LogicST: A Logical Self-Training Framework for Document-Level Relation Extraction with Incomplete Annotations

Shengda Fan[1*], **Yanting Wang**[1*], **Shasha Mo**[1*†], **Jianwei Niu**[2,3]
[1] School of Cyber Science and Technology, Beihang University, Beijing 100191, China
[2] Zhongguancun Laboratory, Beijing 100191, China
[3] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{fanshengda, wangyanting, moshasha, niujianwei}@buaa.edu.cn

Abstract

Document-level relation extraction (DocRE) aims to identify relationships between entities within a document. Due to the vast number of entity pairs, fully annotating all fact triplets is challenging, resulting in datasets with numerous false negative samples. Recently, self-training-based methods have been introduced to address this issue. However, these methods are purely black-box and sub-symbolic, making them difficult to interpret and prone to overlooking symbolic interdependencies between relations. To remedy this deficiency, our insight is that symbolic knowledge, such as logical rules, can be used as diagnostic tools to identify conflicts between pseudo-labels. By resolving these conflicts through logical diagnoses, we can correct erroneous pseudo-labels, thus enhancing the training of neural models. To achieve this, we propose LogicST, a neural-logic self-training framework that iteratively resolves conflicts and constructs the minimal diagnostic set for updating models. Extensive experiments demonstrate that LogicST significantly improves performance and outperforms previous state-of-the-art methods. For instance, LogicST achieves an increase of **7.94**% in F1 score compared to CAST (Tan et al., 2023a) on the DocRED benchmark (Yao et al., 2019). Additionally, LogicST is more time-efficient than its self-training counterparts, requiring only **10**% of the training time of CAST. Code is available at https://github.com/XingYing-stack/LogicST.

1 Introduction

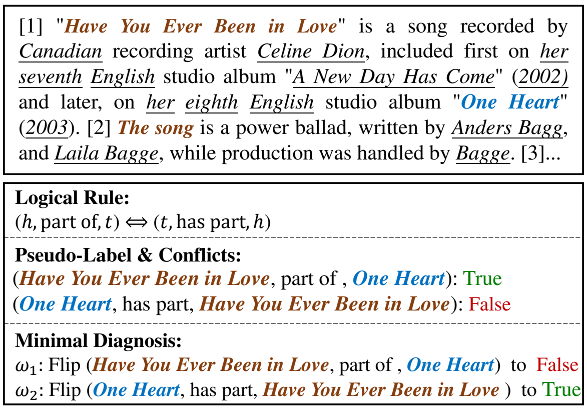
Document-level relation extraction (DocRE) aims to extract relational facts between entities within a document, playing a critical role in knowledge graph construction (Trisedya et al., 2019) and question answering (Yih et al., 2015).

Unlike sentence-level relation extraction, which focuses on individual entity pairs (Stoica et al., 2021), DocRE is challenged by the vast number of potential entity pairs. This number increases quadratically with the number of entities, making it nearly impossible for annotators to meticulously verify the validity of each triplet. Although semi-automatic strategies, such as the *recommend-revise* annotation method (Yao et al., 2019), can alleviate annotators' workload, they still fail to provide gold-quality datasets. Consequently, these datasets are prone to contain numerous false negative samples. For example, over 60% triplets are not annotated in DocRED (Huang et al., 2022). Therefore, training models from incompletely annotated datasets is crucial and practical for DocRE.

There has been extensive research aimed at alleviating the impact of false negative samples (Li et al., 2021; Wang et al., 2022a, 2024). One of the most advanced strategies is self-training (Lee et al., 2013), wherein the model reassigns labels to annotated negative triplets based on its predictions. These adjusted labels are then used iteratively to refine the model's training process. However, self-training is highly vulnerable to confirmation bias (Arazo et al., 2020). Specifically, inaccurately predicted pseudo-labels may impair the model's subsequent training. Previous works have attempted to mitigate this issue by sampling the pseudo-labels based on class frequencies (Wei et al., 2021) or scores calculated on the development set (Tan et al., 2023a). However, these pseudo-labeling methods are still far from satisfactory. **First**, they are purely sub-symbolic approaches. They benefit from the powerful representations provided by language models (Devlin et al., 2019), but struggle with symbolic reasoning among entity pairs. Therefore, they are prone to making mistakes when logical reasoning is required, leading to conflicts with the inherent interdependencies among relations. **Second**, the purely data-driven nature of these pseudo-

[*] The first three authors contributed equally.
[†] Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5496–5510
November 12-16, 2024 ©2024 Association for Computational Linguistics



[1] "***Have You Ever Been in Love***" is a song recorded by Canadian recording artist Celine Dion, included first on her seventh English studio album "*A New Day Has Come*" (2002) and later, on her eighth English studio album "***One Heart***" (2003). [2] ***The song*** is a power ballad, written by Anders Bagg, and Laila Bagge, while production was handled by Bagge. [3]...

Logical Rule:
$(h, \text{part of}, t) \Leftrightarrow (t, \text{has part}, h)$

Pseudo-Label & Conflicts:
(***Have You Ever Been in Love***, part of , ***One Heart***): True
(***One Heart***, has part, ***Have You Ever Been in Love***): False

Minimal Diagnosis:
ω_1: Flip (***Have You Ever Been in Love***, part of , ***One Heart***) to False
ω_2: Flip (***One Heart***, has part, ***Have You Ever Been in Love***) to True

Figure 1: An illustration of conflicts between pseudo-labels and logical rules. To revolve this conflict, two potential minimal diagnostic solutions are proposed, each involving the flip of a binary pseudo-label.

labeling methods makes them hard to interpret. **Third**, to achieve optimal performance, they require multiple rounds of training across various folds, significantly increasing the time consumption and limiting their practical application. Transcending these limitations calls for a fundamental paradigm shift: **i)** moving away from independent classification of each triplet to structured prediction; and **ii)** moving away from pure representation learning towards neural-symbolic computing.

Our key insight is that symbolic knowledge, such as logical rules, can be utilized as *diagnostic tools* to identify conflicts between pseudo-labels. For example, in Figure 1, we can identify conflicts such as asserting that ***Have You Ever Been in Love*** is part of ***One Heart*** while simultaneously claiming that ***One Heart*** does not have ***Have You Ever Been in Love*** as a part, which conflicts with the logical rule $(h, \texttt{part of}, t) \Leftrightarrow (t, \texttt{has part}, h)$. By correctly flipping certain pseudo-labels to resolve this conflict, we can enhance the quality of pseudo-labels and mitigate the pervasive issue of confirmation bias. Building upon this insight, we propose LogicST, a novel self-training framework that uses logical rules to diagnose pseudo-labels. LogicST is implemented within a teacher-student framework (Tarvainen and Valpola, 2017), where the teacher model is first pre-trained to establish a robust initial state. Then, the diagnosed pseudo-labels from the teacher iteratively update the student, whose parameters are in turn used to gradually update the teacher. Given the multitude of potential candidates and the high time complexity of computing diagnoses, LogicST employs a sequential diagnosis approach. Specifically, LogicST defines a scoring function that dynamically evaluates the probabilities and rewards of each diagnosis. It eliminates those with lower scores at each updating step and ultimately uses the highest-scoring diagnosis to update the student model. By doing so, our LogicST framework **i)** introduces symbolic reasoning into representation learning, **ii)** achieves better performance and interpretability, and **iii)** reduces the need for multiple rounds of training and pseudo-labeling, thus significantly improving time efficiency. Our main contributions are listed as follows:

- We propose to use logical rules as diagnostic tools to identify and correct potential errors in pseudo-labels.
- We introduce a sequential diagnosis approach that accelerates the training process while addressing the issues of imbalance and incompleteness in the weakly-supervised training corpus.
- Extensive experiments demonstrate that LogicST significantly improves performance and surpasses previous state-of-the-art methods by a large margin. Additionally, LogicST requires only **10%** of the training time of CAST (Tan et al., 2023a).

