The Journal of Supercomputing (2025) 81:1618
https://doi.org/10.1007/s11227-025-08112-x

SELF: a multi-task learning framework for robust person-centric relation extraction

Hailin Wang[1,2,3] · Wentong Niu[1] · Hangyi Ren[1] · Jiahao Li[1] · Jingxuan Tian[1] · Dan Zhang[1,2,3]

Received: 9 August 2025 / Accepted: 22 November 2025

Abstract

Relation extraction (RE) plays a vital role in transforming unstructured text into structured knowledge. Person-centric relations are prevalent in large-scale user-generated content and real-world applications. Extracting such relations requires fine-grained semantic modeling and high computational efficiency. Despite their importance, person-centric relations remain underexplored in current RE research. Existing models often miss implicit semantic cues linked to person entities, leading to poor performance. To address this, we propose SELF, a multi-task learning framework, which enhances extraction performance for person-centric relationships. SELF introduces a simple, parameter-free auxiliary task that models the semantic space of person-related entities. This enables effective capture of sparse, crucial semantic information in person-centric contexts. Additionally, SELF incorporates a parallel hierarchical feature retainer module. This module adaptively maintains shallow and deep semantic representations from the auxiliary task, enriching the understanding of person-centric relationships. Extensive experiments on TACRED, Re-TACRED, and SemEval-2010 Task-8 show that SELF outperforms existing models in classifying person-centric relations and maintains computational efficiency in multi-task settings.

Keywords Person-centric relation extraction · Multi-task learning · Sparse entity label · Attention mechanism

1 Introduction

Relation extraction (RE) is a subtask of information extraction and a critical research area in natural language processing (NLP) [1]. Its goal is to extract semantic triplets from unstructured text. For example, given the sentence ‘Barack Obama served as the 44th president of the United States,’ RE extracts the triplet (Barack Obama,

Extended author information available on the last page of the article

Published online: 02 December 2025

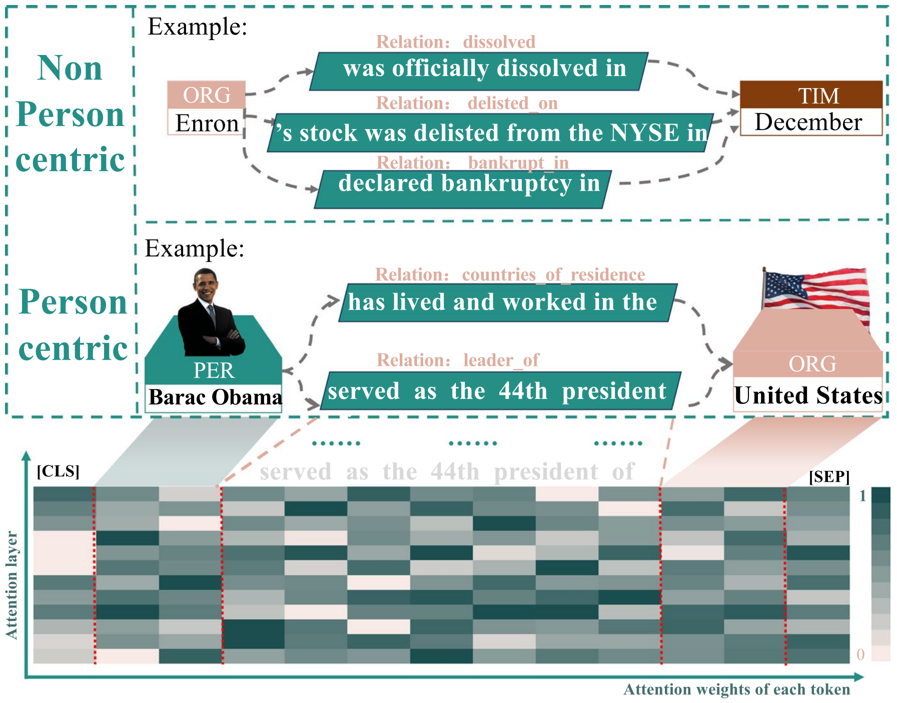


Fig. 1 Examples of person-centric and non-person-centric relations, and the corresponding attention weights of each token from a pre-trained language model. Person-centric instances typically involve person-related entities that provide strong implicit semantic cues, serving as valuable prior knowledge for relation classification. In contrast, non-person-centric relations often involve non-person entities that yield weaker or more ambiguous semantic signals. The bottom half of the figure is attention weights, derived from a BERT-base model [18] with 12 layers, predominantly neglecting the entity tokens at each layer

leader_of, United States). These structured relations play key roles in downstream applications. Such uses include social network analysis [2, 3], financial intelligence [4], medical text analysis [5], information retrieval [6, 7], relation reasoning [8, 9], security intelligence [10], and knowledge base construction [11, 12].

While general RE has been widely studied, a critical yet underexplored challenge lies in extracting *person-centric* relationships, where at least one argument involves a person entity. For instance, as illustrated in Fig. 1, the co-occurrence of entities like *Barack Obama* (PER) and *United States* (ORG) strongly suggests person-centric relations such as *leader_of* or *countries_of_residence*. These person-related entities often provide strong implicit semantic cues that serve as valuable prior knowledge for person-centric relation classification. In contrast, relations involving non-person entities, such as *Enron* (ORG) and *December 2001* (DATE), often exhibit weaker or more ambiguous signals, where multiple candidate relations may coexist. These person-centric relations (as shown in Table 10 of "Appendix") are prevalent in real-world corpora and are particularly crucial in many practical applications, such as building a person-centric knowledge graph at enterprise, which could promote understanding person entities' roles, affiliations, and actions, enabling finer-grained reasoning about social dynamics, organizational structures, and individual behavior patterns. When this process relies on large language models (LLMs) or heavyweight NLP, it incurs significant GPU or CPU costs [13]. Thus, designing an efficient and specific model for these applications has very practical significance. Despite the importance of person-centric relations, existing models often struggle to capture the implicit, fine-grained semantics embedded in person-related entity mentions, thereby limiting the accuracy of relation classification.

Recently, numerous researchers [14–17] have utilized pre-trained language models (PLMs) or LLMs for RE tasks. These methods use explicit entity markers or syntactic encoders to localize relation arguments. However, these markers cannot capture the deep, latent semantic cues associated with person entities, mainly in long-tailed or imbalanced datasets. Moreover, person-centric relations tend to exhibit sparse contextual clues that require reasoning beyond surface-level information to resolve, making it difficult for PLMs or LLMs to learn these distinctions from limited supervision. Although introducing auxiliary tasks within multi-task learning frameworks could provide additional training signals for model reasoning, PLMs fine-tuned on these tasks often fail to effectively absorb and retain the semantic knowledge introduced by auxiliary objectives, due to their tendency to overwrite earlier representations during gradient updates and the high computational cost associated with such training. This issue motivates the need for a dedicated mechanism to preserve and propagate such information across different model layers.

To address these challenges, we propose SELF, a novel multi-task learning (MTL) framework. SELF is designed to improve person-centric RE with minimal computational overhead. It introduces two key innovations. First, SELF adds a simple, parameter-free auxiliary task: Sparse Entity Position Prediction (SPP), guided by Implicit Entity Labels (IEL). This helps the model learn the semantic space of person-related entities without extra annotation or trainable parameters. Second, SELF uses a parallel Hierarchical feature Retainer (HR) module. This module integrates multi-layer semantic representations from PLMs and adaptively preserves both shallow and deep features learned from the auxiliary task. It supports parallel computation and maintains efficient execution. By retaining SPP-induced semantics across all layers, HR boosts the model's capacity for subtle, entity-centric relational understanding.

This work highlights the overlooked value of person-centric RE and offers an effective solution through the SELF framework. Our main contributions are:

- Proposing SELF, an MTL framework that introduces implicit entity positions into a simple, parameter-free SPP auxiliary task guided by IEL. This design enables the model to better capture the latent semantics of person-related entities without requiring additional annotation.
- Introducing a parallel HR module that adaptively integrates and preserves multi-layer entity-related features while maintaining computational efficiency. By explicitly retaining SPP-induced signals across all transformer layers, HR

enhances the model's ability to maintain fine-grained semantic cues essential for person-centric RE.

- Conducting extensive experiments on the TACRED dataset, a large-scale RE dataset rich in person-centric entity relationships. Results demonstrate that SELF significantly outperforms existing models in classifying person-centric relationships. Additional evaluations on the Re-TACRED and SemEval-2010 Task-8 (SemEval) datasets further verify the robustness of our approach.

2 Related work

Person-centric RE, which focuses on extracting relations where person entities participate as arguments, represents a widely occurring but underexplored subset of RE problems. While RE has been extensively studied in both general and domain-specific contexts, most existing work has not explicitly addressed the unique challenges posed by person-centric relations, such as implicit semantics, fine-grained entity distinctions, and contextual variability. This section reviews related work from general RE models as well as domain-specific applications where person-centric information is particularly critical.

2.1 General-purpose RE

The majority of prior RE research has focused on improving general-purpose models that target overall relation classification across diverse relation types. A dominant line of work has focused on enhancing entity representations, often through attention mechanisms such as multi-head attention [19, 20], which dynamically capture contextual interactions between entities and surrounding text. However, these approaches often suffer from attention dispersion [21], making it challenging to capture localized, entity-specific cues, especially for person-centric relations where subtle semantic signals are crucial. To better highlight entities, several methods have introduced auxiliary entity-aware features, including entity markers [22, 23], external descriptions [24], and entity masking strategies [25]. While these methods improve entity salience, they are often limited in fully capturing implicit person-related semantics. MTL has also been widely explored to address semantic sparsity and implicit relation cues. CREST [26] leverages multi-task signals to guide RE learning, while TemPrompt [27] incorporates temporal reasoning into RE via joint modeling. Recent work on classical Chinese by Tu et al. [28] applies multi-task decoding to address prediction gaps in long-distance dependencies. These MTL approaches demonstrate strong potential for addressing implicit semantics, aligning with our goal of enhancing person-centric RE through auxiliary learning objectives. In parallel, recent feasibility studies [29] show that large language models (LLMs), despite their general strengths in language understanding, are not yet robust enough for complex RE scenarios, further underscoring the need for targeted modeling strategies.

2.2 Person-centric RE in financial applications

The financial domain inherently involves human decision-makers, making person-centric relations crucial to interpreting market activities. Nevertheless, most financial RE systems focus on specific relationships, such as events or performance indicators, rather than person-centric semantics. For example, Deußer et al. [30] explored extracting KPIs from financial documents, where a word-level weighting scheme models the inherently fuzzy borders of entity pairs and their corresponding relations. Wu et al. [31] focused on Chinese Financial entity recognition and RE and proposed a mixed pattern with POS tagging to generate the quadruples from the unstructured financial text. Jabbari et al. [32] presented a domain-specific ontology for financial entities and relations in French news and created a corpus to build a knowledge base of financial relations. Recent benchmarks such as FinRED [33] and REFinD [34] further highlight that general-domain RE models are inadequate for finance due to differences in relation sets and semantics. Notably, both datasets contain substantial person-centric relations, underlining the demand for models that can robustly capture person-related meanings in financial texts.

