fulfilling $s_{r,a,c_2,d}^{(1,2)} \geq 1$. In the same way, we can show that there exists relation r_u and entity c_u such that $(c_{u-1},r_u,c_u) \in \mathcal{K}_d$ and $s_{r,a,c_u,d}^{(1,u)} \geq 1$ for $u = 3,\ldots,L-1$ in turn, while there exists relation r_L such that $(c_{L-1},r_L,b) \in \mathcal{K}_d$. Hence there exists a sequence of entities $c_1, \ldots, c_{L-1}$ and a sequence of relations $r_1, \ldots, r_L$ such that (a,r_1,c_1), $(c_1,r_2,c_2) \ldots$, $(c_{L-1},r_L,b) \in \mathcal{K}_d$. These two sequences constitute a ground instance R_g of R such that $H_R(a,b) = H_{R_g}$ and $B_{R_g} \subseteq \mathcal{K}_d$, contradicting $\mathcal{G}_d \not\models H_R(a,b)$. Thus $s_{r,a,b,d}^{(1,L)} < 1$. By Equation (1), Condition 1 in Definition 1 and $\forall (x,r,y) \in \mathcal{E}_d \times \mathcal{R}_\dagger \times \mathcal{E}_d : [\sigma(\mathcal{F}(x,y,d))]_r \in \{0,1\}$, we further have $s_{r,a,b,d}^{(1,L)} = 0$. Therefore, we have $s_{r,a,b,d}^{(1,L)} = 0$ if $\mathcal{G}_d \not\models H_R(a,b)$. □

A.2 Proof of Theorem 1

Proof. Lemma 1 implies that, for all $R_k \in \Sigma$, $s_{r,a,b,d}^{(k,L)} \geq 1$ if $\mathcal{G}_d \models H_{R_k}(a,b)$ and $s_{r,a,b,d}^{(k,L)} = 0$ otherwise.

($\Rightarrow$) Suppose $s_{r,a,b,d}^{(N,L)} \geq 1$. Then by Equation (2) and Condition 2 in Definition 1, there exists at least one r-specific L-CR $R_k \in \Sigma$ such that $s_{r,a,b,d}^{(k,L)} \geq 1$. By Lemma 1 we have $\mathcal{G}_d \models H_{R_k}(a,b)$. Since $\mathcal{G}_d \models H_{R_k}(a,b)$ and $R_k \in \Sigma$, we have $\mathcal{G}_d \models_\Sigma (a,r,b)$.

($\Leftarrow$) Suppose $\mathcal{G}_d \models_\Sigma (a,r,b)$. Then we have $\mathcal{G}_d \models H_{R_k}(a,b)$ for some $R_k \in \Sigma$. By Lemma 1 we have $s_{r,a,b,d}^{(k,L)} \geq 1$ and for all $k' \neq k$, $s_{r,a,b,d}^{(k',L)} \geq 0$. By Equation (2) and Condition 2 in Definition 1, we have $s_{r,a,b,d}^{(N,L)} \geq 1$. □

B Rule Extraction

Based on the theoretical result of Theorem 1, we can interpret chain-like rules (CRs) from the parameter assignment of the rule reasoning module in JMRL. The process of interpretation is shown in Algorithm 1. Intuitively, Algorithm 1 interprets CRs from the parameter assignment of the rule reasoning module in JMRL using beam search, where b is the beam size, f_l is the set of (R', ψ)-pairs for the l^{th} atom, and where R' is the currently interpreted (partial) rule and ψ its estimated score. It should be noted that the process for interpreting r-specific L-CRs outputs up to b interpreted rules for a target rule, where all interpreted rules for the k^{th} target rule share the same confidence score $\alpha_r^{(k)}$.

C Formalization of Loss Functions

Due to space limitation, we omit the detailed formalization of the BCE loss function and the AT loss function in Section 4. In the following, we supplement these formalization as follows.