2 Related Work

Document-Level Relation Extraction. Since the advent of pre-trained language models (Devlin et al., 2019; Liu et al., 2019), research in DocRE has experienced significant growth. Substantial progress has been made through the development of complex neural networks (Zhou et al., 2021; Jiang et al., 2022; Tan et al., 2022a), the integration of evidence sentences (Huang et al., 2021; Xie et al., 2022), and the exploration of loss functions (Zhou and Lee, 2022). More recently, the use of large language models (LLMs) (Brown et al., 2020), has emerged as a promising direction for further advancements (Li et al., 2023a; Gao et al., 2023). Despite these advancements, most existing methods fail to account for the rich logical structures among relations and lack an explicit mechanism for symbolic reasoning. While some approaches attempt to incorporate logical rules, they are typically designed for fully supervised settings and perform poorly in the presence of numerous false negative samples (Fan et al., 2022). To the best of our knowledge, the LogicST framework is the first to integrate symbolic knowledge into DocRE with incomplete annotations.

Document-Level Relation Extraction with Incomplete Annotations. Existing efforts to address incomplete labeling problems in DocRE can be categorized into: *negative sampling* (Li et al., 2021), *positive-unlabeled (PU) learning* (Wang et al., 2022a, 2024), and *sub-symbolic self-training* (Tan et al., 2023a). Negative sampling avoids overfitting to false negatives but fails to effectively utilize the semantic information of unsampled samples. PU learning adjusts the loss weights assigned to relational classes but overlooks the intra-class variability among distinct samples. Sub-symbolic self-training, representing the current state-of-the-art, iteratively re-annotates negative samples, fully utilizing all sample information and accounting for within-class differences. However, it neglects the informative structures between entity pairs, leading to sub-optimal extraction performance and extended training times. To address above three limitations, this work integrates symbolic knowledge into the self-training framework, providing a novel perspective on DocRE with incomplete annotations.

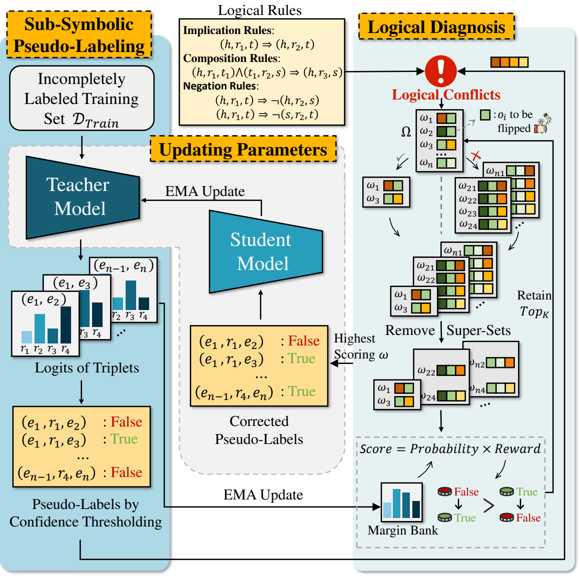

Sub-Symbolic Pseudo-Labeling
Incompletely Labeled Training Set $\mathcal{D}_{Train}$
Logical Rules
Implication Rules:
$(h, r_1, t) \Rightarrow (h, r_2, t)$
Composition Rules:
$(h, r_1, t_1) \wedge (t_1, r_2, s) \Rightarrow (h, r_3, s)$
Negation Rules:
$(h, r_1, t) \Rightarrow \neg(h, r_2, s)$
$(h, r_1, t) \Rightarrow \neg(s, r_2, t)$
Logical Diagnosis
Logical Conflicts
Updating Parameters
Teacher Model
EMA Update
Student Model
(e_{n-1}, e_n)
(e_1, e_3)
(e_1, e_2)
r_1 r_2 r_3 r_4
Logits of Triplets
(e_1, r_1, e_2) : False
(e_1, r_1, e_3) : True
...
(e_{n-1}, r_4, e_n) : False
Pseudo-Labels by Confidence Thresholding
(e_1, r_1, e_2) : False
(e_1, r_1, e_3) : True
...
(e_{n-1}, r_4, e_n) : True
Corrected Pseudo-Labels
Highest Scoring ω
Ω
ω_1 ω_2 ω_3 ω_n
: o_i to be flipped
ω_1 ω_3
ω_{n1} ω_{21} ω_{22} ω_{23} ω_{24}
Retain Top_K
Remove Super-Sets
ω_{22} ω_{24} ω_{n2} ω_{n4}
$Score = Probability \times Reward$
EMA Update
Margin Bank
False → True > True → False

Figure 2: The workflow of the LogicST framework involves three main steps. First, the teacher model pseudo-labels triplets using confidence thresholding. Next, sequential diagnosis refines these labels. Finally, the refined labels are used to update the models.

3 Methodology

Given the training set $\mathcal{D}_{Train} = \{d_i\}_{i=1}^{|\mathcal{D}_{Train}|}$, where each document contains n named entities $\{e_i\}_{i=1}^{n}$, the objective of DocRE with incomplete annotations is to train a model that fully utilizes both the annotated positive triplets G^P and the annotated negative triplets G^N. Note that G^P only contains true positives, while G^N contains both true negatives and numerous false positives. The size of G^P is usually small, resulting in an insufficient learning signal. To address this challenge, self-training is proposed to assign pseudo-labels to G^N based on model predictions.

This paper introduces the LogicST framework, which integrates symbolic knowledge into the self-training process and constructs minimal diagnostic sets to refine pseudo-labels. The overall architecture is described in Section 3.1. The method for developing minimal diagnostic sets is detailed in Section 3.2. To streamline the diagnostic procedure and identify the optimal diagnosis, the sequential diagnosis approach and the scoring function employed are elaborated in Section 3.3. Figure 2 visually outlines the workflow of LogicST.

3.1 Overview

Algorithm 1 details the implementation of LogicST, which includes a teacher model for pseudo-labeling and a student model for online learning. Inspired by semantic segmentation studies (Wang et al., 2022b), both models share the same architecture but have different parameters. LogicST is compatible with any DocRE backbone network. Following prior work (Tan et al., 2023a), we adopt ATLOP (Zhou et al., 2021) as the backbone and NCRL (Zhou and Lee, 2022) as the loss function.

LogicST adopts a two-stage training paradigm. First, the teacher model is pre-trained on $\mathcal{D}_{\text{Train}}$ to establish a robust initial state. In the training stage, the teacher model pseudo-labels each triplet in G^N using the method $\psi(\cdot)$. Typically, existing self-training methods (Tan et al., 2023a) label triplets as true if their logits exceed a threshold f_0, i.e., $\psi(X) = \mathbb{I}(f_{\theta_t}(X) > f_0)$. However, this approach treats triplets independently, ignoring their logical interdependency, which may lead to conflicts.

To address this limitation, LogicST employs a conflict resolution strategy to correct errors in pseudo-labels produced by confidence thresholding, thereby improving the training of neural backbones. Specifically, the student model updates its parameters using the corrected pseudo-labels, while the teacher model's weights are updated via an exponential moving average (EMA) of the student model's weights, ensuring a dynamic yet stable learning.