2.3 Person-centric RE in legal applications

In the legal domain, human actors such as plaintiffs, defendants, and judges are central to legal discourse, making person-centric RE critical. Earlier work combined statistical and rule-based techniques for legal entity and RE [35]. Hendrycks et al. [36] released the CUAD dataset for contract analysis. Others have studied implicit legal relations such as criminal acts [37], clause dependencies [38], and cross-sentence legal reasoning [39]. Thomas and Sangeetha [40] used semi-supervised bootstrapping with OBIE for judicial fact extraction. Wang et al. [39] introduced Bi-FLEET, integrating clause element and classification tasks via bidirectional feedback. Xu et al. [38] proposed ConReader to model clause-level relations, including term-definition and cross-references. More recently, Wang et al. [41] proposed prompt-based methods to reduce label mismatch across legal subdomains. Although these models have achieved notable success within their respective legal datasets, most do not explicitly address person-centric semantic challenges, especially when entity roles extend beyond predefined legal positions. Furthermore, models optimized for legal texts often struggle to generalize to broader RE tasks, underscoring the importance of more flexible, person-aware modeling approaches.

2.4 Other application domains

Beyond finance and law, person-centric RE also plays a crucial role in many other fields where entities naturally center around individuals [42]. In the biomedical domain, patient-centric relations are critical for extracting clinical facts and treatment histories [43]. In social network analysis [3], social media platforms contain rich person-centric information in user-generated content, where relations are often informal, implicit, or multi-modal. In business intelligence [4], it helps track key individuals and their affiliations from unstructured texts such as news and reports. In information security and public opinion monitoring [44], person-centric relationships help reveal risk propagation networks. Other fields also benefit. In journalism [45], it aids in understanding event narratives. In scientific domains, it supports knowledge base construction [46] involving researchers, affiliations, and citation networks. These applications underscore the necessity of RE systems that can accurately model the semantics of person-related entities. Our proposed framework builds on these needs, aiming to provide a generalizable solution that enhances person-centric RE across various domains.

Despite the progress across various domains, current RE models still face significant limitations in effectively capturing person-centric semantic information, particularly when such cues are implicit or sparsely distributed. This situation underscores the need for a framework specifically designed to model person-related semantics in a more targeted and generalizable way. In the next section, we introduce our proposed framework, SELF, which leverages MTL and a hierarchical retainer mechanism to address these challenges.

3 Method

3.1 Task definition

This section describes our SELF framework, as shown in Fig. 2. The overall model architecture includes two tasks: the SPP task and the RE task.

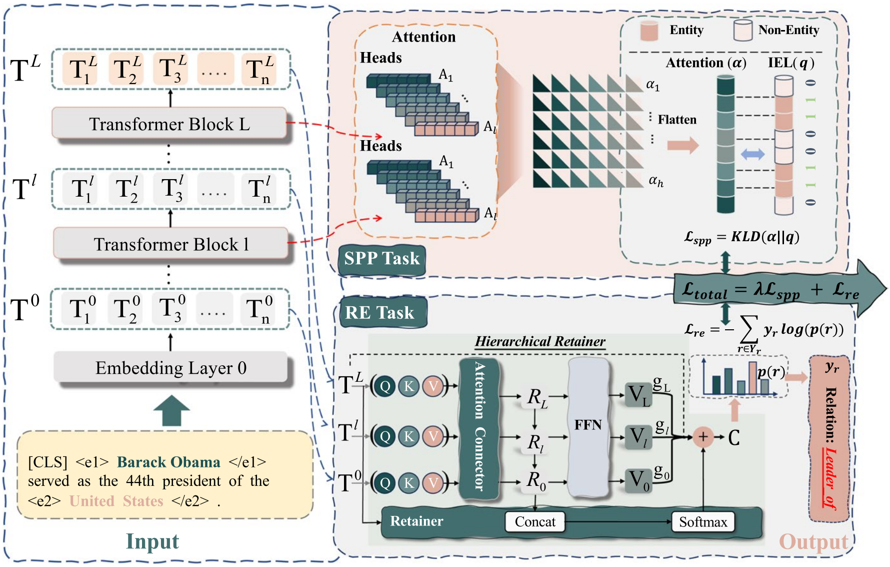


Fig. 2 Architecture of SELF. This model consists of two tasks: the SPP task and the RE task

SPP task The SPP task is proposed as an auxiliary objective to enhance the semantic representation of person-related entities by predicting the positions of sparsely distributed entities without any extra parameters. Specifically, the SPP task utilizes the IEL to measure the distributional discrepancy between predicted entity positions and the model's attention. Minimizing this discrepancy through backpropagation enables the model to focus more precisely on relevant token positions, thereby enhancing the model's ability to capture semantic cues associated with person-related content.

RE task Following prior work, we define the RE task as follows: given a sentence $\mathcal{S}$ comprising words $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$, and entities $\mathcal{E}_S = (e_s, e_o)$, the goal of RE is to predict the relationship $r \in \mathcal{R}$ for each entity pair (e_s, e_o), where $\mathcal{R}$ is a predefined set of relation types. For entity pairs without an explicit relationship, a special *No_relation* label is assigned. Note that different datasets adopt distinct labels for this category: TACRED uses *No_Relation*, Re-TACRED employs *NA*, and SemEval marks it as *Other*.

3.2 Framework incorporating the SPP auxiliary task

We introduce a parameter-free SPP task within the SELF framework to capture better the sparse and implicit semantic signals associated with person-related entities, without incurring the high computational cost on auxiliary objectives. This auxiliary task guides the model's attention toward entity-relevant tokens by leveraging IEL as supervision without any extra parameters. The SPP task encourages the model to align its attention distribution with the entity position distribution, enabling it to encode better subtle semantic cues crucial for RE in person-centric contexts. This section details how to construct IEL, formulate attention-based supervision, and define the loss function to support this auxiliary task.

3.2.1 Supervisory signal via IEL

To help the model better recognize semantically important regions in input text, especially those involving person-related entities, we introduce a weak supervisory signal called the IEL. Rather than labeling entity types or relations, IEL provides a binary indication of whether each token in the sentence belongs to an entity mention.

This simple signal acts as a lightweight yet effective form of supervision. While it does not specify the type or role of the entity, it implicitly guides the model to focus on token positions that frequently carry semantic significance in RE, especially in person-centric cases. For example, person names, regardless of specific roles, tend to appear in informative positions, and helping the model learn to localize them can be helpful in downstream tasks.

Based on IEL, we design the auxiliary task of SPP, where the model learns to predict the positions of all entity tokens in a sentence. Although the task targets position prediction, our ultimate goal is to guide the model toward learning the semantic distribution of different entity types, especially person entities, based on their contextual environments. This auxiliary signal enhances the encoder's sensitivity to person-centric semantics, providing indirect supervision that benefits the downstream RE task. Hence, IEL serves as a supervisory signal that reinforces the model's ability to capture implicit, position-sensitive semantic cues associated with person-related entities, as shown in Fig. 3.

To determine the implicit entity position distribution, we define the IEL as a binary vector $\mathcal{Q} \in \mathbb{R}^n$, which signifies whether each word $x_i \in \mathcal{X}$ in a sentence is an entity token. Specifically, for IEL, if x_i is an entity token, we designate $\mathcal{Q}_i$ as 1; otherwise, it is designated as 0. This binary vector $\mathcal{Q}$ effectively identifies the presence of all entity tokens.

Subsequently, the vectors are normalized to derive the implicit entity position distributions.

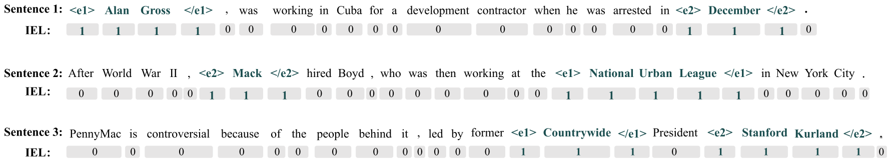
$$q = \frac{\mathcal{Q}}{\mathbf{1}^{\mathrm{T}}\mathcal{Q}} \tag{1}$$

where $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^n$ represents an all-ones vector. $q \in \mathbb{R}^n$ represents the probability distribution of entity positions in the sentence.

3.2.2 Supervision via attention alignment

Leveraging the previously defined IEL as a supervisory signal, the SELF framework must identify salient features to guide supervision. One promising candidate is the attention mechanism, which can highlight task-relevant tokens in NLP [47] . This function aligns well with the SPP task's objectives, which aim to identify key token positions and learn a distribution that captures their positional importance. Therefore, in this study, we evaluate the SELF framework using the multi-head attention mechanism employed in Transformer-based models [18].

Multi-head attention, constructing a Transformer block, is a pivotal element in PLMs and LLMs, generating a cross-dependency importance vector for each token across multiple layers. However, the objective is to derive a position-centric attention weight that closely aligns with the position distribution of the IEL. To achieve this, attention weights are extracted from different Transformer layers, and their average is computed by:



Sentence 1: <e1> Alan Gross </e1> , was working in Cuba for a development contractor when he was arrested in <e2> December </e2> .
IEL: 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0

Sentence 2: After World War II , <e2> Mack </e2> hired Boyd , who was then working at the <e1> National Urban League </e1> in New York City .
IEL: 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0

Sentence 3: PennyMac is controversial because of the people behind it , led by former <e1> Countrywide </e1> President <e2> Stanford Kurland </e2> .
IEL: 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 1 1 1 1 0

Fig. 3 Examples of IEL. These IELs are used to mark the specific positions of entities in sentences

$$\begin{aligned} a^m &= \frac{\sum_{l=1}^{L} A_l}{L} \\ a^h &= \frac{\sum_{j=1}^{H} a_j^m}{H} \end{aligned} \tag{2}$$

where $A_l \in \mathbb{R}^{H\times n\times n}$ represents the attention weights of the l-th Transformer block, and H denotes the number of attention heads, n is the length of the sentence. $a^h \in \mathbb{R}^{n\times n}$ is the average attention score across each layer and each attention head. To obtain a probability distribution vector consistent with the dimensionality of the IEL, a^h is compressed along the vertical axis and normalized using the Softmax function. The specific computation is as follows:

$$\begin{aligned} \alpha' &= \frac{1}{n}\sum_{i=1}^{n} a_{ij}^h \\ \alpha &= Softmax(\alpha') \end{aligned} \tag{3}$$

where $\alpha = \{a_1, \ldots, a_n\} \in \mathbb{R}^n$ represents the model's attention scores concerning the sentence. Here, n corresponds to the length of the input token sequence. $a_i \in \alpha$ represents the attention score of the i-th token in the sentence.

