

ESLA: Empirical and Semantic Label Alignment for Multi-Source NER Transfer in Unlabeled Target Domains

Xiaobo Zhang

Universiti Sains Malaysia

Congqing He

Huzhou Normal University

Ying He

Universiti Sains Malaysia

Jian Peng

Jiangxi Vocational College of Finance & Economics

Dajie Fu

Jiangxi Vocational College of Finance & Economics

Tien-Ping Tan

`tienping@usm.my`

Universiti Sains Malaysia

Article

Keywords: Named Entity Recognition, Label Alignment, Schema Matching, Heterogeneous Datasets

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10594816/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

ESLA: Empirical and Semantic Label Alignment for Multi-Source NER Transfer in Unlabeled Target Domains

Xiaobo Zhang^{1,2}, Congqing He³, Ying He^{1,2}, Jian Peng²,
Dajie Fu², Tien-Ping Tan^{1*}

¹School of Computer Sciences, Universiti Sains Malaysia, 11700,
Penang, Malaysia.

²School of Information Engineering, Jiangxi Vocational College of
Finance & Economics, Jiujiang, 332000, China.

³School of Information Engineering, Huzhou Normal University, 313000,
Huzhou, China.

*Corresponding author(s). E-mail(s): tienping@usm.my;
Contributing authors: zhangxiaobo@student.usm.my;
hecongqing@zjhu.edu.cn; heyings11@student.usm.my;
2007010@jxvc.jx.cn; 2004034@jxvc.jx.cn;

Abstract

Heterogeneous label schemas limit the reuse of public named entity recognition (NER) datasets for cross-domain transfer, particularly when labeled target-domain training data are unavailable. We present ESLA (Empirical and Semantic Label Alignment), a framework that aligns labels by combining empirical cross-dataset prediction consistency with contextual semantic representations. ESLA uses stability-constrained optimization to preserve source-side performance and context-aware label completion to mitigate missing-label conflicts before merging datasets into a unified training corpus. We evaluate ESLA by merging three Chinese NER benchmarks (OntoNotes 5.0, CLUENER, and BosonNER) and transferring to the financial-domain FinReportNER benchmark. FinReportNER annotations are used only for evaluation, not for alignment, training, threshold tuning, or model selection. Experimental results show that ESLA-Merge achieves the highest Macro-F1 of 0.59, while matching the best Micro-F1 of 0.56 and improving transferable label coverage over single-source training. Across all configurations, it also gets notable gains on categories such as **PRODUCT** and **TIME**.

These results show that empirical and semantic label alignment can integrate heterogeneous public datasets for cross-domain NER without target-domain training supervision.

Keywords: Named Entity Recognition, Label Alignment, Schema Matching, Heterogeneous Datasets

1 Introduction

NER aims to extract and classify named entities in text and plays an important role in various Natural Language Processing (NLP) downstream tasks, including information extraction, knowledge graph construction, and question answering [1, 2]. Recent advances in deep learning, particularly pre-trained language models, have significantly improved NER performance. However, these data-driven approaches still rely heavily on high-quality, large-scale annotated corpora. In specific domains such as finance and healthcare, constructing such datasets is expensive, time-consuming, and expertise-intensive [3–6]. As a result, researchers often face a reality where large volumes of domain-specific texts are available but unlabeled.

A natural way to alleviate this problem is to reuse existing public NER datasets. However, relying on a single-source dataset usually provides limited transferable supervision because most public datasets are constructed for general-domain scenarios such as news or social media. While categories with relatively stable semantics (e.g., company or date) can often be transferred successfully, domain-specific categories may remain poorly covered. Consequently, single-source transfer often suffers from limited target label coverage.

To overcome this limitation, recent studies have explored learning from multiple NER corpora or standardizing heterogeneous public NER datasets to leverage complementary supervision across sources [7–9]. However, existing datasets are constructed under heterogeneous label schemas, making them difficult to use as a unified training resource. For example, the definition of label **TIME** in OntoNotes 5.0 [10] is different from **time** in BosonNER [11], and temporal expressions in financial reports are closer to **DATE** than to clock-time expressions. Such label inconsistencies complicate joint training and introduce conflicting supervision signals, ultimately degrading NER performance.