Algorithm 1: LogicST Framework

```
Input: Incompletely labeled training set $\mathcal{D}_{\text{Train}}$,
  development set $\mathcal{D}_{\text{Dev}}$, DocRE backbone model $f$,
  pseudo-labeling method $\psi$, maximum training steps
  $\text{MaxStep}$, coefficient for updating the teacher $\lambda_1$ ;
Parameters: Teacher model parameters $\theta_t$, student
  model parameters $\theta_s$ ;
Output: Optimal model parameters $\theta_{\text{best}}$ ;
Initialize parameters $\theta_{\text{pre}}$                ▷Pre-training stage
while F1 score of $f_{\theta_{pre}}$ can be improved do
    Fetch a batch of positive triplets $G^P$ and negative
      triplets $G^N$ from $\mathcal{D}_{\text{Train}}$ ;
    Update $\theta_{\text{pre}}$ using $G^P$ and $G^N$;
end
$\theta_t \leftarrow \theta_{\text{pre}}$, $\theta_s \leftarrow \theta_{\text{pre}}$                ▷Training stage
for step $= 1$ to $\text{MaxStep}$ do
    Fetch a batch of positive triplets $G^P$ and negative
      triplets $G^N$ from $\mathcal{D}_{\text{Train}}$ ;
    for $X := (h, r, t)$ in $G^N$ do
        if $f_{\theta_t}(X) == \text{True}$ then
            $G^P \cup = \{X\}, G^N \setminus = \{X\}$;
        end
    end
    Update $\theta_{\text{pre}}$ using $G^P$ and $G^N$;
    $\theta_t = (1 - \lambda_1) \cdot \theta_t + \lambda_1 \cdot \theta_s$
end
Evaluate $\theta_t$, $\theta_s$ on the $\mathcal{D}_{\text{Dev}}$ dataset;
return the better model between $\theta_t$ and $\theta_s$;
```

3.2 Logical Correction of Pseudo-Labels

LogicST integrates symbolic knowledge into its pseudo-labeling method $\psi(\cdot)$, which uses logical rules to identify *conflicts* between pseudo-labels, and generate *diagnostic sets* to resolve *conflicts* (Reiter, 1987). LogicST considers $\mathcal{S} =< G^P \cup G^N, f_{\theta_t}, \mathcal{K} >$ as the input for resolving conflicts:

- $G^P \cup G^N$ includes all possible relational triplets.
- f_{θ_t} is the teacher model that computes logits and binary pseudo-labels $\mathcal{O} = \{o_{tri}\}_{tri \in G^P \cup G^N}$ for all triplets, where o_{tri} is defined as a key-value pair in the form *triplet: boolean value*.
- $\mathcal{K}$ is a finite set of first-order logical rules that symbolically capture the logical dependencies between relations. To ensure the quality of the logical rules, we use a frequency-based approach (Fan et al., 2022) to construct $\mathcal{K}$ from the development set, including implication, composition, and negation rules. Definitions and examples of used rules are provided in Appendix A.

Conflicts arise when the pseudo-labels $\mathcal{O}$ produced by the teacher model are not compatible with the logical rules $\mathcal{K}$:

$$\mathcal{K} \wedge \mathcal{O} \wedge \bigwedge_{o \in \mathcal{O}} \text{keep}(o) \vdash \bot, \tag{1}$$

where $\vdash$ denotes logical entailment and $\bot$ signifies a logical contradiction. The unary predicates keep$(\cdot)$ and flip$(\cdot)$ indicate whether to keep or flip the boolean value of o, respectively. Furthermore, LogicST endeavors to identify and flip a subset ω of $\mathcal{O}$ to resolve conflicts, a process termed as a *diagnosis* (de Kleer and Williams, 1987):

$$\mathcal{K} \wedge \mathcal{O} \wedge \bigwedge_{o \in \mathcal{O} \setminus \omega} \text{keep}(o) \wedge \bigwedge_{o \in \omega} \text{flip}(o) \vdash \top, \tag{2}$$

where $\top$ signifies that the resulting condition is logically consistent. As shown in Figure 1, for a logical conflict involving predicates $o_1, o_2, \ldots, o_\ell$, the conflict can be resolved by flipping any individual predicate $o_i \in \{o_1, \ldots, o_\ell\}$. Specifically, flipping a predicate in the antecedent will make the rule no longer applicable, whereas flipping the consequent predicate will make the rule logically consistent. Based on Occam's razor (Domingos, 1999), we only consider the *minimal diagnosis*, where no subset of the minimal diagnostic set can resolve the conflict. Inspired by circuit diagnosis theory (Reiter, 1987), we calculate minimal diagnoses by identifying conflicts. As illustrated in Algorithm 2 and Figure 1, LogicST employs an iterative process to identify and resolve logical conflicts between the pseudo-labels $\mathcal{O}$ and the rules $\mathcal{K}$.

During each iteration, the algorithm checks whether the current minimal candidate set ω can resolve the conflict between the pseudo-labels $\mathcal{O}$ and the logical rules $\mathcal{K}$. This process involves evaluating whether the intersection $\omega \cap \{o_1, \ldots, o_\ell\}$ is an empty set, where $\{o_1, \ldots, o_\ell\}$ represents the set of conflicting pseudo-labels in this iteration:

1. **Case 1: Conflict Already Resolved**

If the intersection $\omega \cap \{o_1, \ldots, o_\ell\} \neq \emptyset$, this means that ω already contains at least one new conflicting pseudo-label to be flipped. In this case, the minimal candidate set ω can resolve the conflict without further modification, and no changes are needed. This situation is marked by the ✓ in Figure 1.

2. **Case 2: Conflict Not Resolved**

If the intersection $\omega \cap \{o_1, \ldots, o_\ell\} = \emptyset$, it indicates that none of the conflicting pseudo-labels are present in the current set ω. Therefore, ω cannot resolve the conflict, and the algorithm replaces ω with its supersets. Each superset contains one new conflicting pseudo-label o_i from $\{o_1, \ldots, o_\ell\}$. This situation is indicated by the ✗ in Figure 1.

Once the minimal candidate sets are expanded and evaluated, LogicST proceeds to remove any

Algorithm 2: Resolution of Logical Conflicts via Minimal Diagnosis

```
Input: Logical rules $\mathcal{K}$, pseudo-labels $\mathcal{O}$ ;
Output: The complete set of minimal diagnostic sets
    $\Omega = \{\omega_j\}_{j=1}^{|\Omega|}$ ;
Initialize the set of diagnostic sets $\Omega = \{\emptyset\}$;
for $k$ in $\mathcal{K}$ where $k$ involves $\ell$ triplets do
    foreach $(o_1, o_2, \cdots, o_\ell)$ in $\mathcal{O}$ do
        if $\neg Check(o_1, o_2, \cdots, o_\ell | k)$ then
            $\Omega \leftarrow \{\omega \cup \{o_i\} : \forall o_i, \omega \in \Omega, \omega \cap \{o_1, \dots, o_\ell\} = \emptyset\} \cup \{\omega : \omega \in \Omega, \omega \cap \{o_1, \dots, o_\ell\} \neq \emptyset\}$ ;
        end
        for $\omega$ in $\Omega$ do
            if $\exists \omega' \in \Omega$, s.t. $\omega' \subset \omega$ then
                $\Omega \leftarrow \Omega \setminus \{\omega\}$ ;
            end
        end
    end
end
return $\Omega$ ;
```

supersets that are redundant (as shown by the removal of ω_{21} and other supersets in Figure 1). This ensures that only the minimal diagnosis sets, those that can resolve the conflict with the fewest changes, remain. Each iteration of this process updates the pseudo-labels accordingly, refining the labels used for subsequent model updates.

3.3 Sequential Diagnosis

Although Algorithm 2 can compute all minimal candidate sets, it incurs a time complexity of $O(|\Omega|^2)$ in each iteration to remove non-minimal sets, which is impractical when dealing with numerous conflicts. Moreover, as shown in Figure 1, there are usually multiple minimal candidate sets that meet the definition, but only one specific set will be used to flip the pseudo-labels during training. To address these challenges, LogicST introduces a scoring function $F(\omega|\mathcal{S})$ to evaluate a diagnosis ω given the input $\mathcal{S}$. After updating the minimal diagnostic sets in each iteration, only the Top_K diagnoses with the highest scores are retained, significantly reducing the time required. Finally, the highest-scoring candidate set is used to update the student model, a process termed *best-first sampling* in model-based diagnosis (Rodler, 2022).

We design the scoring function considering the following two properties:

Probabilities of Diagnoses. LogicST assumes independent failure (de Kleer, 1991), using the multiplication rule to calculate probabilities:

$$P(\omega|\mathcal{S}) = \prod_{o \in \omega^+} P(o|\mathcal{S}) \prod_{o \in \omega^-} P(\neg o|\mathcal{S}), \quad (3)$$

where ω^+ and ω^- are sets of triplets to be flipped to true and false, respectively. $P(o|\mathcal{S})$ is the probability of the corresponding triplet predicted to be true given the input $\mathcal{S}$, and $P(\neg o|\mathcal{S}) = 1 - P(o|\mathcal{S})$.