3.2.3 Loss in auxiliary task

The auxiliary SPP task reduces the distributional discrepancy between the model's attention and the IEL. The Kullback–Leibler Divergence (KLD) measures the discrepancy between two distributions. The following formula defines the KLD:

$$KLD(P||Q) = \sum_{j=1}^{m} P(x_j)\log\left(\frac{P(x_j)}{Q(x_j)}\right) \tag{4}$$

where P and Q represent two different distributions. The SPP task employs the KLD to compute the distributional discrepancy loss between the model's attention (α) and the IEL(q):

$$\mathcal{L}_{spp} = KLD(\alpha||q) \tag{5}$$

The SPP task enhances entity position semantics through position prediction and provides crucial person-related semantic guidance for the following encoder through supervision via attention alignment (Sect. 3.2.2). We introduce a retainer module to integrate further and exploit the hierarchical semantic signals captured across different layers of the following encoder. In the next section, we describe how the retainer module leverages the guidance from SPP to perform dynamic cross-layer feature preservation, enabling fine-grained representation for person-centric relations.

3.3 Framework incorporating the HR module

While the SPP task explicitly enhances the model's awareness of person-related entities by injecting supervision into attention patterns, its influence tends to weaken in deeper encoder layers due to gradient vanishing and semantic drift. As a result, the entity position cues and person-related semantics introduced by SPP may not be fully preserved throughout the encoding process.

To mitigate this limitation, we design the HR module that serves two complementary purposes: (1) it builds a cross-layer connection to maintain the influence of the SPP task across all encoder layers using a parallel attention mechanism, and (2) it dynamically integrates multi-level semantic representations to maintain a refined and context-aware feature space for robust relation classification. In this way, the HR module ensures that the supervision signals from SPP are not only injected early in the model but also propagated and selectively emphasized during relation prediction.

The following subsections present the HR module in detail, outlining how to generate hierarchical contextual features and how HR adaptively selects and fuses them under SPP guidance to preserve person-related semantics throughout the encoding process.

3.3.1 Text encoder

PLMs typically utilize stacked Transformer blocks as their core architecture for text encoding. Given a token sequence X, it first passes through the embedding layer to obtain the initial feature representation T^0, which is then processed through L layers to produce the final output T^L. The following formula can express it:

$$\begin{aligned} T^0 &= Embedding(X) \\ T^l &= TransformerBlock(T^{l-1}) \end{aligned} \tag{6}$$

where $T^l \in \mathbb{R}^{n\times d}$ represents the text semantic feature extracted from the lth layer. n is the sentence length and d is the embedding dimension. Meanwhile, the transformer block generates the A_l in Eq. 2, representing the attention weights of the lth Transformer block, which is supervised by the SPP task.

3.3.2 Hierarchical feature retainer

We propose an HR module to reinforce and retain the implicit person-related semantic signals introduced by the SPP task across different encoder layers. It consists of two key components: an Attention Connector (AC) to align multi-layer semantics and a Retainer to adaptively integrate the most informative representations across layers under SPP guidance.

Attention Connector Since the introduction of the SPP task, the final-layer output T^L is most influenced by attention weights. To propagate this influence throughout the encoder, we aim for the attention and representations of all intermediate layers to be directly affected by the final layer. We design the AC to achieve this, which links each intermediate layer with the final-layer output.

Specifically, by treating the final-layer output $T^L \in \mathbb{R}^{n\times d}$ as the query and an intermediate layer $T^l \in \mathbb{R}^{n\times d}$ as key and value, the AC computes attention heads and produces aligned features R_l as follows:

$$\begin{aligned} M_T^{(l,m)} &= Softmax(\frac{[T^L W_q^m][T^l W_k^m]^T}{\sqrt{d/M}}) \\ head_m &= M_T^{(l,m)}[T^l W_v^m] \\ R_l^{'} &= [head_1;head_2;\ldots;head_M]W_h \\ R_l &= LayerNorm(R_l^{'}) \end{aligned} \tag{7}$$

where $\{W_q^m, W_k^m, W_v^m\} \in \mathbb{R}^{d\times d/M}$ are the weight matrices specific to $head_i$ for generating key, query, and value vectors. $M_T^{(l,m)} \in R^{n\times n}$ represents the connector attention map of the l-th layer. $W_h \in \mathbb{R}^{d\times d}$ aggregates all heads. And all heads can be parallel operated.

Next, the features of the final and current layers are jointly fed into a Fully Connected Feed-Forward Network (FFN). The specific formulas are as follows:

$$\begin{aligned} V_l &= FFN([T^L;R_l]) + T^L \\ &= h([T^L;T_l]W_1^l + b_1^l)W_2^l + b_2^l + T^L \end{aligned} \tag{8}$$

where $\{W_1^l, W_2^l\} \in \mathbb{R}^{2d\times d}$ are the transformation matrix for the l-th layer. $h(*)$ is the activation function.

Retainer We apply a discriminative retainer component to select the most informative layer-wise representations adaptively. Specifically, this component assesses whether the semantic information encoded at various layers aligns well with the feature representations from the final layer. The retainer's input is from the above CA, a hierarchical set of textual features R_T and V_T:

$$\begin{aligned} R_T &= \{R_0, R_1, \ldots, R_l, R_L\} \\ V_T &= \{V_0, V_1, \ldots, V_l, V_L\} \end{aligned} \tag{9}$$

All texts embedding in R_T are concatenated and fed into the retainer, a linear input transformation followed by a Softmax layer. Then, the retainer predicts the importance weights for each layer by:

$$G = \sigma(R_T, T^L) = Softmax([T^L;R_0;R_1;\ldots;R_L]W_g) \tag{10}$$

where $W_g \in \mathbb{R}^{((L+2)\times d)\times(L+1)}$ is the trainable projection matrix and the output $G \in R^{n\times(L+1)}$ preserves the predicted retainer weights for diverse textual tokens.

The retainer weight matrix G is employed to maintain the V_T from different layers, resulting in a more refined feature representation encompassing all layers. The specific formulas are as follows:

$$\begin{aligned} C &= \sum_{l=0}^{L} \sigma(R_T, T^L)_l \cdot V_l + \sigma(R_T, T^L)_{text} \cdot T^L \\ &= \sum_{l=0}^{L} \boldsymbol{G}_{[:,l+1]} \cdot V_l + \boldsymbol{G}_{[:,0]} \cdot T^L \end{aligned} \tag{11}$$

This retainer process generates a rich, context-aware representation that adaptively integrates shallow and deep semantic features for robust RE. Specifically, $\sigma(R_T, T^L)_l$ and $\sigma(R_T, T^L)_{text}$ represent the integration retainer weights assigned to the l-th layer and the final layer T^L, respectively. These weights are organized into a unified gating vector $\boldsymbol{G} = \{G_{[:,0]}, G_{[:,1]}, \dots, G_{[:,L+1]}\}$, which governs the contribution of each layer to the final representation. Through this dynamic retainer, the model learns to emphasize the most informative layer-wise features while suppressing redundant or noisy representations, which is rarely influenced by the SPP task.

At the training phase, SELF requires additional computational resources compared to conventional single-task RE models. The HR module, in particular, introduces extra cost due to its cross-layer attention and retainer operations. These computations are parallelized across GPUs to maintain training efficiency and scalability. In contrast, the SPP task remains lightweight and incurs no additional parameter or computational overhead, making it a cost-effective auxiliary mechanism to improve PLMs' semantic sensitivity without requiring expensive task-specific fine-tuning.

3.3.3 Loss in RE task

We use the final feature representation for relation prediction. A common approach is to concatenate the features corresponding to the entity positions and the [CLS] position, pass them through a linear layer, and then apply the Softmax function to predict the relationship probabilities:

$$P_r = Softmax([C_{[CLS]};C_{<e1>};C_{<e2>}]W_r) \tag{12}$$

where $W_r \in \mathbb{R}^{3d \times R}$ is the trainable projection matrix. P_r denotes the predicted probability distribution over all relation classes, where the values sum to one. Subsequently, we compute the loss for the RE task using cross-entropy:

$$\mathcal{L}_{re} = -\sum_{r \in Y_r} y_r \cdot \log(p(r)) \tag{13}$$

where $p(r)$ refers to the predicted probability of a specific relation r obtained from P_r, concretely, $p(r) = P_r[r]$. Meanwhile, y_r represents the ground-truth label of relation r, where $y_r = 1$ if r is the correct relation and $y_r = 0$ otherwise. For example, if relation r is the ground-truth class and is predicted with probability a in P_r, then $p(r) = a$ and $y_r = 1$.

3.4 Total loss construction

SELF adopts the parallel-shared MTL paradigm [48], jointly optimizing the SPP and RE tasks. In MTL, the choice of optimization strategy is as critical as the model architecture itself. A common approach is to linearly combine the task-specific losses into a unified global loss, enabling the training of the entire model using standard optimization algorithms. The total loss is defined as follows:

$$\mathcal{L}_{totall} = \mathcal{L}_{re} + \lambda \cdot \mathcal{L}_{spp} \tag{14}$$

where λ is a hyperparameter that balances the importance of the SPP task relative to the RE task.

4 Experiments

4.1 Experimental setting

Datasets To evaluate the effectiveness of our framework for classification involving person-related entities, we conduct experiments on three benchmark datasets: TACRED [49], RE-TACRED [50], and SemEval [51]. TACRED and RE-TACRED provide comprehensive coverage of person-centric relations, while SemEval is an auxiliary benchmark for assessing generalizability. Detailed dataset statistics are summarized in Table 1.

- **TACRED** is a large-scale RE dataset in which each sentence is annotated with entity types and the corresponding relation. It includes a broad range of person-related entities and relation types. For example, *per:cities_of_residence* indicates a relation between a person and a location, while *org:founded_by* reflects a relation between a person and an organization. The dataset comprises 42 relation categories, including a *No_Relation* class.
- **Re-TACRED** is a refined version of TACRED, which corrects label noise and ambiguous annotations. Specifically, it addresses 23.9%
- **SemEval** is a widely used relation classification benchmark comprising 8000 training and 2717 test instances. Each instance contains two annotated noun phrases with directional semantic relations, so (e1, e2) and (e2, e1) are treated

Table 1 Statistical Overview of the Datasets

Dataset	#Train	#Dev	#Test	#Rel	#Ent	#PER	PER%
TACRED	61,128	22,631	15,509	42	136,248	58,303	43%
Re-TACRED	58,465	19,584	13,418	40	116,930	51,734	44%
SemEval-2010 Task-8	8000	–	2,717	19	–	–	–

Train, Dev, Test: number of instances in each datasets. Rel: number of relation types; Ent: number of total entity mentions; PER: number of person entities; PER%: proportion of person-related entities among all entities as distinct. This results in 19 relation classes (2 × 9 directional relations and one *Other* class).

As shown in Table 1, both TACRED and Re-TACRED exhibit many person-related entities, validating their suitability for evaluating person-centric RE models.

Metrics To evaluate the model's performance, we adopt precision, recall, and F1-score as the primary evaluation metrics.

- **Precision** (Prec) measures the proportion of correctly predicted positive instances among all instances predicted as positive.
- **Recall** (Rec), also known as sensitivity, indicates the proportion of actual positive instances the model correctly identifies.
- **F1-score** (F1) is the harmonic mean of precision and recall, balancing both metrics and serving as the primary performance indicator in all experiments.