Let $\mathcal{D} = \{d_i\}_{1 \leq i \leq N_\mathcal{D}}$ be the set of documents for training, $\mathcal{E}_d$ the set of mentioned entities in document $d \in \mathcal{D}$, and $\mathcal{G}_d = \{(e_h,r,e_t)_i\}_{1 \leq i \leq N_{\mathcal{G}_d}}$ the set of annotated facts in document $d \in \mathcal{D}$, where $e_h, e_t \in \mathcal{E}_d$, $r \in \mathcal{R}_+$, $N_\mathcal{D}$ denotes the number of documents in $\mathcal{D}$, and $N_{\mathcal{G}_d}$ the number of facts in $\mathcal{G}_d$. Then the BCE loss function $\mathcal{J}_{\text{BCE}}^{(x,y,d)}$ for the

Algorithm 1: Interpreting r-specific L-CRs

Input: beam size $b \geq 1$ and a parameter assignment of the rule reasoning module in JMRL for the head relation r, namely $\theta_r^{(N,L)} = \{w_i^{(r,k,l)}\}_{1\leq k\leq N, 1\leq l\leq L, 1\leq i\leq m} \cup \{\alpha_r^{(k)}\}_{1\leq k\leq N}$ where $m = 2n+3$ if $r = \perp$ or $m = 2n+1$ otherwise.
Output: a set of up to bN r-specific L-CRs
$\mathbb{R} \leftarrow \emptyset$;
for $1 \leq k \leq N$ **do**
  $f_0 \leftarrow \{(\Delta^L, 1)\}$ where Δ denotes a placeholder to be filled;
  $\forall 1 \leq l \leq L : f_l \leftarrow \emptyset$;
  for $1 \leq l \leq L$ **do**
    for $(R, \psi) \in f_{l-1}$ **do**
      for $1 \leq i \leq m$ **do**
        $R' \leftarrow R$ with the l^{th} placeholder replaced with r_i;
        $f_l \leftarrow f_l \cup \{(R', w_i^{(r,k,l)}\psi)\}$;
    sort $f_l = \{(R,\psi)_j\}_{1\leq j\leq bm}$ in the descending order of ψ and preserve the top-b in f_l;
  $\mathbb{Q} \leftarrow \{R'$ rewritten from R to the form of a CR $\mid (R, \psi) \in f_L\}$;
  $\mathbb{R} \leftarrow \mathbb{R} \cup \mathbb{Q}$;
return $\mathbb{R}$;

entity pair (x, y) in document d is defined as

$$\mathcal{J}_{\text{BCE}}^{(x,y,d)} = -\sum_{r\in\mathcal{R}_+} \mathbb{I}((x,r,y) \in \mathcal{G}_d) \log \sigma(\phi_r^{(x,y,d)}) + \mathbb{I}((x,r,y) \notin \mathcal{G}_d) \log(1 - \sigma(\phi_r^{(x,y,d)})) \quad (4)$$

where σ denotes the sigmoid function, and $\mathbb{I}(C)$ is an indicator function that returns 1 if C is true or 0 otherwise. The adaptive thresholding (AT) loss $\mathcal{J}_{AT}^{(x,y,d)}$ for the entity pair (x, y) in d is defined as

$$\mathcal{J}_{\text{AT}}^{(x,y,d)} = -\sum_{r\in\mathcal{R}_{\text{pos}}} \frac{\exp(\phi_r^{(x,y,d)})}{\sum_{r'\in\mathcal{R}_{\text{pos}}^d\cup\{\perp\}} \exp(\phi_{r'}^{(x,y,d)})} - \frac{\exp(\phi_\perp^{(x,y,d)})}{\sum_{r'\in\mathcal{R}_{\text{neg}}^d\cup\{\perp\}} \exp(\phi_{r'}^{(x,y,d)})} \quad (5)$$

where $\mathcal{R}_{\text{pos}}^d = \{r \mid (x,r,y) \in \mathcal{G}_d, r \in \mathcal{R}\}$ and $\mathcal{R}_{\text{neg}}^d = \{r \mid (x,r,y) \notin \mathcal{G}_d, r \in \mathcal{R}\}$. Then the entire loss function is calculated by:

$$\mathcal{L}_\Delta = \sum_{d\in\mathcal{D}} \sum_{x,y\in\mathcal{E}_d, x\neq y} \mathcal{J}_\Delta^{(x,y,d)} \quad (6)$$

where $\Delta \in \{\text{BCE}, \text{AT}\}$.