A naive method to compute this probability is by applying the sigmoid function to the margin between the classification logits and the threshold logits:

$$P(o|\mathcal{S}) = \sigma(f_i - f_0), \quad (4)$$

where f_i is the teacher model's logit for the relation class i corresponding to o, and f_0 is the classification threshold score for that triplet. However, DocRE tasks often suffer from severe class imbalance (Tan et al., 2022a), causing logits to be biased towards popular classes (Menon et al., 2021) and introducing confirmation bias (Cho and Roy, 2004).

To address this, LogicST adopts an adaptive post-hoc logit adjustment to compensate for minority classes. We maintain a *margin bank* to dynamically record the training status of each class. At the step-th training iteration, we first use the batch-wise mean margin to evaluate the instant performance of class i:

$$\text{Margin}^i_{\text{step}} = \frac{1}{|G_i^P| + |G_i^N|} \sum_{G_i^P \cup G_i^N} (f_i - f_0), \quad (5)$$

where G_i^P and G_i^N denote the triplets belonging to class i in G^P and G^N, respectively. Then we use the EMA of $\text{Margin}^i_{\text{step}}$ to stably measure the performance of each class:

$$\text{Margin}^i = (1 - \lambda_2) \cdot \text{Margin}^i + \lambda_2 \cdot \text{Margin}^{\text{i}}_{\text{step}}, \quad (6)$$

where $\lambda_2 \in [0, 1)$ is the momentum coefficient. We use the difference between the sample-wise margin $f_i - f_0$ and the class-wise margin Margin^i to more fairly calculate the probability:

$$P(o|\mathcal{S}) = \sigma(f_i - f_0 - \tau \cdot \text{Margin}^i), \quad (7)$$

where τ is a hyper-parameter controlling the intensity of compensation.

Rewards of Diagnoses. In the early phase of the training stage, the teacher models pre-trained under numerous false-negative samples tend to generate high-precision but low-recall pseudo-labels (Tan et al., 2022b). Thus, the reward of flipping pseudo-labels from false to true is greater. As training progresses and the teacher model is updated, the importance of precision will gradually increase, while the importance of recall will decrease. Therefore, the reward of a diagnosis is defined as:

$$R(\omega|\mathcal{S}) = \gamma^{\frac{|\omega^{+}|}{1+\text{epoch}}}, \tag{8}$$

where $\gamma > 1$ is a hyper-parameter that measures the importance of the reward.

Combining the two properties above, the score of a diagnostic set ω is defined as the product of its probability and its reward:

$$F(\omega|\mathcal{S}) = P(\omega|\mathcal{S}) \cdot R(\omega|\mathcal{S}). \tag{9}$$

4 Experiment

4.1 Experimental Setup

Datasets. The experiments are conducted on the DocRED (Yao et al., 2019) and DWIE (Zhang et al., 2022a) datasets. DocRED is a large-scale and widely-used benchmark, but it is known to have numerous missing annotations. We use the incompletely labeled training set of DocRED and the revised development and test sets of Re-DocRED (Tan et al., 2022b) and DocGNRE (Li et al., 2023b) to validate the models' effectiveness. Additionally, we experiment on the extremely incomplete training set, DocRED_ext (Wang et al., 2022a), where the number of labels for each relation type in a document is limited to one. DWIE (Zaporojets et al., 2021) is a human-annotated dataset. To create incompletely annotated training sets, we uniformly sample 20%, 40%, 60%, and 80% of positive triplets to build labels. The original development and test sets of DWIE are used for evaluation. The dataset statistics are provided in Appendix B.

Evaluation Metrics. We utilize F1, Ign F1, precision, and recall as the primary metrics, where the Ign F1 score excludes triplets shared between the training and test sets to avoid data leakage. We also compute F1 scores for frequent classes (the top 10 most common relation types) and long-tail classes (all other relation types), denoted as Freq_F1 and LT_F1, respectively. Additional implementation details are provided in Appendix C.

4.2 Baselines

We compare LogicST to the following six types of baselines: 1) vanilla baselines, including various top-performing models under fully supervised settings, such as GAIN (Zeng et al., 2020), ATLOP (Zhou et al., 2021), and KD-DocRE (Tan et al., 2022a); 2) negative sampling methods (Li et al., 2021); 3) PU learning-based methods, including SSR-PU (Wang et al., 2022a) and P^{3}M (Wang et al., 2024); 4) sub-symbolic self-training methods, such as VST (Jie et al., 2019), CREST (Wei et al., 2021), and CAST (Tan et al., 2023b); 5) methods based on large language models (LLMs), including LLaMA2-7B (Touvron et al., 2023), GPT-3.5 (OpenAI, 2022), and GPT-4o (OpenAI, 2024), as well as techniques utilizing in-context learning (ICL) [1] for task-specific adaptation (Dong et al., 2022), natural language inference (NLI) models for fuzzy matching (Li et al., 2023a), and data programming for label denoising (Gao et al., 2023); and 6) logical frameworks designed for supervised DocRE, including LogiRE (Ru et al., 2021), MILR (Fan et al., 2022), and JMRL (Qi et al., 2024).

4.3 Main Results

Results on DocRED. Table 1 presents the quantitative comparisons on Re-DocRED[2], from which we draw four observations: **First**, even with the inclusion of sophisticated adaptation techniques, all LLMs still underperform compared to specialized models. This may be due to LLMs' difficulty in handling complex reasoning, and domain-specific nuances (Pang et al., 2023), as well as the lack of task-specific tuning and sufficient labeled data for relation extraction tasks (Zhang et al., 2023). Furthermore, we observe that adding more noisy in-context samples can mislead LLMs, degrading their performance. **Second**, LogicST surpasses all baselines by a large margin, achieving a **7.94%** absolute F1 improvement over CAST, establishing new state-of-the-art results with **69.26%** F1 using BERT-base and **73.29%** F1 using RoBERTa-large. This significantly narrows the gap with fully supervised ATLOP-BERT's 74.02% performance. **Third**, while all weakly-supervised methods aim to enhance overall performance, often at the expense of precision, LogicST maintains a superior balance between precision and recall. Remarkably, LogicST either achieves or closely approaches the best performance in both precision and recall among all these methods. **Finally**, by incorporating symbolic logic, LogicST mitigates the confirmation bias inherent in self-training and the class imbalance problem in the training set, thereby improving perfor-

[1] The detailed configurations can be found in Appendix D.

[2] The results on DocGNRE can be found in Appendix E.