Baselines We compare our model against a series of representative and fully reproducible baselines across three benchmark datasets. Although some of these baselines are not the most recent, they were deliberately selected based on three considerations shown in Table 2:

The baselines include:

- **Att-Bi-LSTM**(Att) [52]: A bidirectional LSTM model with a soft attention mechanism that assigns different weights to each token in the sentence.
- **R-BERT** [22]: A BERT-based model that marks entities explicitly and uses the representations of entity markers for classification.
- **RoBERTa-marker** (RoBERTa) [23]: An enhanced BERT variant with entity markers, using RoBERTa as the encoder.
- **Causal** [25]: A causal reasoning-based method that incorporates external descriptions of entities into the input sentence to assist in relation classification.
- **TCohPrompt** [53]: A model that leverages prompt tuning and entity coherence to inject extra entity-related semantics.

Table 2 Justification for baseline selection

Criterion	Justification
Reproducibility	All baseline models provide open-source implementations; this ensures consistent, transparent, and verifiable experiments
Person-Centric Relevance	Each method uses attention or entity-aware design, making them appropriate for capturing person-related semantics
Methodological Diversity	The selected baselines cover various paradigms, including BiLSTM, PLMs, prompt tuning, and LLMs, supporting a broad evaluation

- **PTR** [24]: A recent method that incorporates human-defined label-specific rules into a prompt-based tuning framework, bridging the gap between symbolic cues and PLMs learning.
- **GPT-2** [54]: A representative large language model used here as a baseline for fine-tuning comparison.

Training Details The SELF model is trained based on the RoBERTa-base architecture [55], with configurations tailored to the characteristics of each dataset. Across all datasets, we set the batch size to 32 and the learning rate to 3×10^{-5}. However, the number of training epochs and the hyperparameter λ vary depending on the dataset. The detailed hyperparameter settings are summarized in Table 3. These parameters were selected through grid search and fine-tuning on the respective development sets to achieve optimal performance.

All experiments were conducted on a high-performance computing cluster equipped with NVIDIA RTX 3090 GPUs. We employed 2 GPUs in parallel and implemented distributed training using PyTorch's DistributedDataParallel (DDP) framework. Each GPU is equipped with 24 GB of memory, which enabled us to train deep Transformer-based models with relatively large batch sizes and to efficiently perform multi-task optimization (SPP and HR). This environment significantly reduced training time, facilitated large-scale hyper-parameter tuning , and ensured reproducibility across repeated experiments.

4.2 Main results

Tables 4 and 5 present a comprehensive performance comparison between our proposed SELF model and several baselines on the TACRED, Re-TACRED, and SemEval datasets. Across all relation types, SELF achieves generally competitive performance; its significant advantage emerges in person-centric RE, where it consistently outperforms other models in terms of F1.

In Table 4, which reports category-level results for selected person-centric relations in TACRED, SELF achieves the highest scores in multiple key categories. For example, SELF achieves an F1 of 81.38% on *per:employee_of*, representing a +1.47% improvement over the next-best model TCohPrompt. Similarly, for *per:countries_of_residence*, SELF obtains 60.44%, surpassing competitors by +2.56%. Meanwhile, to show this superiority, Fig. 4 complements these findings by visualizing the F1 values of all relation categories on TACRED and Re-TACRED. For clarity, this figure separates person-centric relations (inner arc) from

Table 3 Detailed hyperparameter settings for SELF across different datasets

Dataset	Batch size	Learning rate	Epochs	λ
TACRED	32	3×10^{-5}	10	0.4
Re-TACRED	32	3×10^{-5}	15	0.7
SemEval	32	3×10^{-5}	20	0.2

Table 4 Comparison of F1 across models for partial person-centric relations in the TACRED dataset

Person-centric relations (F1%)	Att	RoBERTa	R-BERT	PTR	Causal	TCohPrompt	SELF
per:employee_of	71.66	74.07	77.82	77.98	75.45	79.91	**81.38**$_{(+1.47)}$
per:countries_of_residence	16.47	42.71	51.16	52.12	49.54	57.88	**60.44**$_{(+2.56)}$
per:cities_of_residence	66.66	69.12	66.43	63.97	66.2	66.66	**70.46**$_{(+3.80)}$
org:founded_by	0	83.87	82.25	80.29	85.26	86.76	**89.06**$_{(+2.30)}$
org:top_members/employees	85.75	85.76	86.66	85.18	85.23	85.46	**87.55**$_{(+0.89)}$
per:parents	0	83.87	83.66	56.84	84.41	82.11	**85.89**$_{(+1.48)}$
per:other_family	0	48.19	57.44	20.75	60.86	55.61	**64.34**$_{(+3.48)}$
per:state_of_residence	63.76	60.65	61.66	61.53	61.78	56.57	**70.76**$_{(+7.00)}$
per:schools_attended	66.66	68.08	68.08	69.23	78.43	77.85	**79.24**$_{(+0.81)}$

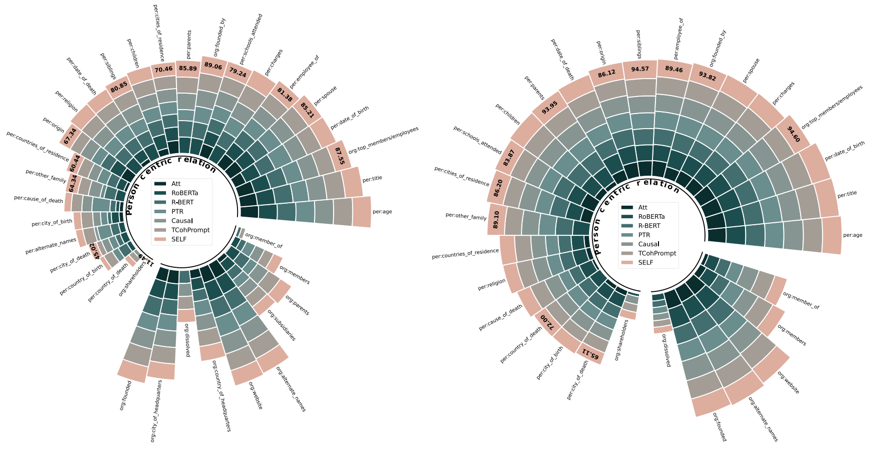
Table 5 Overall results of the model on different datasets

Models	TACRED			Re-TACRED			SemEval		
	Prec (%)	Rec (%)	F1 (%)	Prec (%)	Rec (%)	F1 (%)	Prec (%)	Rec (%)	F1 (%)
Att [52]	60.25	49.83	54.55	72.52	73.61	73.06	81.44	85.32	83.34
RoBERTa [23]	73.28	65.08	68.93	88.34	86.04	87.18	87.08	89.49	88.27
GPT2 [54]	54.80	35.84	43.34	61.08	52.49	56.46	76.77	81.35	78.99
PTR [24]	65.67	70.84	68.15	86.47	86.26	86.37	-	-	-
Causal [25]	71.65	68.66	70.12	89.02	89.02	89.02	87.64	88.68	88.16
R-BERT [22]	72.02	65.80	68.77	87.53	86.31	86.92	88.37	88.37	88.37
TCohPrompt [53]	73.44	71.99	72.71	92.17	88.89	90.50	94.12	93.30	93.71
SELF	71.25	70.28	70.76	89.31	89.49	89.40	88.02	89.30	88.65

non-person-centric ones in the charts. Bars where SELF leads are in bold, clearly show that most of its advantages concentrate in person-centric categories.

While SELF's overall performance across all relation types is competitive but not the highest, its advantage becomes most pronounced in person-centric RE. As shown in Table 5, SELF achieves an overall F1 of 70.76% on TACRED, comparable to Causal (70.12%) and RoBERTa (68.93%), yet clearly surpasses them in person-centric categories. The heightened category sensitivity enables SELF to capture subtle semantic cues between person-related entities, a capability less evident in other models.

Although SELF achieves strong performance on person-centric relations, Table 5 shows that TCohPrompt obtains a higher overall F1. This advantage stems from its injection of entity coherence prompts and external entity-level semantic priors, which are especially useful when relevant entity semantics cannot be recovered from the context alone. In contrast, SELF is designed to strengthen contextual attention to entity positions and preserve these signals across layers; it therefore excels when person-centric cues are present in surrounding text but may be less effective than



(a) Results on TACRED

(b) Results on Re-TACRED

Fig. 4 Category-level performance comparison on **a** TACRED and **b** Re-TACRED. Each circular stacked bar chart visualizes the F1 values of different models across all relation types. Relations are divided into person-centric relations (labeled along the inner arc in the figure) and non-person-centric relations (not labeled in the figure for display clarity), separated by a visible gap in the ring. Bars are stacked by model, with heights representing model-specific performance. Bars where the SELF model outperforms all other models in person-centric relations are annotated in bold

prompt-based entity description methods when the needed semantic priors are missing from the sentence.

5 Analysis and discussion

5.1 Effectiveness of SPP with IEL

To evaluate the effectiveness of the proposed SPP task guided by IEL, we analyze its influence on attention allocation within PLMs. Specifically, we visualize the attention maps for several representative examples from the TACRED dataset, as illustrated in Fig. 5.

The figure compares attention distributions from two settings: the baseline model without SPP supervision and the proposed model enhanced by IEL guidance. Across all examples, the baseline attention maps reveal a tendency to over-concentrate on the sentence's final tokens, with broadly dispersed focus elsewhere. This pattern suggests that, without explicit supervision, the model fails to highlight entity-relevant regions effectively.

In contrast, when the model is trained with the auxiliary SPP task using IEL as supervision, the attention distribution becomes noticeably more focused around entity-related tokens, especially those related to persons, suggesting that the model is better guided toward semantically informative regions of the sentence. Notably, in sentences 1–3, attention shifts from irrelevant tokens to those marking or surrounding entities such as *Mack* or *Thai Petrochemical Industry*.

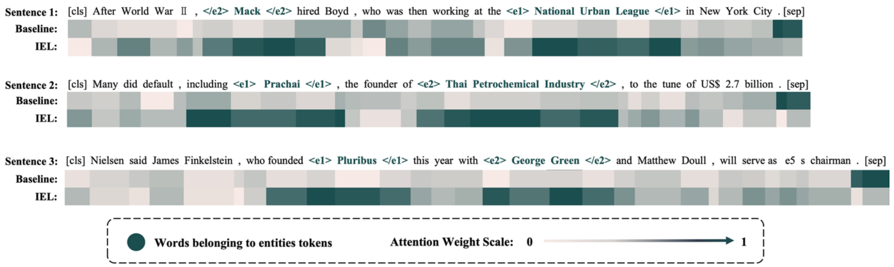


Fig. 5 Attention visualization for person-related entities on the TACRED dataset. The baseline represents the model without the SPP task, while IEL denotes the model trained with SPP-guided supervision to enhance attention alignment with entity positions. The comparison highlights how SPP encourages the model to focus more precisely on person-related tokens

These results demonstrate that IEL-guided supervision enables the model to learn a more meaningful positional attention distribution, thereby enhancing its ability to capture fine-grained semantic signals. Consequently, this improves entity representation and the classification accuracy for person-centric relations.