Previous studies have attempted to address label inconsistency by constructing domain knowledge-based label graphs [12] or adopting pseudo-labeling [13] strategies. These methods often lack interpretability and rely on prior expert knowledge, limiting their adaptability to new domains. More recently, prompting-based unified frameworks like PUnifiedNER [14] bypass alignment through instruction tuning. However, these strategies treat heterogeneous labels in isolation, failing to resolve underlying semantic conflicts.

The problem of label alignment is closely related to research on ontology alignment. Related efforts in ontology alignment have shown success in database schema matching, but traditional term-based methods often struggle in NER contexts. These

methods mainly focus on matching concept definitions across structured knowledge bases through lexical similarity, hierarchical relations, or external ontologies [15–17]. While effective in domains with well-defined ontologies, these methods depend on explicit concept definitions, which are often unavailable in NER datasets.

To address these challenges, we introduce ESLA, an automatic label alignment framework that improves the accessibility and reusability of existing data resources. Instead of relying on manual heuristic rules, we formulate the label alignment task as a stability-constrained optimization problem. ESLA jointly models two complementary signals: the empirical signal, which measures the proportion of entity overlap across different datasets, revealing commonalities in annotation standards and label granularity; and the semantic signal, which leverages contextual embeddings to measure the spatial distance between the contextual label prototypes. By dynamically integrating these two signals, our framework can search for the optimal label alignment schema that maximizes the dataset merging capacity while strictly safeguarding model performance.

We perform our study in Chinese NER, where label ambiguity and schema heterogeneity are more prevalent across Chinese datasets. Our experiments first merge three mainstream Chinese NER datasets (OntoNotes 5.0, CLUENER2020, and BosonNER) into a unified corpus and evaluate the unified corpus on a self-annotated financial target-domain benchmark FinReportNER. FinReportNER annotations are only used for evaluation and are not used for source-side label alignment, model training, label completion, threshold tuning, or model selection. The main contributions of this paper are summarized as follows:

1. We propose ESLA, an interpretable label alignment framework for cross-domain NER. ESLA formulates heterogeneous label alignment as a dual-signal optimization problem by jointly exploiting empirical prediction consistency and contextual semantic similarity, without relying on external knowledge bases or heavy manual intervention.
2. We develop a scalable multi-source dataset-merging strategy to unify heterogeneous label schemas, effectively alleviating the limitations of insufficient label coverage in single-source datasets.
3. We construct a unified Chinese NER corpus by aligning and merging three heterogeneous NER datasets. Based on this corpus, we compare ESLA with several alignment baselines, including rule-based, expert-defined, and instruction-based LLM alignment methods.
4. We validate ESLA in a target-unlabeled cross-domain transfer setting on a financial domain NER benchmark. Experimental results show that ESLA-Merge improves label coverage and achieves the best Macro-F1 across all source training settings, demonstrating its effectiveness for cross-domain generalization.

2 Related Works

2.1 Evolution of NER Paradigms and Cross-Domain Transfer

Early NER approaches progressed from rule-based systems to statistical models like conditional random fields (CRFs), which offered interpretability but relied on manual feature engineering. The advent of deep learning shifted the paradigm toward automated representation learning, with architectures evolving from convolutional neural network (CNN) and recurrent neural network (RNN) based models to pre-trained language models (PLMs) [3, 4]. Meanwhile, alternative paradigms have arisen. Encoder-Decoder models like T5 and BART treat NER as a sequence generation task, with the encoder extracting features and the decoder generating entity labels, and are typically trained or fine-tuned on task-specific data [14, 18]. By contrast, recent instruction-based large language models (LLMs), such as GPT and Gemini, perform NER through prompt-based inference without task-specific fine-tuning, enabling zero-shot or low-resource adaptation [19–21]. Despite this architectural diversity, most existing methods remain model-driven and rely heavily on consistent, high-quality annotations, which motivates data-driven solutions such as label alignment and dataset integration [22]. Therefore, cross-domain NER requires not only stronger models but also effective mechanisms for reusing heterogeneous annotated resources. Recent studies have explored learning from multiple NER corpora or standardizing heterogeneous public NER datasets to leverage complementary supervision across sources [7–9]. In contrast to model-centric transfer methods, this work focuses on label alignment and multi-source dataset integration for target-unlabeled cross-domain NER.