D Experiments on Re-DocRE and DocGNRE

Dataset	Split	#Doc.	#Rel.	#Ent.	#Facts.
Re-DocRED	train	3053	96	59,359	85,932
	dev	500		9,684	17,284
	test	500		9,779	17,448
DocGNRE	GPT	3,053	96	59,359	96,505
	mGPT	3,053		59,359	103,561
	test	500		9,779	19,526

Table 6: Statistics on datasets, where Doc. (resp. Rel or Ent) abbreviates documents (resp. relations or entities).

Method	Ign F1	F1
DREEAM	77.34	77.94
JMRL-DREEAM (this work)	**77.98**	**78.61**

Table 7: Comparison results on Re-DocRED.

Due to the space limitation, the experiments on the Re-DocRED (Tan et al., 2022b) and DocGNRE (Li et al., 2023) datasets are reported in this section. Statistical details of Re-DocRED and DocGNRE are reported in Table 6. Note that we have done comparisons on Re-DocRED (Table 3), following the setting used in MILR for a fair comparison. This setting uses the training set of DocRED for training and uses the test set of Re-DocRED for test. To further verify the effectiveness of JMRL, we conducted experiments on the Re-DocRED dataset under the original setting, as reported in Table 7. Results show that the proposed JMRL framework pushes DREEAM by an absolute gain of 0.67% (resp. 0.64%) in terms of F1-scores (resp. Ign F1-scores). These results demonstrate that JMRL is able to enhance the SOTA DocRE method DREEAM on the Re-DocRED dataset under the original setting.

Furthermore, we also conducted experiments on the DocGNRE dataset. Note that DocGNRE is a new dataset that constitutes three training sets and a test set, where two of three training sets are enhanced by distant supervision using the large language model ChatGPT, and the test set is also enhanced by distant supervision using ChatGPT

Training data	Test data	PLM	Method	P	R	F1
Re-DocRED	DocGNRE	$\text{BERT}_{\text{base}}$	DREEAM	81.45	56.98	67.05
Re-DocRED	DocGNRE	$\text{BERT}_{\text{base}}$	JMRL-DREEAM	88.02	57.52	69.57
Re-DocRED	DocGNRE	$\text{RoBERTa}_{\text{large}}$	DREEAM	85.00	64.29	73.21
Re-DocRED	DocGNRE	$\text{RoBERTa}_{\text{large}}$	JMRL-DREEAM	**89.31**	63.12	**73.96**
Re-DocRED (GPT)	DocGNRE	$\text{BERT}_{\text{base}}$	DREEAM	83.66	57.62	68.24
Re-DocRED (GPT)	DocGNRE	$\text{BERT}_{\text{base}}$	JMRL-DREEAM	84.55	59.16	69.61
Re-DocRED (GPT)	DocGNRE	$\text{RoBERTa}_{\text{large}}$	DREEAM	84.92	63.86	72.90
Re-DocRED (GPT)	DocGNRE	$\text{RoBERTa}_{\text{large}}$	JMRL-DREEAM	83.83	65.92	73.81
Re-DocRED (mGPT)	DocGNRE	$\text{BERT}_{\text{base}}$	DREEAM	81.71	58.23	68.00
Re-DocRED (mGPT)	DocGNRE	$\text{BERT}_{\text{base}}$	JMRL-DREEA	82.55	59.39	69.08
Re-DocRED (mGPT)	DocGNRE	$\text{RoBERTa}_{\text{large}}$	DREEAM	80.93	**66.98**	73.29
Re-DocRED (mGPT)	DocGNRE	$\text{RoBERTa}_{\text{large}}$	JMRL-DREEAM	84.24	64.84	73.28

Table 8: Comparison results on the DocGNRE dataset.