5.2 Effectiveness of HR module

This section investigates how the HR module preserves and amplifies the entity-related information introduced by the SPP task. Although HR consists of the AC and Retainer components, they operate as an integrated mechanism; therefore, we analyze their effect jointly. The goal is to assess whether HR can maintain consistent entity sensitivity across layers and prevent the gradual loss of SPP-induced semantic cues.

To investigate this, we visualize the retainer weights learned by HR under two configurations: without SPP supervision and with SPP supervision. Figure 6 presents five representative examples, each shown as a pair of heatmaps (left: without SPP, right: with SPP). In these heatmaps, rows correspond to Transformer layers, columns to token positions, and darker cells indicate stronger preference for incorporating semantic features from a specific layer for a given token. This design allows direct observation of how SPP influences HR's token-layer selection patterns.

Across all examples, the left-side heatmaps (without SPP) display diffuse attention distributions, with no consistent emphasis on entity-centric positions. In contrast, the right-side heatmaps (with SPP) exhibit markedly sharper focus on semantically important tokens, such as *Mark Fisher* and *Dayton Daily News* in Fig. 6a. This shift indicates that SPP supervision guides HR to prioritize entity-related positions and preserve them consistently across layers.

These visualizations collectively demonstrate that the HR module, when jointly optimized with the SPP task, effectively learns to retain discriminative token-layer combinations critical for relation prediction. By adaptively aligning shallow and deep semantics throughout the encoding process, HR enhances the model's robustness in capturing subtle person-centric relationships.

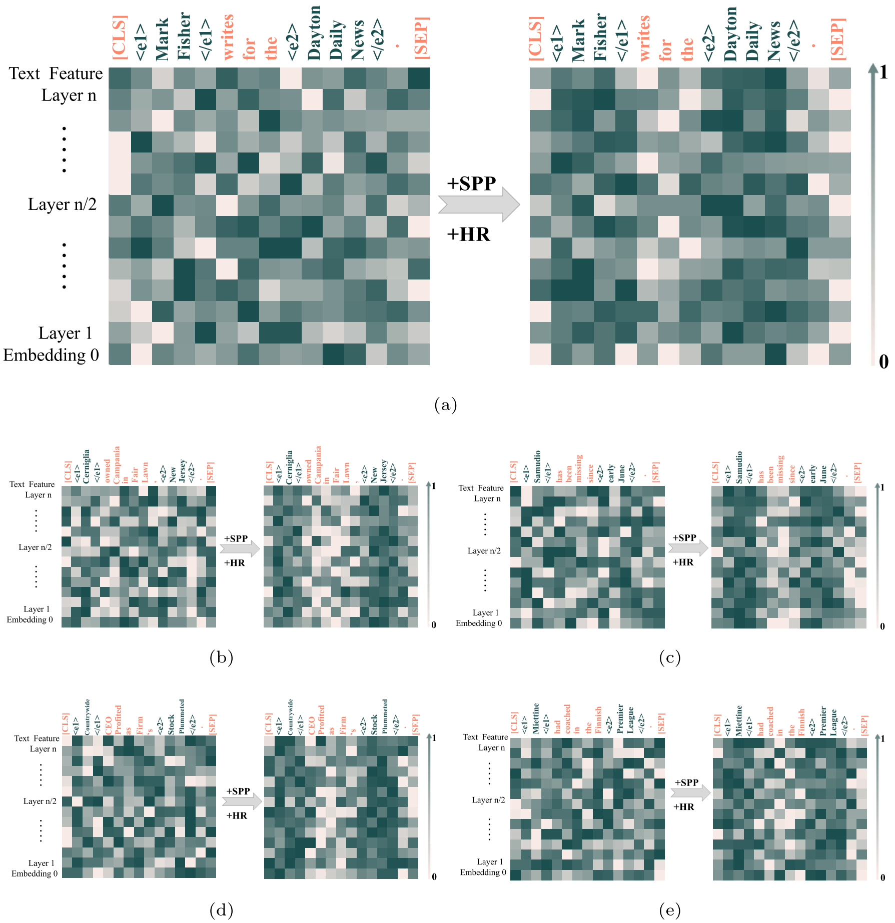


Fig. 6 Retainer heatmaps from the HR module before (left) and after (right) incorporating SPP supervision. Darker cells indicate higher retainer weights. Without SPP, attention is diffuse across tokens and layers; with SPP, it concentrates on entity-related tokens (e.g., Mark Fisher, Dayton Daily News in (**a**)), showing HR's enhanced ability to preserve entity-focused features across layers

5.3 Ablation study on person-centric RE

Given the focus of this work on enhancing the extraction of person-centric relations, we design targeted ablation experiments to evaluate the contributions of different components within the SELF framework. All variants are assessed primarily on person-centric relation categories to ensure alignment with the central objective of our study.

We construct the following model variants: (1) *Baseline*: The original architecture without the SPP or HR modules, using only the encoder's final-layer features for relation classification. (2) +*SPP*: Extends the Baseline with the SPP module.

(3) *+HR*: Extends the Baseline with the HR module to fuse multi-layer encoder features. (4) *SELF*: The complete model integrating both SPP and HR.

Table 6 reports the results on the TACRED, Re-TACRED, and SemEval datasets. SELF consistently achieves the best performance, with particularly notable gains on person-centric relations in TACRED and Re-TACRED, improving F1 by nearly two percentage points over the baseline. Introducing either SPP or HR alone also yields clear improvements, with SPP showing a slightly stronger effect, indicating that guiding the model to attend to implicit entity positions is especially beneficial for RE. Nevertheless, the full SELF model delivers the most consistent gains, confirming the complementary nature of SPP and HR. On SemEval, where person-centric relations are less dominant, the improvements are smaller but still demonstrate the method's generalizability.

While the HR module improves overall F1, we note a slight drop in precision on TACRED and SemEval. This phenomenon can be attributed to the following reasons: (1) The advantage of the HR module depends on the SPP signals. The main strength of HR lies in its ability to preserve and adaptively integrate hierarchical representations so that the entity position supervision from SPP can be effectively retained at deeper layers. Without such explicit guidance, HR merely fuses intermediate representations in a mechanical manner, which may not always be beneficial and can even introduce instability. (2) Impact of annotation noise. TACRED has been reported to contain a considerable amount of noisy or erroneous labels. In this setting, HR may inadvertently preserve and propagate these noisy signals when aggregating multi-layer representations, which diminishes its effectiveness. By contrast, Re-TACRED corrects such labeling errors, providing cleaner supervision and allowing HR to stably enhance performance. (3) Sample sparsity in SemEval. SemEval is relatively small in scale, and several relation categories are sparsely represented. In this context, the additional complexity introduced by HR may increase the risk of overfitting or amplify data sparsity effects, leading to marginal performance degradation.

These findings indicate that HR alone may not always be beneficial under noisy or low-resource conditions. However, when combined with SPP in the full SELF framework, HR effectively retains the explicit entity position signals provided by SPP, thereby yielding consistent and significant improvements across datasets.

Table 6 Performance of the SELF model on person-centric categories under different ablation settings

Modules	TACRED			Re-TACRED			SemEval		
	Prec (%)	Rec (%)	F1 (%)	Prec (%)	Rec (%)	F1 (%)	Prec (%)	Rec (%)	F1 (%)
Baseline	91.82	67.72	77.95	96.19	87.72	91.76	87.64	87.28	87.46
+HR	91.27	68.97	78.57	96.93	88.21	92.37	87.12	88.47	87.79
+SPP	93.47	68.71	79.20	97.67	88.95	93.11	87.87	88.59	88.23
SELF	**93.93**	**69.10**	**79.63**	**98.04**	**89.77**	**93.72**	**88.02**	**89.30**	**88.65**

SemEval is tested under all categories

5.4 Case study

We conduct a fine-grained case study on selected categories from the Re-TACRED dataset to further assess SELF's effectiveness in person-centric RE. Specifically, we compare the F1 of multiple models on four representative relations: *per:cities_of_residence*, *per:employee_of*, *per:schools_attended*, and *org:founded_by*, as shown in Fig. 7.

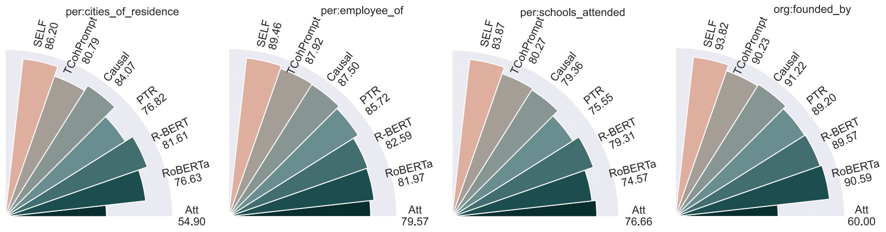
SELF consistently achieves the highest performance across all four relations. For example, on *per:cities_of_residence*, it surpasses the second-best model, Causal [25], by 2.13%. On *per:employee_of* and *per:schools_attended*, SELF outperforms TCohPrompt [53] by 1.54% and 3.50%, respectively. These relation types often demand nuanced reasoning over contextual associations between persons, locations, or organizations. SELF's advantage stems from the combined effect of the hierarchical retainer and the attention supervision provided by the SPP task, enabling it to capture both shallow and deep semantic cues more effectively.

5.5 Error analysis

To better understand the sources of misclassification in person-centric RE, we perform a detailed error analysis based on SELF's learned representation space and prediction outcomes.

We first apply t-SNE to visualize the high-dimensional semantic representations learned by SELF, projecting them into a two-dimensional space for inspection (Fig. 8). Most relation categories, such as *per:employee_of*, *org:top_members*, and *org:founded_by*, form compact and well-separated clusters, indicating that SELF effectively captures discriminative features for these relation types. However, certain categories show ambiguous boundaries and substantial overlaps, for instance, *per:other_family* and *per:spouse*, as well as *per:origin* and *per:countries_of_residence*, exhibit notable entanglement (highlighted with dashed circles), mainly due to their high semantic similarity, which challenges fine-grained relation discrimination.