	Model	Precision	Recall	F1	Ign_F1	Freq_F1	LT_F1
LLMs	GPT-3.5	13.12	2.85	4.68	-	-	-
	GPT-3.5 + NLI	23.57	6.14	9.74	-	-	-
	LLaMA2	5.70	25.50	9.32	8.04	-	-
	LLaMA2 + DP	6.56	27.00	10.56	9.03	-	-
	GPT-4o	23.57	6.14	21.41	21.17	-	-
	GPT-4o + ICL (1 shot)	47.45	19.61	27.75	27.36	-	-
	GPT-4o + ICL (3 shot)	39.51	20.67	27.14	26.53	-	-
BERT	GAIN	88.11	30.98	45.82	45.57	-	-
	ATLOP	**88.39** ±0.39	28.87 ±0.34	43.52 ±0.25	43.28 ±0.24	45.49 ±0.24	40.46 ±0.28
	NS-ATLOP	74.79 ±0.31	46.33 ±0.34	57.22±0.25	56.28 ±0.21	59.23 ±0.23	54.13 ±0.24
	SSR-PU-ATLOP	65.10 ±0.90	50.53 ±0.89	56.84 ±0.72	55.45 ±0.59	60.21 ±0.64	51.84 ±0.82
	P^{3}M-ATLOP‡	64.50 ±0.49	58.65 ±0.33	61.43 ±0.06	60.17 ±0.08	68.20 ±0.20	52.54 ±0.17
	VST-ATLOP	63.53 ±1.17	56.41 ±0.86	59.56 ±0.16	58.03 ±0.25	63.17 ±0.46	55.61 ±0.25
	CREST-ATLOP	69.34 ±1.55	50.58 ±1.35	58.48 ±0.30	57.33 ±0.21	60.31 ±0.64	56.33 ±0.15
	CAST-ATLOP	70.49 ±1.12	54.34 ±1.07	61.36 ±0.67	60.16 ±0.79	63.66 ±0.44	58.12 ±0.36
	LogicST-ATLOP	74.68±0.65	**63.95**±0.49	**69.26**±0.31	**68.49**±0.33	**72.74**±0.45	**64.26**±0.18
RoBERTa	KD-DocRE	92.08	32.07	47.57	47.32	-	-
	ATLOP	92.62 ±0.35	33.61 ±0.48	49.32 ±0.29	49.16 ±0.27	51.49 ±0.51	45.36 ±0.43
	NS-ATLOP	68.39 ±2.23	56.05 ±0.98	61.58 ±0.48	60.43 ±0.55	65.35 ±0.12	57.16 ±0.44
	SSR-PU-ATLOP	65.71 ± 0.28	57.01 ±0.47	61.05 ±0.21	59.48 ±0.18	62.85 ±0.10	58.19 ±0.54
	P^{3}M-ATLOP‡	69.22 ±0.42	57.95 ±0.26	63.06 ±0.36	61.98 ±0.35	69.55 ±0.64	54.58 ±0.78
	VST-ATLOP	62.85 ±0.48	63.58 ±0.62	63.21 ±0.39	61.83 ±0.41	65.68 ±0.43	60.09 ±0.45
	CREST-ATLOP	73.09 ±0.79	55.06 ±0.86	62.81 ±0.35	61.90 ±0.33	63.71 ±0.41	61.75 ±0.49
	CAST-ATLOP	72.83 ±0.50	59.22 ±0.61	65.32 ±0.22	64.25 ±0.15	66.99 ±0.29	63.05 ±0.11
	LogicST-ATLOP	78.77±0.64	**68.54**±0.48	**73.29**±0.21	**72.64**±0.22	**76.61**±0.26	**68.57**±0.19

Table 1: Experimental results on the test set of Re-DocRED (%). The reported results are the average of five runs. Results marked with ‡ are reproduced from Wang et al. (2024) using the dev set of Re-DocRED.

mance for both frequent and long-tail classes. Additionally, we compare the performance of CAST and LogicST across various relation classes in Appendix F.

Results on DocRED_ext. The experimental results using the DocRED_ext training set and the Re-DocRED test set are shown in Table 2. The proposed LogicST framework consistently outperforms all strong baselines, surpassing the previous state-of-the-art, P^{3}M, by **7.94**% in F1 score.

	Model	Precision	Recall	Ign F1	F1
BERT	ATLOP	**88.42**	12.19	21.37	21.42
	NS	69.83	34.36	45.50	46.05
	SSR-PU	59.52	39.18	46.47	47.24
	P^{3}M	61.12	53.44	56.17	57.02
	CAST	68.94	48.17	56.10	56.71
	LogicST	58.01	**71.74**	**62.72**	**64.15**
RoBERTa	ATLOP	**90.78**	12.43	21.82	21.86
	NS	70.29	34.31	45.47	46.11
	SSR-PU	61.57	41.75	48.98	49.74
	P^{3}M	63.04	57.01	59.02	59.86
	CAST	67.86	52.08	58.44	58.93
	LogicST	60.97	**76.40**	**66.44**	**67.82**

Table 2: Experimental results on Re-DocRED under extremely unlabeled settings with ATLOP (%).

Results on DWIE. As illustrated in Figure 3, LogicST consistently surpasses all baseline models across different sampling ratios, with its superiority becoming increasingly evident in scenarios of limited annotations. Remarkably, it approaches the fully supervised performance of 74.36%. This verifies the ability of logical diagnosis to mitigate false negative issues. However, LogicST's improvement is less significant compared to DocRED, which can be attributed to differences in dataset construction. The incomplete DWIE dataset is generated through uniform sampling, whereas the missing annotations in DocRED result from distant supervision, leading to biases towards popular classes and entities (Huang et al., 2022). Consequently, the DWIE dataset is simpler, reducing the performance gap between different frameworks.

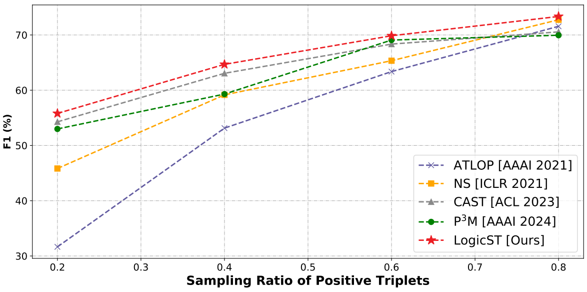


Figure 3: Comparison of F1 scores on the DWIE dataset with varying positive sampling ratios.

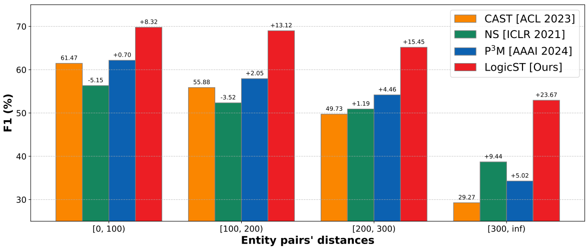


Figure 4: F1 comparison of CAST, NS, P^{3}M, and LogicST with entity pairs' distances on the test set of DocRED. The ATLOP model is used as the backbone.

4.4 Analysis & Discussion

Comparison with Other Logical DocRE Frameworks. We compare LogicST with LogiRE (Ru et al., 2021), MILR (Fan et al., 2022), and JMRL (Qi et al., 2024) using BERT-base as the encoder. The results in Table 3 show that these baselines only marginally improve the backbone's performance due to the use of noisy labels for calculating the classification loss, which inevitably leads to overfitting. In contrast, LogicST aims to correct these noisy labels, performing significantly better under conditions of incomplete annotations.

Model	Test	
	Ign F1	F1
ATLOP	43.28	43.52
+ LogiRE	44.05(+0.77)	44.33(+0.81)
+ MILR	46.08(+2.80)	46.38(+2.86)
+ JMRL	47.32(+3.27)	47.54(+3.21)
+ LogicST	**69.26(+25.98)**	**69.26(+24.47)**

Table 3: Comparison with other logical frameworks on the test set of Re-DocRED (%).

Performance with respect to Entity Pairs' Distances. We break down the relation extraction performance into four groups based on the distance between entity pairs to analyze the long-range dependency capture capabilities. As shown in Figure 4, the LogicST framework consistently outperforms all strong baselines across all groups. Moreover, the performance gains from CAST to LogicST increase as the distance grows. For distances in the ranges $[200, 300)$ and $[300, \infty)$, LogicST achieves **15.45**% and **23.67**% F1 enhancement, respectively. These results demonstrate the superiority of LogicST in incorporating rules to capture long-range dependencies and alleviate confirmation bias.

Efficiency Comparison. Table 4 presents the training time of various frameworks. Notably, only self-training frameworks such as CAST and LogicST require an additional pre-training stage to get better initial state. Unlike CAST, which necessitates multiple rounds and splits of training, our LogicST framework pre-trains the teacher model only once, significantly reducing this time overhead. During the training stage, LogicST incurs only an additional forward propagation and logical diagnosis step, resulting in an acceptable time increase compared to the vanilla training pipeline.

Framework	Pre-Training	Training	Toal
Vanilla	-	1h 1m	1h 1m
SSR-PU	-	1h 15m	1h 15m
NS	-	1h 8m	1h 8m
CAST	20h 32m	1h 1m	21h 33m
LogicST	11 m	1h 41m	1h 52m

Table 4: Training time of ATLOP-backbone frameworks training for 30 epochs with batch size 4 on DocRED.