2.2 NER Datasets and Multi-Source Data Reuse

NER research has been supported by a variety of benchmark datasets, ranging from general-domain and domain-specific scenarios. General-domain benchmarks like CoNLL-2003 [23] and MSRA-NER [24] mainly cover formal news in English and Chinese, respectively. OntoNotes 5.0 expanded this scope with a trilingual corpus (English, Chinese, and Arabic) from various media channels. In contrast, CLUENER2020 and BosonNER cover more diverse and informal texts, including online news, user comments, and social media texts, which usually contain colloquial expressions and more noise. In vertical fields, Zhang et al. [25], Loukas et al. [26], and Shah et al. [27] constructed datasets that focused on extracting financial entities from annual reports and eXtensible Business Reporting Language (XBRL) annotations.

Although these datasets provide valuable annotated resources, reusing these multi-source heterogeneous data faces severe alignment challenges. First, entity distributions vary substantially across domains. For instance, BosonNER originates from social media, and its company names often appear as Weibo handles (e.g., Apple), which differ from the formal, standardized company names in financial reports. Secondly, semantic discrepancies are widespread. That is, the same term may correspond to different label definitions across datasets. For example, the word **apple** in CLUENER may be labeled as **product** or **movie**, whereas in financial datasets it is usually labeled as **organization**. Conversely, the same label name may have different semantic scopes

Table 1: Overview of the source NER datasets used for multi-source label alignment.

Dataset	Language / Domain	Size (Train)	#Labels	Label characteristics
OntoNotes 5.0 (ZH)	Chinese / Mixed media	~50K	18	Ontology-driven schema with general-purpose entity types, such as <code>GPE</code> , <code>FAC</code> , and <code>ORG</code>
CLUENER 2020	Chinese / Online text	~13K	10	Domain-specific flat schema with fine-grained entity types, such as <code>book</code> , <code>movie</code> , and <code>game</code>
BosonNER	Chinese / Social media and news	~2K	6	Flat schema with informal and application-oriented labels, such as <code>product_name</code> , <code>company_name</code> , and <code>government</code>

in different annotation systems. For instance, BosonNER marks `March 29th` or `1929` as `TIME`, while in OntoNotes’ definition, `TIME` only refers to a specific moment (such as `11:19`). Lastly, the datasets vary greatly in size and label granularity. From the coarse-grained classification of CoNLL-2003 to the fine-grained hierarchy of OntoNotes to the ultra-fine-grained division of Finer-139, merging these inconsistencies directly will introduce substantial noise and severely imbalance the category distribution. Table 1 summarizes the main features of the representative NER datasets included in this study. These datasets are selected to represent different domains, annotation styles, and label granularities, which will affect the label alignment of multi-source NER datasets.

2.3 Label Alignment and Schema Matching

Manual and structure-based mapping. Early studies on dataset alignment primarily relied on manual mapping, in which domain experts defined rules to align labels. Although this method is interpretable, it is labor-intensive and lacks scalability. To achieve automation, Zhao et al. [12] proposed a knowledge-driven label graph method that leverages external hierarchical structures (e.g., the pet taxonomy). However, this structure-based approach heavily relies on prior domain knowledge, which may not be available for flattened or low-resource NER patterns.

Semantic and LLM-based alignment methods. Unifying label schemas is closely related to ontology alignment (OA), which aims to match concepts across heterogeneous knowledge bases. Traditional OA methods rely on string similarity or graph structures, but recent research has integrated deep learning and generative AI. BERTMap represents an important shift by fine-tuning BERT to capture semantic context for ontology classes, outperforming string-based baselines [28, 29]. Recently, LLMs have been studied for schema alignment through prompt-based reasoning, showing a strong ability to manage complex one-to-many or many-to-many mappings in

zero-shot or few-shot scenarios. However, these approaches depend on external knowledge implicitly encoded in LLMs and may face stability problems when dealing with noisy or domain-specific NER data [15, 30].

Instance-based and Unified Frameworks. Consequently, recent research has focused on utilizing data overlaps to integrate datasets. Pseudo-labeling methods address label inconsistency by cross-dataset training [13]. In terms of frameworks, Ushio and Camacho-Collados introduced T-NER, a library for aggregating diverse NER datasets. However, instead of algorithmically resolving schema conflicts, T-NER primarily addresses heterogeneity by simplifying the task to span detection or simply concatenating datasets with disjoint label spaces [31]. To further advance automation, prompting-based unified frameworks such as PUnifiedNER have emerged, attempting to bypass explicit alignment via instruction tuning. While flexible, these implicit strategies treat heterogeneous labels in isolation, failing to resolve underlying semantic conflicts between overlapping definitions, which has been shown to induce negative transfer in multi-domain settings [14]. Other approaches utilize semantic embeddings [32], but they ignore the specific contextual semantics of entities and often struggle with partial overlaps.