We examine the confusion matrix of the 20 most frequently misclassified relations, focusing on person-centric types in the TACRED test set (Fig. 9). Many confusion pairs involve subtle distinctions between closely related categories such as *per:employee_of*, *per:title*, and *org:top_members/employees*. These relations



(a) (b) (c) (d)

Fig. 7 F1 comparison of models for person-related relations in Re-TACRED

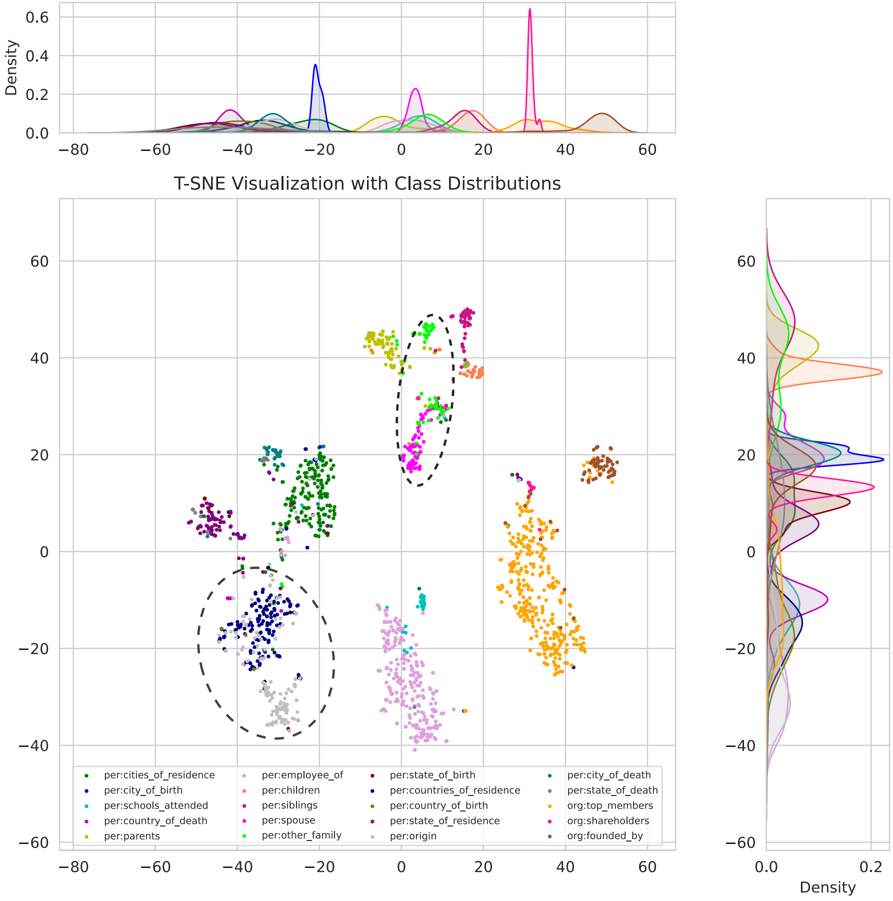
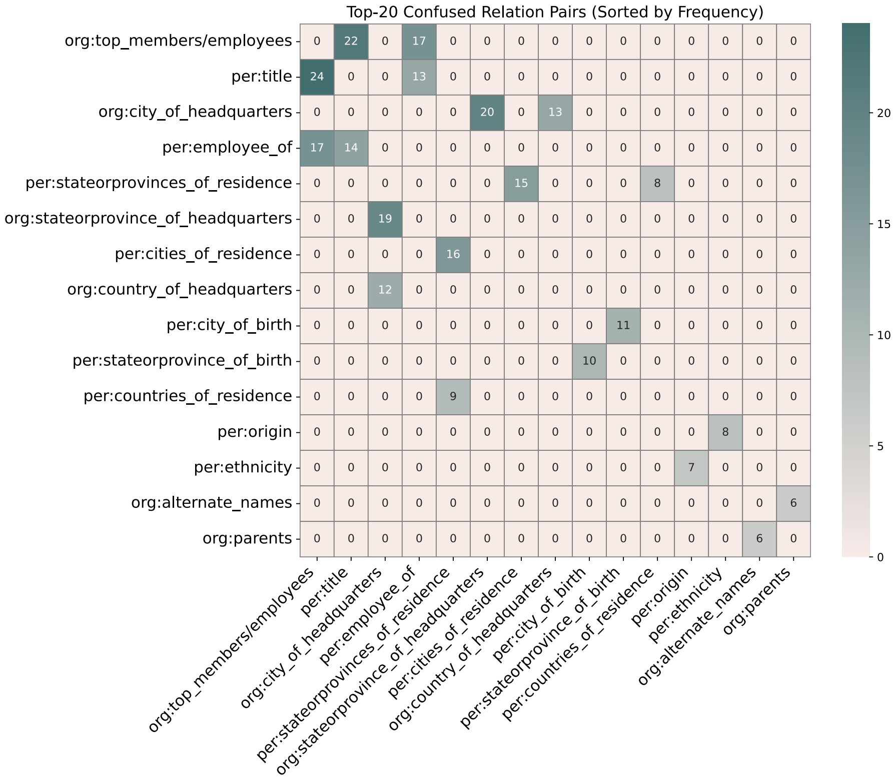


Fig. 8 T-SNE visualization of the SELF model on person-centric relation categories in the TACRED dataset

often co-occur in similar contexts, making them difficult to separate. For example, “Barack Obama, the president of the USA” should be labeled as *per:title*, yet it may be misclassified as *per:employee_of* or *org:top_members/employees* due to overlapping semantic cues. Geographic person-centric relations, such as *per:stateorprovince_of_residence* and *per:country_of_residence*, are also frequently confused, largely because of subtle granularity differences and shared lexical indicators. These cases highlight the importance of token-level distinctions and accurate modeling of entity roles for reliable prediction.

This error analysis reveals two major challenges in person-centric RE: (1) fine-grained semantic ambiguity among closely related relations and (2) contextual overlap in semantically similar expressions. While SELF demonstrates strong capability in learning discriminative features, their performance still is relative



	org:top_members/employees	per:title	org:city_of_headquarters	per:employee_of	per:stateorprovinces_of_residence	org:stateorprovince_of_headquarters	per:cities_of_residence	org:country_of_headquarters	per:city_of_birth	per:stateorprovince_of_birth	per:countries_of_residence	per:origin	per:ethnicity	org:alternate_names	org:parents
org:top_members/employees	0	22	0	17	0	0	0	0	0	0	0	0	0	0	0
per:title	24	0	0	13	0	0	0	0	0	0	0	0	0	0	0
org:city_of_headquarters	0	0	0	0	0	20	0	13	0	0	0	0	0	0	0
per:employee_of	17	14	0	0	0	0	0	0	0	0	0	0	0	0	0
per:stateorprovinces_of_residence	0	0	0	0	0	0	15	0	0	0	8	0	0	0	0
org:stateorprovince_of_headquarters	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0
per:cities_of_residence	0	0	0	0	16	0	0	0	0	0	0	0	0	0	0
org:country_of_headquarters	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0
per:city_of_birth	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0
per:stateorprovince_of_birth	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0
per:countries_of_residence	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0
per:origin	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0
per:ethnicity	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0
org:alternate_names	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
org:parents	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0

Fig. 9 Top-20 relation confusion submatrix of SELF on the TACRED test set

lower in these scenarios, underscoring the need for more context-sensitive modeling and explicit semantic disentanglement in future work.

5.6 Category sensitivity analysis

As highlighted in the preceding error analysis, certain person-centric relations exhibit significant semantic overlap and are prone to misclassification. These include *per:spouse* versus *per:other_family*, and *per:origin* versus *per:countries_of_residence*. Although SELF demonstrates strong capability in capturing discriminative relational features, these fine-grained categories remain challenging due to subtle contextual differences and overlapping semantic cues. To further examine SELF's effectiveness in such scenarios, we conduct two complementary analyses. First, we compare SELF against strong baselines on representative confusing relation categories to evaluate its fine-grained discriminative power. Second, we assess whether SELF's advantage is sensitive to the availability of person-centric examples in the training data.

Performance on confusing relations We begin by evaluating SELF on a set of person-centric relation types that are frequently confused due to semantic proximity (discussed in the Sect. 5.5). Table 7 reports class-wise F1 on challenging relation pairs such as *per:spouse* versus *per:other_family* and *per:origin* versus *per:countries_of_residence* with previous best models. Despite the inherent ambiguity of these categories, SELF consistently achieves the highest performance across all listed relations, except *per:title*. For instance, it improves F1 by + 7.00 on *per:stateorprovinces_of_residence* and +3.05 on *per:spouse* over the best baseline. This discovery indicates that our model has made significant progress in these easily confused relationships, further demonstrating the contribution of our model.

Sensitivity to data availability To verify whether SELF's improvements rely on large amounts of person-centric data, we progressively downsample person-centric training samples to 75%, 50%, and 25% while keeping the test distribution fixed. Figure 10 shows that SELF consistently outperforms baselines across all reduced settings, even when only 25% of the original person-centric data is available. This demonstrates that SELF's advantage stems from its enhanced semantic modeling ability rather than dependence on data quantity, confirming its robustness under data scarcity.

5.7 Hyperparameter sensitivity analysis on λ

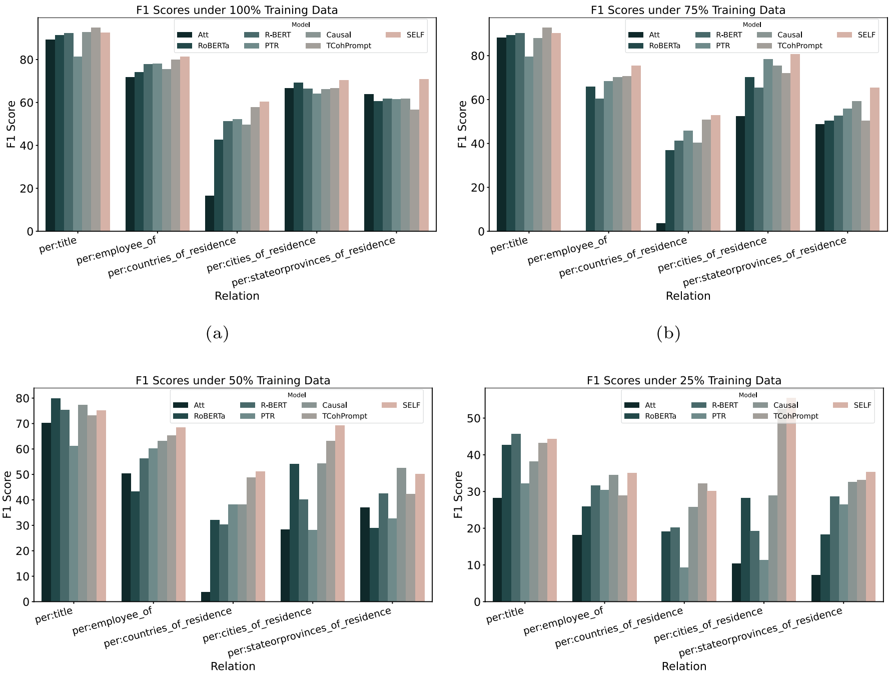
In MTL, a central challenge is how to assign appropriate weights to loss function of each task. As noted in prior work [56], one common and practical strategy is to treat these weights as hyperparameters and tune them via grid search. Following this established practice, we adopt a grid search method to set the hyperparameter λ, which balances the losses of the main RE task and SPP task. **Our analysis shows that SELF is not overly sensitive to λ and maintains strong performance across a wide range of λ values.**

We systematically analyze the influence of λ on model performance from three perspectives: (1) performance under different λ values, (2) robustness across random seeds, and (3) the effectiveness of an automatic task weighting method. Our goal is to demonstrate that the proposed SELF model, particularly when equipped with the

Table 7 SELF improves performance on confusing person-centric relation categories

Confusing relations	Best-model	Best-F1	SELF-F1	△
per:stateorprovinces_of_residence	Att	63.76	70.76	+7.00
per:countries_of_residence	TCohPrompt	57.88	60.44	+2.56
per:other_family	Cause	60.86	64.34	+3.48
per:spouse	Cause	82.16	85.21	+3.05
per:origin	RoBERTa	65.48	67.34	+1.86
per:cities_of_residence	RoBERTa	69.12	70.46	+1.34
per:employee_of	TCohPrompt	79.91	81.38	+1.47
org:top_members/employees	R-BERT	86.66	87.55	+0.89
per:title	TCohPrompt	94.80	92.40	−2.40

△ refers to the difference between SELF-F1 and Best-F1. Best-model and Best-F1 are the previous best baselines and best performance of a specific relation, respectively



(a)

(b)

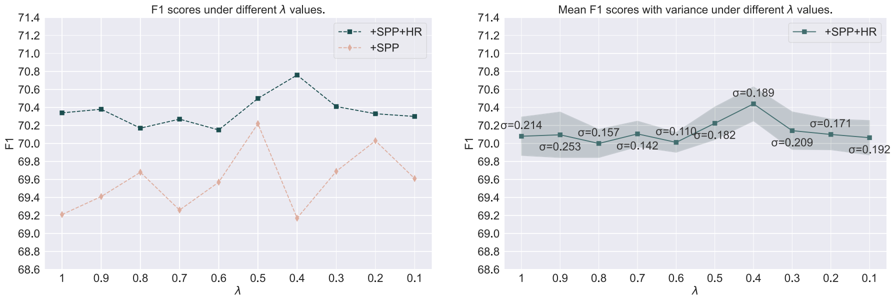
(c)

(d)

Fig. 10 Model performance under different proportions of person-centric samples in training, evaluated on the full test set

SPP and HR modules, is not overly sensitive to λ and can maintain strong and stable performance under a broad range of settings.