Effect of Scoring Functions. To assess the impact of the scoring function, we plot the F1 scores of different variants over training epochs in Figure 5. The results reveal that: 1) Without considering the probability of diagnoses, the model struggles to correct false positives. Initially, it corrects false negatives, leading to a performance increase, but subsequently fits extra introduced errors, causing a performance decline; 2) Without considering the reward of diagnoses, the framework fails to correct enough false negatives in the early training stages, resulting in sub-optimal performance. We also perform a case study of margin bank in Appendix G.

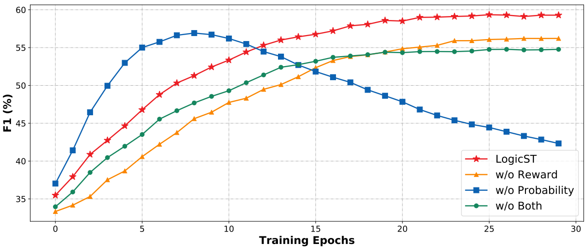


Figure 5: F1 vs. the number of training epochs on the development set of DWIE with 40% sampling ratios.

Ablation Study. We conduct an ablation experiment to assess the efficacy of LogicST's components. Additionally, we introduce a baseline termed "Fixed Diagnosis", which employs implication and composition rules to complement training labels before the two-stage training, and keeping these labels fixed. The experimental results in Table 5 reveal three key observations. **First**, the EMA teacher is essential for mitigating the impact of noisy labels and stabilizing the training process. **Second**, logi-

Model	Dev		Test	
	Ign F1	F1	Ign F1	F1
LogicST-ATLOP	**67.69**	**68.50**	**68.45**	**69.18**
- EMA Teacher	47.16	48.54	48.21	49.08
- Diagnosis	59.05	59.94	58.81	59.60
ATLOP	44.16	44.35	43.64	43.81
+ Fixed Diagnosis	61.56	62.08	61.25	61.94

Table 5: Ablative experiments on DocRED (%).

cal diagnosis significantly improves the quality and coverage of pseudo-labels, enhancing performance. **Third**, while simply adding missing triplets via logical rules yields competitive results, the "Fixed Diagnosis" method falls short of LogicST due to its inability to account for additional false negatives identified by the backbone during training.

Case Study. To further illustrate the effect of logical diagnosis and its contribution to interpretability, we present a typical example in Figure 6, where the pseudo-labels generated by confidence thresholding conflict with symbolic knowledge. For instance, ***Have You Ever Been in Love*** is labeled as part of ***One Heart***, while simultaneously, ***One Heart*** is claimed to not have part of ***the song***, which contradicts the rule: $(h,\ \texttt{part of}, t) \Leftrightarrow (t,\ \texttt{has part}, h)$. LogicST resolves this conflict by adding the missing label instead of negating a true positive, aligning pseudo-labels with logical rules and enhancing interpretability. Additionally, Figure 6 provides two similar cases. Another case study of predictions is provided in Appendix H.

[1] "***Have You Ever Been in Love***" is a song recorded by *Canadian* recording artist ***Celine Dion***, included first on ***her*** *seventh* *English* studio album "*A New Day Has Come*" (*2002*) and later, on ***her*** *eighth* *English* studio album "***One Heart***" (*2003*). [2] ***The song*** is a power ballad, written by *Anders Bagg*, and *Laila Bagge*, while production was handled by *Bagge*. [3]...

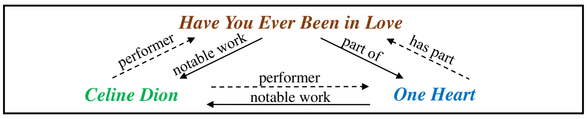


Figure 6: Case study of LogicST correcting pseudo-labels on the DocRED. Solid lines indicate pseudo-labels by confidence thresholding, while dotted lines indicate extra facts added by diagnosis.

5 Conclusion

In this work, we introduce the LogicST framework for DocRE with incomplete annotations. LogicST utilizes logical rules to identify conflicts among pseudo-labels and develop minimal diagnoses to correct potential errors. Experiments on various benchmarks demonstrate that LogicST achieves state-of-the-art results. The effectiveness and efficiency of LogicST highlight the potential of the neural-logic paradigm in incompletely labeled information extraction. We believe this paper opens a new avenue for future exploration.

Limitations

Although making some progress, our LogicST framework still has several limitations. First, the scoring function of LogicST is designed for settings with incomplete annotations and is not applicable to distant supervision settings (Mintz et al., 2009; Liu et al., 2022). Secondly, LogicST is only applicable to datasets with clear logical relationships between relations, making it unsuitable for binary datasets such as CDR (Li et al., 2016) and biomedical datasets such as ChemDisGene (Zhang et al., 2022b). Specifically, if a chemical affects the expression of a gene and that gene can be used as a marker for a disease, we cannot assume that this chemical is a therapeutic for that disease. Thirdly, LogicST assumes that entities and their mentions are identified beforehand (Li and Ji, 2014), which falls short of real-world applications. We will address these limitations in future work.

Ethics Statement

ChatGPT and Grammarly were used for parts of the writing. Compared to their sentence-level counterparts, DocRE models including the proposed LogicST, demonstrate enhanced capabilities for analyzing vast volumes of online text and identifying private information across different users. Aware of the associated privacy concerns, we ensure that all data utilized in this study is public and devoid of any personal information. Furthermore, we strongly advise against using the proposed framework for analyzing data that contains personal privacy elements in future applications.

Acknowledgements

This work was supported by National Natural Science Foundation of China (62106013).

References

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *Proc. of IJCNN*, pages 1–8. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*, pages 1877–1901. Curran Associates, Inc.

Junghoo Cho and Sourashis Roy. 2004. Impact of search engines on page popularity. In *Proc. of WWW*, pages 20–29. ACM.

Johan de Kleer. 1991. Focusing on probable diagnoses. In *Proceedings of the 9th National Conference on Artificial Intelligence, Anaheim, CA, USA, July 14-19, 1991, Volume 2*, pages 842–848. AAAI Press / The MIT Press.

Johan de Kleer and Brian C. Williams. 1987. Diagnosing multiple faults. *Artif. Intell.*, pages 97–130.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186. Association for Computational Linguistics.

Pedro Domingos. 1999. The role of occam's razor in knowledge discovery. *Data mining and knowledge discovery*, 3:409–425.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

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

Chufan Gao, Xulin Fan, Jimeng Sun, and Xuan Wang. 2023. Promptre: Weakly-supervised document-level relation extraction via prompting-based data programming. *CoRR*.

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 *Proc. of ACL*, pages 6241–6252. Association for Computational Linguistics.

Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. 2021. Three sentences are all you need: Local path enhanced document relation extraction. In *Proc. of ACL*, pages 998–1004. Association for Computational Linguistics.

Feng Jiang, Jianwei Niu, Shasha Mo, and Shengda Fan. 2022. Key mention pairs guided document-level relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1904–1914.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proc. of NAACL*, pages 729–734. Association for Computational Linguistics.

Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896.

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

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

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

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proc. of ACL*, pages 402–412. The Association for Computer Linguistics.

Yangming Li, Lemao Liu, and Shuming Shi. 2021. Empirical analysis of unlabeled entity problem in named entity recognition. In *Proc. of ICLR*. OpenReview.net.

Ruri Liu, Shasha Mo, Jianwei Niu, and Shengda Fan. 2022. Ceta: A consensus enhanced training approach for denoising in distantly supervised relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2247–2258.

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

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. Long-tail learning via logit adjustment. In *Proc. of ICLR*. OpenReview.net.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. of ACL*, pages 1003–1011. The Association for Computer Linguistics.

OpenAI. 2022. OpenAI: Introducing ChatGPT.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-10-03.

Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. 2023. Guideline learning for in-context information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15372–15389.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Kunxun Qi, Jianfeng Du, and Hai Wan. 2024. End-to-end learning of logical rules for enhancing document-level relation extraction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7247–7263.

Raymond Reiter. 1987. A theory of diagnosis from first principles. *Artif. Intell.*, pages 57–95.

Patrick Rodler. 2022. Random vs. best-first: Impact of sampling strategies on decision making in model-based diagnosis. In *Proc. of AAAI*, pages 5869–5878. AAAI Press.