Unlike existing methods, ESLA combines semantic signal with instance-level empirical statistics to handle label schema heterogeneity across multiple NER datasets. During dataset fusion, ESLA explicitly introduces stability constraints to prevent overly aggressive label merging and to preserve source-domain NER performance. Compared with traditional methods that rely solely on semantic matching, structural alignment, external knowledge, or prompt-driven approaches, ESLA provides a data-driven and interpretable solution for cross-domain transfer.

3 Methodology

As discussed in Section 2, heterogeneous label schemas across NER datasets pose challenges for multi-source data reuse and cross-domain transfer. Existing methods have clear limitations: manual mapping and label graph construction rely on domain expertise, which severely limits their scalability. On the other hand, approaches that rely solely on label semantics may overlook commonalities in dataset annotation practices and the contextual distributions of entities. To address these gaps, we propose ESLA, a label alignment framework constructing a unified multi-source training corpus for cross-domain transfer. In this work, ESLA is used exclusively to align and integrate the source corpora. It is not applied between the unified source corpus and FinReportNER, because target-domain entity annotations are assumed to be unavailable during model development.

As illustrated in Figure 1, the overall architecture of ESLA consists of three core modules: 1) Data Input & Shared Encoding: Utilizing a shared pre-trained language model, the contextual feature representations of entities in multiple labeled source datasets are uniformly extracted, providing the basic features for subsequent alignment. 2) Dual-Signal Matrix Computation: ESLA calculates label correspondence using two complementary signals. An empirical signal matrix captures similarities in

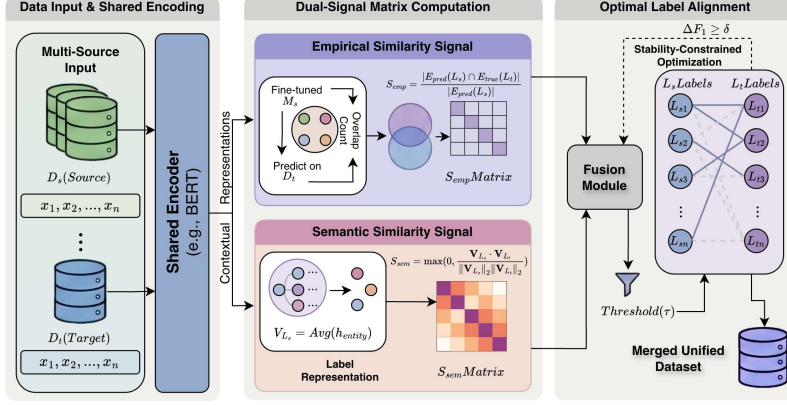


Fig. 1: The overall structure of the proposed framework

annotation behavior across datasets, while a semantic signal matrix measures the similarity between contextualized label prototypes. 3) Optimal Schema Alignment: The empirical and semantic signal matrices are integrated to compute a combined alignment score. The optimal alignment threshold is selected through stability-constrained optimization, balancing label coverage against acceptable degradation in source-domain NER performance. Based on the result of label alignment, ESLA constructs a unified multi-source corpus for downstream NER training and cross-domain transfer.

Theoretically, relationships between labels from different source datasets can be categorized into four relation types: (1) equivalence, (2) subset/superset, (3) partial overlap, and (4) disjointness. However, explicitly classifying these relations typically requires prior knowledge. Unlike previous approaches, our data-driven method bypasses the rigid classification process. Instead, it calculates a combined alignment score $S_{merge}(L_s, L_t)$ between each label pair. The final merging decision is dynamically controlled by the optimization threshold τ derived from our stability constraint module (see 3.3 for details). Although these four relationship types are not used as explicit supervisory signals, in practice, they naturally map onto the continuous score range as follows:

- High $S_{merge}(L_s, L_t)$ alignment scores typically correspond to equivalent labels (e.g., **name** in CLUENER and **PERSON** in OntoNotes);
- Intermediate $S_{merge}(L_s, L_t)$ scores generally reflect partial overlap or subset/superset relations (e.g., **company_name** in BosonNER and **ORG** in OntoNotes);
- Low $S_{merge}(L_s, L_t)$ scores typically indicate unrelated label pairs (e.g., **location** in BosonNER and **FAC** in OntoNotes).