Performance across λ values Figure 11a illustrates the F1 of different model variants on the TACRED dataset under varying λ values. The orange curve represents the F1 obtained at different λ values when only utilizing the SPP task without the HR



(a)

(b)

Fig. 11 F1 (left) and Mean F1 values with standard deviation σ (right) of SELF variants on TACRED across different λ values

module. We can observe that the F1 fluctuates across different λ values, indicating that the parameter λ impacts model performance in MTL. The dark green line represents the F1 obtained after further incorporating the HR module and the SPP task. Although the model's performance still exhibits fluctuations across different λ values, introducing the HR module enhances performance compared to using the SPP task alone, leading to more stable variations. This variation indirectly confirms the effectiveness of the HR method incorporating the SPP task in supporting RE, and it can also reduce the reliance on the λ parameter to a certain extent.

Robustness under random seeds In the main experiment, to ensure the reproducibility of the experimental results, we used fixed seeds. To further examine robustness, we repeat experiments across 10 values of λ (from 0.1 to 1.0) under 5 random seeds each. As shown in Fig. 11b, the average F1 remains tightly bounded between 70.00 and 70.44, and standard deviations are consistently small (average std ≈ 0.18), indicating that performance is stable and not highly sensitive to the exact choice of λ. Especially, the model performs robustly within the range $\lambda \in [0.4, 0.8]$, allowing flexible deployment.

Effectiveness of automatic task weighting To test whether manual tuning of λ can be replaced by adaptive strategies, we also implement the uncertainty-based multi-task loss [57], which introduces trainable observation noise variables σ_{re} and σ_{spp}:

$$\mathcal{L} = \frac{1}{2\sigma_{\mathrm{RE}}^2}\mathcal{L}_{\mathrm{re}} + \frac{1}{2\sigma_{\mathrm{spp}}^2}\mathcal{L}_{\mathrm{spp}} + \log \sigma_{\mathrm{re}} + \log \sigma_{\mathrm{spp}} \tag{15}$$

This allows the model to automatically adjust task weights during training based on uncertainty. As shown in Table 8, while the performance remains stable across seeds (Mean F1 70.0), it is slightly inferior to our best manually tuned (random seeds) results ($\lambda = 0.4$ yields 70.44). This suggests that manual tuning still offers a performance advantage for this task.

Although model performance may vary slightly with different λ settings, our results show that SELF consistently achieves strong and stable performance across a wide λ range. More importantly, for our core goal, enhancing the extraction of person-centric relations, SELF remains robust and effective. This confirms that the model's key innovations, including the auxiliary SPP task and the HR module, contribute substantially to its practical utility and novelty, even under hyperparameter variations.

5.8 Model efficiency

To demonstrate SELF's suitability for large-scale, high-throughput RE in high-performance computing environments, we comprehensively evaluate its computational efficiency. Specifically, we compare SELF against several baselines on the TACRED

Table 8 Performance of uncertainty-based multi-task loss across 5 random seeds

Seeds	42	52	62	72	82
F1	70.05	69.94	69.88	70.01	70.13

Table 9 Comparison of model efficiency on TACRED

Dataset	Models	Params (K)	FLOPs (G)	TT (s)	IT (ms)	GPU (GB) T/I
TACRED	Att [52]	154,904	20.14	112.63	0.13	2.16/1.33
	RoBERTa [23]	432,483	41.40	802.77	4.95	4.50/2.46
	GPT2 [54]	3,190,666	353.74	3556.31	32.55	15.31/5.57
	PTR [24]	432,375	261.66	3661.49	23.98	9.45/3.56
	Causal [25]	642,515	44.08	737.79	4.61	5.51/2.45
	R-BERT [22]	432,471	261.63	3596.29	23.41	8.92/3.21
	TCohPrompt [53]	1,344,429	232.46	4320.33	15.52	9.46/2.28
	SELF	732,948	37.22	789.77	4.73	4.72/2.53

TT and IT denote training time (s) and inference time (ms); T/I indicates GPU memory usage (GB) during training and inference

dataset across multiple dimensions: time complexity, number of parameters, floating-point operations (FLOPs), training and inference time, and GPU memory usage.

Following prior work [58, 59], all experiments were conducted under a unified hardware setting using RTX 3090 GPUs.[1] We use a batch size of 6 for training and 1 for inference. Training time (in seconds) denotes the time per epoch, while inference time (in milliseconds) indicates processing time per instance. GPU memory usage is reported separately for training and inference phases. Table 9 provides detailed results.

Moreover, the design of SELF, featuring parameter-free auxiliary tasks (SPP) and parallelizable hierarchical fusion modules (HR), facilitates efficient distributed execution. This enables real-time processing of large-scale person-centric relation data, which is essential for intelligent financial systems, large-scale social media monitoring, and biomedical literature mining.

As shown in Table 9, SELF outperforms many recent methods in computational efficiency while maintaining comparable F1 performance. Compared to multi-stage or prompt-based models such as R-BERT, PTR, or TCohPrompt, SELF achieves over 4× faster training and inference speeds and requires significantly less memory and FLOPs. These characteristics make SELF highly suitable for deployment in high-performance computing environments or latency-sensitive settings where model efficiency is critical.

5.9 Limitations of SELF

Despite the promising performance of the proposed SELF framework, we acknowledge the following limitations:

- **Task-specific hyperparameter sensitivity** While our experiments demonstrate that SELF performs robustly across a wide range of λ values, the performance of auxiliary tasks may still fluctuate slightly depending on dataset characteris-

[1] https://github.com/Lyken17/pytorch-OpCounter.

tics. Although grid search is a widely accepted tuning strategy, integrating more adaptive weighting mechanisms could further improve generalizability.

- **Single-instance forward dependency** SELF processes each instance independently without leveraging document-level or dialog-level context. This may limit its capability in handling relations that span across multiple sentences or depend on global discourse structures.
- **Uneven gains across relation subtypes** Although SELF significantly improves performance on most person-centric relations, we observe that a few fine-grained subcategories, such as *per:other_family* and *per:countries_of_residence*, still exhibit relatively modest improvements compared to simpler categories like *per:title*. This suggests that further refinement in modeling subtle semantic distinctions, possibly through external knowledge integration or contrastive learning, remains a worthwhile direction.

6 Conclusions

This paper presented SELF, a multi-task learning framework that (1) injects explicit supervision for person-related semantics through the parameter-free SPP task and (2) preserves and refines these signals across encoder layers via the parallel HR module. Working in synergy, these components mitigate the loss of fine-grained semantic cues in deep encoders, thereby enhancing category sensitivity to person-centric relations and stabilizing multi-task optimization, even under substantial reductions in training data. Extensive experiments on TACRED and Re-TACRED show that SELF consistently surpasses strong baselines in person-centric relation classification while maintaining competitive performance across all relation types. These findings confirm that its advantage derives from heightened semantic sensitivity rather than dependence on data quantity. Future work will explore extending SELF to a wider range of relation types and domains and incorporating external knowledge to further improve adaptability and generalization.

Appendix

See Table 10.

Table 10 TACRED statistics and classification of relation types

Relations	Train	Dev	Test	Rel-type
per:parents	152	56	88	Person-Centric
per:employee_of	1524	375	264	
per:children	211	99	37	
per:siblings	165	30	55	
per:spouse	258	159	66	
per:other_family	179	80	66	
per:state_of_birth	38	26	8	
per:countries_of_residence	445	226	148	
per:country_of_birth	28	20	5	
per:state_of_residence	331	72	81	
per:cities_of_residence	374	179	189	
per:city_of_birth	65	33	5	
per:schools_attended	149	50	30	
per:country_of_death	6	46	9	
per:origin	325	210	132	
per:city_of_death	81	118	28	
per:state_of_death	49	41	14	
per:date_of_birth	63	31	9	
per:cause_of_death	117	168	52	
per:age	390	243	200	
per:date_of_death	63	206	54	
per:religion	53	53	47	
per:charges	72	105	103	
per:title	2443	919	500	
per:alternate_names	104	38	11	
org:top_members/employees	1890	534	346	
org:shareholders	76	55	13	
org:founded_by	124	76	68	

Table 10 (continued)

Relations	Train	Dev	Test	Rel-type
org:website	111	86	26	Non-Person-Centric
org:political	105	10	10	
org:alternate_names	808	338	213	
org:number_of_employees	75	27	19	
org:subsidiaries	296	113	44	
org:parents	286	96	62	
org:members	170	85	31	
org:member_of	122	31	18	
org:founded	124	38	37	
org:city_of_headquarters	382	109	82	
org:dissolved	23	8	2	
org:country_of_headquarters	468	177	108	
org:state_of_headquarters	229	70	51	

Most relation types are person-related; we call them person-centric relationships

Acknowledgements This research work was supported by the Natural Science Foundation of Xinjiang Uyghur Autonomous Region, China (Nos. 2023D01B12, 2024D01B05, 2024D01B07), the National Natural Science Foundation of China (No. 62376228), the Sichuan Province Science and Technology Support Program, China (Nos. 2023NSFSC1412, 2025ZNSFSC1505), and the Chengdu Science and Technology Program, Sichuan, China (No. 2023-JB00-00016-GX).

Author contributions Hailin Wang helped in conceptualization, methodology, software, investigation. Wentong Niu was involved in writing—original draft preparation, validation, formal analysis. Hangyi Ren helped in visualization, writing—review & editing. Jiahao Li was involved in validation and visualization. Jinxuan Tian and Dan Zhang helped in validation.