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 *Proc. of EMNLP*, pages 1239–1250. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Proc. of AAAI*, pages 13843–13850. AAAI Press.

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

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2023a. Class-adaptive self-training for relation extraction with incompletely annotated training data. In *Proc. of ACL Findings*, pages 8630–8643. Association for Computational Linguistics.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2023b. Class-adaptive self-training for relation extraction with incompletely annotated training data. In *Proc. of ACL Findings*, pages 8630–8643. Association for Computational Linguistics.

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 *Proc. of EMNLP*, pages 8472–8487. Association for Computational Linguistics.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. of NeurIPS*, pages 1195–1204.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bayu Distiawan Trisedya, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proc. of ACL*, pages 229–240. Association for Computational Linguistics.

Ye Wang, Xinxin Liu, Wenxin Hu, and Tao Zhang. 2022a. A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In *Proc. of EMNLP*, pages 4123–4135. Association for Computational Linguistics.

Ye Wang, Huazheng Pan, Tao Zhang, Wen Wu, and Wenxin Hu. 2024. A positive-unlabeled metric learning framework for document-level relation extraction with incomplete labeling. In *Proc. of AAAI*, pages 19197–19205. AAAI Press.

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. 2022b. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proc. of CVPR*, pages 4238–4247. IEEE.

Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan L. Yuille, and Fan Yang. 2021. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proc. of CVPR*, pages 10857–10866. Computer Vision Foundation / IEEE.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Proc. of ACL Findings*, pages 257–268. Association for Computational Linguistics.

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 *Proc.*

of ACL, pages 764–777. Association for Computational Linguistics.

Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In *Proc. of ACL*, pages 1321–1331. Association for Computational Linguistics.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: An entity-centric dataset for multi-task document-level information extraction. page 102563.

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

Dongxu Zhang, Sunil Mohan, Michaela Torkar, and Andrew McCallum. 2022a. A distant supervision corpus for extracting biomedical relationships between chemicals, diseases and genes. In *Proc. of LREC*. European Language Resources Association.

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

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812.

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

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proc. of IJCAI*, pages 4538–4544. ijcai.org.

A Details of Logical Rules

In this paper, we consider implication, composition, and negation rules, which are defined as follows:

- **Implication Rules** define logical consequences between multiple relations within an entity pair. For example, "If `London` is the `capital` of `UK`, then `London` is definitely `located in` `UK`." Formally, implication rules are defined as:

$$(h, r_1, t) \Rightarrow (h, r_2, t). \tag{10}$$

- **Composition Rules** describe how relations between entities can be combined to derive new triples. For example, "If `Paris` is the `capital` of `France`, and `France` is `located` in `Europe`, then `Paris` is `located` in `Europe` too." Formally, two-hop composition rules are represented as:

$$(h, r_1, t_1) \wedge (t_1, r_2, s) \Rightarrow (h, r_3, s). \tag{11}$$

Multi-hop compositions are defined similarly.

- **Negation Rules** express the mutual exclusivity between triples involving the same entities. For example, "If `Beijing` is the `capital` of `China`, then `Beijing` cannot logically be the `spouse` of any entity." Formally, the negation rules are categorized and defined based on the shared entity:

$$(h, r_1, t) \Rightarrow \neg(h, r_2, s) \quad \text{if } h \text{ is shared,} \tag{12}$$
$$(h, r_1, t) \Rightarrow \neg(s, r_2, t) \quad \text{if } t \text{ is shared.} \tag{13}$$

We list several used rules on the DocRED dataset in Table 7.

B Dataset Statistics

The dataset statistics are summarized in Table 6. As observed, the average number of triplets in the incompletely labeled training sets is significantly lower than in the development and test sets. This discrepancy indicates a substantial number of false negatives in the training sets.

C Implementation Details

The proposed LogicST framework is compatible with any DocRE backbone. Consistent with prior work (Tan et al., 2023a; Wang et al., 2024), we adopt the ATLOP model (Zhou et al., 2021) as our backbone. We use BERT-base (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019) as the text encoders. All models are implemented in PyTorch (Paszke et al., 2019) and trained on one Tesla V100 GPU.

For hyper-parameters, we perform a grid search for λ_1 and λ_2 within {0.99, 0.999, 0.9995}, for Top_K within {10, 20, 50}, for τ within {0.1, 0.3, 0.5}, and for γ within {2, 20, 50, 100, 1000}. All hyper-parameters are selected based on the F1 score computed on the development set.

D Experimental Configurations for In-Context Learning

In our experiment, we employ the `sentence-transformers/all-MiniLM-L6-v2`

Dataset		**#Doc.**	**#Rel.**	**Avg.# Ent.**	**Avg. #Triplets**	**#Negative Rate**
DocRED	train	3053	96	19.49	12.51	96.81%
DocRED_ext	train	3053	96	19.49	5.36	98.63%
Re-DocRED	dev	500	96	19.37	34.57	91.05%
	test	500		19.56	34.90	91.22%
DocGNRE	test	500	96	19.56	39.05	90.17%
DWIE	train(20%)	602	65	27.40	5.18	99.42%
	train(40%)				9.97	98.87%
	train(60%)				14.75	98.34%
	train(80%)				19.54	97.81%
	dev	98		28.42	26.78	97.15%
	test	99		26.49	24.83	96.96%

Table 6: Dataset Statistics. **#Doc.** indicates the number of documents in each dataset. **#Rel.** denotes the number of relation classes. **Avg.# Ent.** represents the average number of entities per document. **Avg. #Triplets** indicates the average number of annotated true triplets per document. **#Negative Rate** represents the ratio of negative triplets to the total number of triplets (negative + positive). For the DWIE dataset, the percentages in parentheses specify the sampling ratios of positive samples.

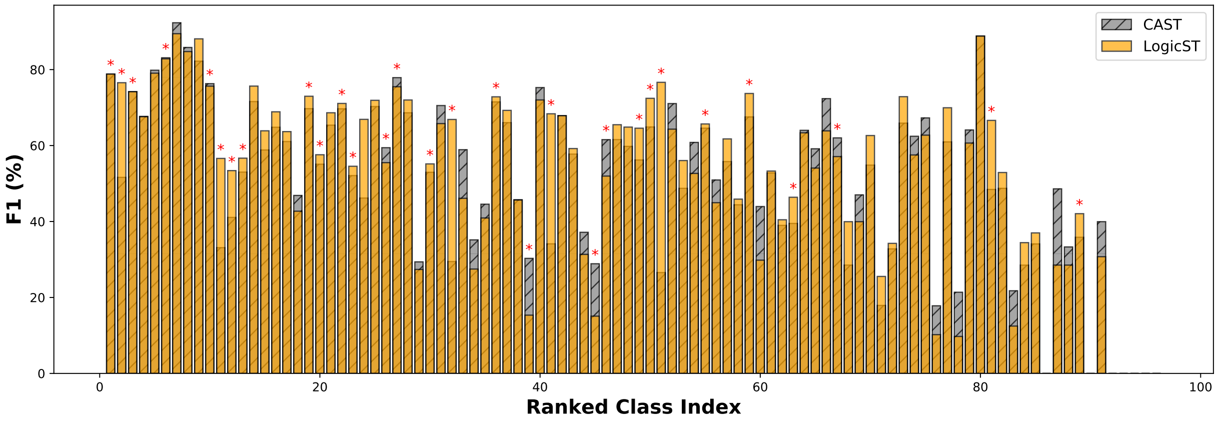


Figure 7: F1 scores (%) for each class (ranked by class frequency: left (high)→right (low)) on the test set of Re-DocRED of CAST and LogicST's best models, which are trained on DocRED. Columns marked with * are relation classes involved in more than 8 logical rules. Better view in color.