Hyper-parameter	DWIE				DocRED			Re-DocRED+DocGNRE	
	LSTM	BiLSTM	GAIN	ALTOP	GAIN	ALTOP	DREEAM	DREEAM†	DREEAM
Number of rules N	20	20	20	20	20	20	20	20	20
Maximum length L	2	2	2	2	2	2	2	2	2
Optimizer for training	Adam	Adam	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Maximum training epoch	300	300	300	300	20	20	10	30	30
Learning rate (DocRE model)	1e-3	1e-3	2e-5	2e-5	2e-5	2e-5	1e-6	5e-5	2e-5
Learning rate (rule module)	1e-1	1e-1	3e-1	3e-1	3e-1	3e-1	1e-2	1e-1	1e-2
Batch size for training	4	4	4	4	4	4	4	4	4
Dropout rate	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Warmup ratio	0.0	0.0	0.0	0.06	0.0	0.06	0.1	0.06	0.06
Weight decay	0.0	0.0	1e-4	0.0	1e-4	0.0	0.0	0.0	0.0
λ for trading-off losses	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 9: Hyper-parameter settings for different datasets, where DREEAM† denotes BERT is used as PLM.

and further revised by human annotators. The comparison results are reported in Table 8. We can observe that JMRL-DREEAM is able to consistently outperform DREEAM for all settings in F1 scores on the DocGNRE set, with the sole exception being trained with Re-DocRED (mGPT) and meanwhile using $\text{RoBERTa}_{\text{large}}$ to calculate contextualized representations. These results further confirm the effectiveness of JMRL. Besides, we also observe from the comparison results that the use of distant data cannot further improve the performance of JMRL-DREEAM. The reason may lie in two-fold. On one hand, the external knowledge within the distant supervision data from ChatGPT may be covered by JMRL. On the other hand, the distant data from ChatGPT may introduce noise to training data, resulting in performance degradation.

E Hyper-parameter Details

To help reproduce our results, we provide the hyper-parameter settings used in our experiments. Table 9 reports the detailed hyper-parameter settings in regard to different baseline models and datasets. These hyper-parameters are set to maximize the Ign F1-scores on the development set.

Method	F1-score
ChatGPT (2-shot ICL)	12.4
Davinci (2-shot ICL)	22.9
GPT-4 (2-shot ICL)	27.9
FLAN-UL2 (2-shot ICL)	1.9
FLAN-UL2 (fine-tuned)	54.5
JMRL-DREEAM (this work)	**67.9**

Table 10: Comparison results on DocRED for LLMs.

F Discussion on LLMs

In this section, we provide detailed discussions on comparing JMRL with the current SOTA LLMs, including ChatGPT, GPT-4, Davinci and FLAN-UL2. Table 10 reports the comparison results on the DocRED dataset, where the results of LLMs are sourced from (Peng et al., 2023). Results show that there is a huge performance gap between the SOTA LLMs and JMRL-DREEAM on DocRED. We can also observe that the performance of FLAN-UL2 significantly improves after being fine-tuned on the

Dataset	Logical rules	Weight
DWIE	$\text{head\_of\_gov}(x, y) \leftarrow \text{head\_of\_state}(x, z) \wedge \text{in}^{-}(z, y)$	0.9999
	$\text{agency\_of}(x, y) \leftarrow \text{agency\_of}(x, z) \wedge \text{based\_in}(z, y)$	0.9999
	$\text{appears\_in}(x, y) \leftarrow \text{player\_of}(x, z) \wedge \text{appears\_in}(z, y)$	0.9999
	$\text{in}(x, y) \leftarrow \text{in}(x, z) \wedge \text{based\_in}^{-}(z, y)$	0.9999
	$\perp(x, y) \leftarrow \text{in}(x, z) \wedge \perp(z, y)$	0.9999
	$\text{mayor\_of}(x, y) \leftarrow \text{citizen\_of}(x, y)$	-0.9774
DocRED	$\text{child}(x, y) \leftarrow \text{father}^{-}(x, z) \wedge \text{sibling}(z, y)$	0.9998
	$\text{production\_company}(x, y) \leftarrow \text{series}(x, z) \wedge \text{production\_company}(z, y)$	0.9976
	$\text{publisher}(x, y) \leftarrow \text{series}(x, z) \wedge \text{developer}(z, y)$	0.9589
	$\text{mother}(x, y) \leftarrow \text{spouse}(x, z) \wedge \text{sibling}^{-}(z, y)$	0.8394
	$\perp(x, y) \leftarrow \perp^{-}(x, y)$	0.5716
	$\text{residence}(x, y) \leftarrow \text{child}(x, z) \wedge \text{residence}(z, y)$	-0.9997