Data availability https://github.com/WentongNiu/SELF.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Xiaoyan Z, Yang D, Min Y, Lingzhi W, Rui Z, Hong C, Wai L, Ying S, Ruifeng X (2023) A comprehensive survey on deep learning for relation extraction: recent advances and new frontiers. arXiv preprint arXiv:2306.02051
2. Zheng C, Wu Z, Feng J, Fu Z, Cai Y (2021) MNRE: a challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, pp 1–6
3. Qiu L, Liang Y, Zhao Y, Lu P, Peng B, Yu Z, Wu YN, Zhu S-C (2021) SOCAOG: incremental graph parsing for social relation inference in dialogues. arXiv preprint arXiv:2106.01006

4. Khaldi H, Benamara F, Siegel G, Pradel C, Aussenac-Gilles N (2022) How's business going worldwide? a multilingual annotated corpus for business relation extraction. In: 13th Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association (ELRA), pp 3696–3705
5. Liu W, Zhang X, Wang Z, Zheng W (2024) Research on causality extraction algorithm for medical text based on bert and graph attention network. In: 2024 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI). IEEE, pp 187–193
6. Yang Z (2020) Biomedical information retrieval incorporating knowledge graph for explainable precision medicine. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, p 2486
7. Zhang H, Zhao Y, Sun B, Wu Y, Fu Z, Xiao X (2025) Large language model based intelligent fault information retrieval system for new energy vehicles. Preprints https://doi.org/10.20944/preprints202503.1268.v1
8. Santoro A, Raposo D, Barrett DG, Malinowski M, Pascanu R, Battaglia P, Lillicrap T (2017) A simple neural network module for relational reasoning. Adv Neural Inf Process Syst 30
9. Liang K, Meng L, Zhou S, Tu W, Wang S, Liu Y, Liu M, Zhao L, Dong X, Liu X (2024) Mines: message intercommunication for inductive relation reasoning over neighbor-enhanced subgraphs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 38, pp 10645–10653
10. Mouiche I, Saad S (2025) Entity and relation extractions for threat intelligence knowledge graphs. Comput Secur 148:104120
11. Wang H, Qin K, Zakari RY, Lu G, Yin J (2022) Deep neural network-based relation extraction: an overview. Neural Comput Appl 1–21
12. Plum A, Ranasinghe T, Jones S, Orasan C, Mitkov R (2022) Biographical semi-supervised relation extraction dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 3121–3130
13. Min C, Mathew R, Pan J, Bansal S, Keshavarzi A, Kannan AV (2025) Efficient knowledge graph construction and retrieval from unstructured text for large-scale rag systems. arXiv preprint arXiv:2507.03226
14. Zaratiana U, Tomeh N, Holat P, Charnois T (2024) An autoregressive text-to-graph framework for joint entity and relation extraction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 38, pp 19477–19487
15. Ding Z, Huang W, Liang J, Xiao Y, Yang D (2024) Improving recall of large language models: a model collaboration approach for relational triple extraction. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp 8890–8901
16. Wang Y, Cao P, Fang H, Ye Y (2025) Span-aware pre-trained network with deep information bottleneck for scientific entity relation extraction. Neural Netw 107250
17. Yang F, Ren M, Kong D, Liu S, Fu Z (2025) Rehearsal-free continual few-shot relation extraction via contrastive weighted prompts. Neurocomputing 129741
18. Devlin J (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
19. Chen T, Zhou L, Wang N, Chen X (2022) Joint entity and relation extraction with position-aware attention and relation embedding. Appl Soft Comput 119:108604
20. Liu Z, Li H, Wang H, Liao Y, Liu X, Wu G (2023) A novel pipelined end-to-end relation extraction framework with entity mentions and contextual semantic representation. Expert Syst Appl 228:120435
21. Gao C, Zhang X, Li L, Li J, Zhu R, Du K, Ma Q (2023) ERGM: a multi-stage joint entity and relation extraction with global entity match. Knowl Based Syst 271:110550
22. Wu S, He Y (2019) Enriching pre-trained language model with entity information for relation classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp 2361–2364
23. Zhou W, Chen M (2021) An improved baseline for sentence-level relation extraction. arXiv preprint arXiv:2102.01373
24. Han X, Zhao W, Ding N, Liu Z, Sun M (2022) PTR: prompt tuning with rules for text classification. AI Open 3:182–192
25. Wang F, Mo W, Wang Y, Zhou W, Chen M (2023) A causal view of entity bias in (large) language models. arXiv preprint arXiv:2305.14695

26. Le T-T, Nguyen M, Nguyen TT, Van LN, Nguyen TH (2024) Continual relation extraction via sequential multi-task learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 38, pp 18444–18452
27. Yang J, Zhao Y, Yang L, Wang X, Chen L, Wang F-Y (2024) Temprompt: Multi-task prompt learning for temporal relation extraction in rag-based crowdsourcing systems. arXiv preprint arXiv:2406.14825
28. Tu C, Zuo J, Hu Y, Wan J, Wang M (2025) Multi-step fusion of relation type information and multi-task decoding for entity relation extraction in ancient Chinese. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 1–5
29. Swarup A, Pan T, Wilson R, Bhandarkar A, Woodard D (2025) LLM4RE: a data-centric feasibility study for relation extraction. In: Proceedings of the 31st International Conference on Computational Linguistics, pp 6670–6691
30. Deußer T, Ali SM, Hillebrand L, Nurchalifah D, Jacob B, Bauckhage C, Sifa R (2022) KPI-EDGAR: a novel dataset and accompanying metric for relation extraction from financial documents. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp 1654–1659
31. Wu H, Lei Q, Zhang X, Luo Z (2020) Creating a large-scale financial news corpus for relation extraction. In: 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, pp 259–263
32. Jabbari A, Sauvage O, Zeine H, Chergui H (2020) A French corpus and annotation schema for named entity recognition and relation extraction of financial news. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp 2293–2299
33. Sharma S, Nayak T, Bose A, Meena AK, Dasgupta K, Ganguly N, Goyal P (2022) FINRED: a dataset for relation extraction in financial domain. In: Companion Proceedings of the Web Conference 2022, pp 595–597
34. Kaur S, Smiley C, Gupta A, Sain J, Wang D, Siddagangappa S, Aguda T, Shah S (2023) REFIND: relation extraction financial dataset. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 3054–3063
35. Andrew JJ (2018) Automatic extraction of entities and relation from legal documents. In: Proceedings of the Seventh Named Entities Workshop, pp 1–8
36. Hendrycks D, Burns C, Chen A, Ball S (2021) CUAD: an expert-annotated NLP dataset for legal contract review. arXiv preprint arXiv:2103.06268
37. Chen Y, Sun Y, Yang Z, Lin H (2020) Joint entity and relation extraction for legal documents with legal feature enhancement. In: Proceedings of the 28th International Conference on Computational Linguistics, pp 1561–1571
38. Xu W, Deng Y, Lei W, Zhao W, Chua T-S, Lam W (2022) ConReader: exploring implicit relations in contracts for contract clause extraction. arXiv preprint arXiv:2210.08697
39. Wang Z, Song H, Ren Z, Ren P, Chen Z, Liu X, Li H, Rijke M (2021) Cross-domain contract element extraction with a bi-directional feedback clause-element relation network. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1003–1012
40. Thomas A, Sangeetha S (2021) Semi-supervised, knowledge-integrated pattern learning approach for fact extraction from judicial text. Expert Syst 38(3):12656
41. Wang Z, Wang H, Ren P, Chen Z, Rijke M, Ren Z (2025) Graph-enhanced prompt learning for cross-domain contract element extraction. ACM Trans Inf Syst 43(3):1–24
42. Zhao X, Deng Y, Yang M, Wang L, Zhang R, Cheng H, Lam W, Shen Y, Xu R (2024) A comprehensive survey on relation extraction: recent advances and new frontiers. ACM Comput Surv 56(11):1–39
43. Li D, Wu P, Dong Y, Gu J, Qian L, Zhou G (2023) Joint learning-based causal relation extraction from biomedical literature. J Biomed Inform 139:104318
44. Guo Y, Liu Z, Huang C, Liu J, Jing W, Wang Z, Wang Y (2021) CyberRel: joint entity and relation extraction for cybersecurity concepts. In: Information and Communications Security: 23rd International Conference, ICICS 2021, Chongqing, China, November 19–21, 2021, Proceedings, Part I 23. Springer, pp 447–463
45. Zhao W, He P, Zeng Z, Xu X (2024) Fake news detection based on knowledge-guided semantic analysis. Electronics 13(2):259

46. Wang H, Zhang D, Liu G, Huang L, Qin K (2024) Enhancing relation extraction using multi-task learning with SDP evidence. Inf Sci 670:120610
47. Niu Z, Zhong G, Yu H (2021) A review on the attention mechanism of deep learning. Neurocomputing 452:48–62
48. Zalmout N, Habash N (2019) Adversarial multitask learning for joint multi-feature and multi-dialect morphological modeling. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp 1775–1786
49. Zhang Y, Zhong V, Chen D, Angeli G, Manning CD (2017) Position-aware attention and supervised data improve slot filling. In: Conference on Empirical Methods in Natural Language Processing
50. Stoica G, Platanios EA, Póczos B (2021) Re-TACRED: addressing shortcomings of the TACRED dataset. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 35, pp 13843–13850
51. Hendrickx I, Kim SN, Kozareva Z, Nakov P, Séaghdha DO, Padó S, Pennacchiotti M, Romano L, Szpakowicz S (2010) Semeval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. ACL 2010:33
52. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 2: Short Papers), pp 207–212
53. Long J, Yin Z, Liu C, Huang W (2024) TCohPrompt: task-coherent prompt-oriented fine-tuning for relation extraction. Complex Intell Syst 10(6):7565–7575
54. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
55. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692
56. Chen S, Zhang Y, Yang Q (2024) Multi-task learning in natural language processing: an overview. ACM Comput Surv 56(12):1–32
57. Kendall A, Gal Y, Cipolla R (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7482–7491
58. Gao C, Zhang X, Li L, Li J, Zhu R, Du K, Ma Q (2023) ERGM: a multi-stage joint entity and relation extraction with global entity match. Knowl Based Syst 271:110550
59. Zhang L, Zheng N (2025) DIREL: joint relational triple extraction through dual implicit relation. Neurocomputing 129374

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Hailin Wang[1,2,3] · Wentong Niu[1] · Hangyi Ren[1] · Jiahao Li[1] · Jingxuan Tian[1] · Dan Zhang[1,2,3]

✉ Dan Zhang
danzhang@swufe.edu.cn

Hailin Wang
wanghl@swufe.edu.cn

Wentong Niu
223081200003@smail.swufe.edu.cn

Hangyi Ren
224081200039@smail.swufe.edu.cn

Jiahao Li
42311137@smail.swufe.edu.cn

Jingxuan Tian
42311103@smail.swufe.edu.cn

[1] School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, Sichuan, China

[2] Engineering Research Center of Intelligent Finance, Ministry of Education, Chengdu, China

[3] Kash Institute of Electronics and Information Industry, Kashi, Xinjiang, China