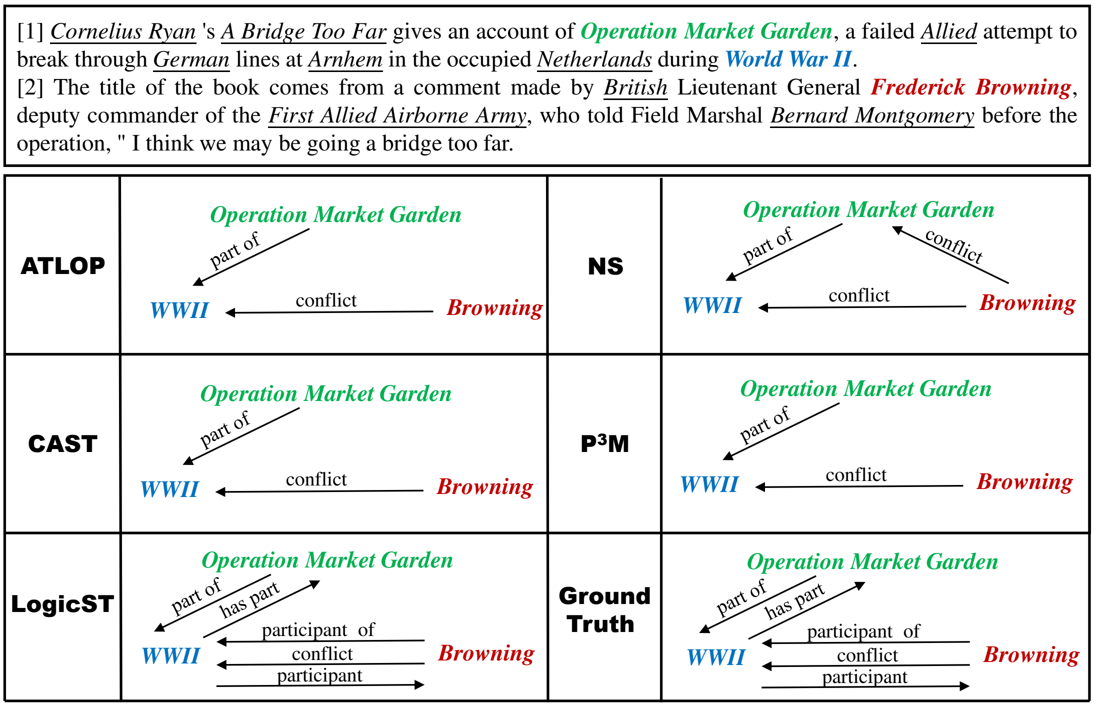
Antecedent	Consequent	Type
capital	located in	Implication
spouse^{-1}	spouse	Implication
spouse^{-1} $\wedge$ continent	continent	Composition
father^{-1} $\wedge$ mother	spouse	Composition
capital	$\neg$ spouse	Negation(h)
sister city	$\neg$ performer	Negation(t)

Table 7: Case study of logical rules. For simplicity, we hide the entity variables and use (h, r^{-1}, t) to represent (t, r, h).

model to find the most similar in-context samples from the incompletely annotated training set. The prompt template used is as follows:

```
Given a target relation type list
    and a document, entities and
    their mentions are marked in
    the document with numbered
    tags. Each entity is
    represented by a unique number
     enclosed in angle brackets.
    Please identify all valid
    given relation types between
    any two given entities in the
    document.

Target relation type list:
{'located in the administrative
    territorial entity': "In the '
    located in the administrative
    territorial entity' relation,
    the subject, a place, event,
    or item, resides or takes
    place in the object, an
    administrative region. Example
    : (Harvard University, located
```



[1] *Cornelius Ryan* 's *A Bridge Too Far* gives an account of ***Operation Market Garden***, a failed *Allied* attempt to break through *German* lines at *Arnhem* in the occupied *Netherlands* during ***World War II***.
[2] The title of the book comes from a comment made by *British* Lieutenant General ***Frederick Browning***, deputy commander of the *First Allied Airborne Army*, who told Field Marshal *Bernard Montgomery* before the operation, " I think we may be going a bridge too far.

Figure 8: Case study of relation extraction results on the test set of DocRED. For clarity, the named entities involved in this case study are marked in color, and other entities are underlined. The specific definition of the relations can be found in the original paper (Yao et al., 2019).

```
   in the administrative
  territorial entity, Cambridge,
   Massachusetts)."
......(all 96 relations and
  corresponding description and
  example) }

All non-duplicate valid "subject
  entity""relation type""object
  entity" triples in the
  document (output format: "
  entity ID""relation type name
  ""entity ID", e.g., <1>-
  country-<2>; one triple per
  line):
[<Entity ID>]-<Relation Name>-[<
  Entity ID>]
[<Entity ID>]-<Relation Name>-[<
  Entity ID>].
Please return the triplets in the
   specified format directly,
  without adding any additional
  information.

Here are some examples.
{ICL examples}

Document: {document to be
   extracted}

Triplets:
```

E Results on DocGNRE

Table 8 presents the performance of various models on the DocGNRE dataset (Li et al., 2023b), which is based on Re-DocRED and enhanced through distant supervision using ChatGPT, followed by human annotation for refinement. As shown in Table 8, LogicST demonstrates significant performance improvements over existing leading methods in both precision and recall. Notably, it surpasses the state-of-the-art P^{3}M model by **7.51%** in F1 score. These experimental results further validate the effectiveness of our proposed framework.

F Detailed Comparison with CAST

We plot the F1 scores of CAST and LogicST for all the classes in Figure 7, which indicates that LogicST surpasses CAST in most classes. Note that CAST samples pseudo labels for all classes, while LogicST only performs logical diagnosis on the classes involved in the rules. It can be seen that for classes involved in many logical rules (marked with *), LogicST usually has better performance.

Model	Precision	Recall	Ign F1	F1
GPT3.5	13.97	2.71	-	4.54
GPT3.5 + NLI	72.71	15.32	-	25.31
ATLOP	**91.56**	27.45	42.10	42.23
NS	69.23	42.34	51.72	52.54
SSR-PU	61.82	47.61	52.64	53.79
P^{3}M	67.93	50.95	57.24	58.23
CAST	65.06	51.46	56.44	57.47
LogicST	75.16	**58.41**	**65.05**	**65.74**

Table 8: Experimental Results on the test set of DocNGRE using BERT-base (excluding GPT-3.5 based methods) and the training set of DocRED (%).

G Case Study of Margin Bank

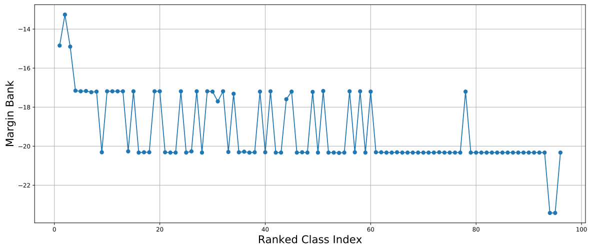


Figure 9: Margin Bank for each class (ranked by class frequency). Most of the margins are negative because the logits of the threshold class are higher.

We present the margin bank after training completion on DocRED in Figure 9. It is evident that the Margin^i of class i generally decreases as the frequency decreases, although there are some exceptions. This may be due to certain classes having fixed patterns, making them easier to recognize.

H Case Study of Relation Extraction Results

Figure 8 illustrates relation extraction cases for ATLOP and various frameworks built upon it, including NS, CAST, P^{3}M, and LogicST. This case study involves three logical rules: $(h, \texttt{part of}, t) \Rightarrow (t, \texttt{has part}, h)$, $(h, \texttt{conflict}, t) \Rightarrow (t, \texttt{participant}, h)$, and $(h, \texttt{participant}, t) \Rightarrow (t, \texttt{participant\_of}, h)$. The results demonstrate that ATLOP, CAST-ATLOP, and P^{3}M-ATLOP successfully extract two true positive relations but fail to infer their logical derivatives. In contrast, the NS-ATLOP method, which drops many true negative samples during training, introduces an additional false positive error. Notably, LogicST-ATLOP extracts all relevant facts using the same architecture and inference method as the other models, highlighting the effectiveness of incorporating logical rules as diagnostic tools to identify and correct pseudo-label errors. These findings underscore the advantages of using LogicST to enhance the robustness and accuracy of relation extraction tasks.