Table 11: Case study of learnt rules, where r^{-} denotes the reverse relation of r.

training data. It implies that LLMs with few-shot ICL can hardly leverage the full domain knowledge within the training data. Besides, it can also be observed that JMRL-DREEAM still significantly outperforms FLAN-UL2 even after FLAN-UL2 was fine-tuned on the training data. The reasons may be two-fold. On one hand, FLAN-UL2 is too general to fit the DocRE task, which is a classification task, when compared with JMRL-enhanced models that are discriminative models. There is a significant gap between the generative training objective and the discriminative training objective for classification tasks. On the other hand, LLMs inherently suffer from the hallucination issue (Ji et al., 2023), e.g., LLMs may generate unexpected relations as the final predictions. This issue cannot be fully addressed by fine-tuning on the training data. In summary, these comparison results demonstrate that JMRL remains an effective solution for the DocRE task with SOTA performances on benchmark datasets. Furthermore, compared with LLMs, the JMRL-enhanced models have evident advantages in terms of memory cost and inference speed.

Nevertheless, combining JMRL with large language models is a promising way to further improve performances. For example, the work (Ma et al., 2023b) has shown that few-shot ICL for LLMs cannot generalize well in the IE tasks, but they found that LLMs are able to address some hard examples. This provides us with an innovate way to combine JMRL with LLMs, by employing the JMRL enhanced model to deal with most simple cases, and employing LLMs to handle some hard examples. We argue that such combination is able to help JMRL to generalize in more knowledge-intensive scenarios. Besides, the work (Luo et al., 2023) has shown that LLMs like ChatGPT can generate logical rules for reasoning, by leveraging the relational paths as input. We argue that the induced logical rules from ChatGPT can be used to initialize the parameter assignment of the rule reasoning module in JMRL. This process is possible to help JMRL learn more logical rules for reasoning, resulting in better convergence and performance.

Method	Total size	Extra size	Ratio
ATLOP	115,087,170	0	0.0%
JMRL-ALTOP	117,369,453	2,282,283	1.9%
Using NeuralLP	175,386,173	60,299,003	34%
Using DRUM	175,485,113	60,397,943	34%

Table 12: Comparison on parameter sizes.

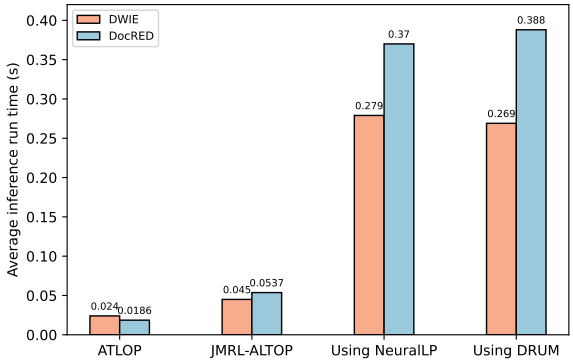

Figure 6: Comparison results on the inference time.

G Analysis on Model Efficiency

JMRL introduces external parameters to learn logical rules. To clarify whether JMRL is efficient in the DocRE task, we analyzed the model efficiency. First, we compared the model parameters of ATOP, JMRL-ATOP and other variant models, as reported in Table 12. It can be seen that JMRL-ATLOP introduces only 1.9% extra model parameters, while the other variants of JMRL-ALTOP that use NeurlLP or DRUM as the rule reasoning module require to

introduce more than 30% extra model parameters. These results indicate that JMRL is parameter efficient. Second, we compared the inference time of ATOP, JMRL-ATOP and other variant models. Figure 6 illustrates the comparison results between different methods on the average inference time in seconds. It can be seen that the employment of JMRL increases the inference time by about 0.03 seconds, whereas both the two variants of JMRL increase the inference time by about 0.3 seconds. These results imply that JMRL is able to significantly improve performance of the DocRE task with a small overhead on the inference time.

In addition, we also analyzed the time complexities of the rule reasoning module in JMRL and other rule-based methods. Specifically, The theoretical inference time complexity of the rule reasoning module in JMRL is $\mathcal{O}(nNL(2n+1)|\mathcal{E}|^2)$, where $n = |\mathcal{R}_*|$. By parallel implementation, the amortized time complexity reduces to $\mathcal{O}(nNL(2n+1))$. The time complexity of the baseline method LogicRE is $\mathcal{O}(nNL(2n+1)|\mathcal{E}|^2 + nNLd^2(2n+1))$, where d is the hidden size. The additional part $\mathcal{O}(nNLd^2(2n+1))$ comes from the Transformer network used in LogicRE for rule generation. Note that LogicRE is a path-based method for rule reasoning, thus it has no parallel implementation for rule reasoning. The amortized time complexity reduces to $\mathcal{O}(nNL(2n+1)|\mathcal{E}|^2 + nNLd^2(2n+1))$ due to the parallelization of Transformer. Besides, the amortized time complexity of previous rule learning methods NeuralLP and DRUM is $\mathcal{O}(nNL(2n+1)+nNLd^2(2n+1))$ due to the use of LSTM for calculating $w_i^{(r,k,l)}$. These analysis shows that the rule reasoning module in JMRL is more efficient than previous rule-based methods.

H Case Study of Learnt Rules

We showcase in Table 11 some logical rules extracted from the parameter assignment of the rule reasoning module in JMRL-ATLOP for both the DWIE and DocRED datasets. These rules are extracted by applying Algorithm 1 with the beam size set to 100 and then simplified by omitting identity body atoms. The weight of each rule is sourced from $\alpha_r^{(r)}$ in Equation (2). It can be observed that expressive logical rules with different weights and different numbers of body atoms can be extracted for both the DWIE and DocRED datasets. Moreover, some rules for inferring the head relation $\perp$ can also be discovered by JMRL, see the fifth rule for DWIE and the fifth rule for DocRED. It should be noted that LogicRE and MILR do not learn rules for the head relation $\perp$. The introduction of logical rules for the head predicate $\perp$ could make the prediction of no-relation between two entities more accurate since extra information is exploited. This is also a potential reason for explaining why JMRL outperforms both LogicRE and MILR.

I Analysis on the Learnt Distribution

We analyzed the learnt distribution of $w_i^{(r,k,l)}$ on the DWIE dataset. Specifically, we utilized the mean symmetric Kullback-Leibler divergence (KLD) score between $w_i^{(r,k,l)}$ and the uniform distribution to represent the distribution of $w_i^{(r,k,l)}$. Note that $w_i^{(r,k,l)}$ is initialized randomly. The mean symmetric Kullback-Leibler divergence (KLD) score between $w_i^{(r,k,l)}$ and the uniform distribution is 2.1e-5 at the initial stage of training phase, and the KLD score increases to 4.3 after training. These results indicate that $w_i^{(r,k,l)}$ becomes sharp after training, implying that the model has learned the discrete distribution of logical rules.

J Discussion on Other Applications

JMRL is an end-to-end framework for jointly learning specific neural models and logical rules. Therefore, we argue that JMRL can be used to enhance the application scenarios where logical rules are useful. For instance, JMRL can be applied to other information extraction tasks such as document-level event argument extraction (Liu et al., 2023) and document-level event causality identification (Chen et al., 2023). Apart from information extraction, we argue that JMRL can also benefit the field of knowledge-aware recommendations (Spillo et al., 2022). The exploration of extending JMRL to these applications is a part of our future work.