3.1 Dual-Signal Matrix Computation

Before computing the alignment matrices, ESLA first unifies the feature space of heterogeneous datasets. Given a source dataset D_s with its label schema $\mathcal{L}_s$ and a target dataset D_t with its label schema $\mathcal{L}_t$, we employ a shared pre-trained language

model (e.g., BERT [33]) as the foundational encoder. For any given entity e composed of a token sequence $\{x_1, x_2, \dots, x_n\}$, the shared encoder maps it into a sequence of contextualized representations $\{H_e = \{h_1, h_2, \dots, h_n\}$, where $h_i \in \mathbb{R}^d$. Throughout this section, the terms source and target refer to the two datasets involved in pairwise label alignment, rather than the financial target domain used later for transfer evaluation.

Based on these unified contextual features, we calculate pairwise signals for each label pair (L_s, L_t) , where $L_s \in \mathcal{L}_s$ and $L_t \in \mathcal{L}_t$. This computation yields two complementary signals: the empirical signal and the semantic signal, as detailed below.

3.1.1 Empirical Signal

The empirical signal, denoted by $S_{emp}(L_s, L_t)$, measures how consistently a model trained on D_s predicts label L_s for instances that are annotated as L_t in D_t . Specifically, we first train an NER model M_s on the source dataset D_s , then evaluate it on the training set of the target dataset D_t to identify entities and compare the predicted labels with the ground-truth annotations. The calculation process is as follows: Suppose L_s is the label in the source dataset (such as **address** in CLUENER), and L_t is the label in the target dataset (such as **GPE** in OntoNotes); we use the NER model trained on D_s to predict all entities on the training set of D_t . For all entities in D_t that are truly labeled as L_t , the model may predict several entities as L_s . The empirical signal is defined as the proportion of entities truly labeled as L_t among the entities predicted as L_s by the model. The formula is as follows:

$$S_{emp}(L_s, L_t) = \frac{|E_{pred}(L_s) \cap E_{true}(L_t)|}{|E_{pred}(L_s)|}. \quad (1)$$

Herein, $E_{pred}(L_s)$ denotes the set of entities predicted as L_s by M_s in D_t , $E_{true}(L_t)$ represents the set of entities with the ground-truth label L_t , and $|\cdot|$ denotes the cardinality of a set. For instance, if M_s predicts 100 entities as **address** in D_t (i.e., $|E_{pred}(\text{address})| = 100$), and 70 of these are ground-truth annotated as **GPE** (i.e., $|E_{pred}(\text{address}) \cap E_{true}(\text{GPE})| = 70$), then $S_{emp}(\text{address}, \text{GPE}) = 70/100 = 0.7$. This signal measures the consistency between the annotation behavior learned from D_s and the annotation schema of D_t .

Notably, this empirical signal is asymmetric, which means that $S_{empirical}(L_s, L_t) \neq S_{empirical}(L_t, L_s)$. This directional discrepancy arises from variations in label granularity, domain shifts, and annotation guidelines between datasets. For instance, a model trained on OntoNotes might frequently predict **GPE** entities as **address**, but the reverse projection from CLUENER to OntoNotes may result in a different distribution. Consequently, the empirical signal must be calculated in both directions, which motivates the formulation of our optimal schema alignment phase (discussed in Section 3.2).

3.1.2 Semantic Signal

Although the empirical signal captures cross-dataset annotation behavior, it may become unreliable when the annotation overlap between datasets is limited. To address

this limitation, we introduce a complementary semantic signal by measuring the similarity between label prototypes in a shared contextual representation space, denoted as $S_{sem}(L_s, L_t)$.

Instead of relying on static word embeddings, we use contextual representations extracted from a shared encoder, Chinese-BERT-wwm-ext in our implementation [34]. This allows entity representations to encode contextual information and alleviate ambiguity caused by polysemy. The calculation process of the semantic signal is as follows.

For each label L from either the source or target dataset, we define its entity support set as $\mathcal{E}(L)$. For each entity span $e \in \mathcal{E}(L)$, we average the contextual token representations corresponding to the span to obtain its original entity representation $\mathbf{e}$. Contextual representations from pre-trained language models are known to exhibit anisotropic geometry [35]. Following representation normalization methods that aim to improve embedding-space isotropy [36], we apply global centering followed by L2 normalization:

$$\tilde{\mathbf{e}} = \frac{\mathbf{e} - \boldsymbol{\mu}}{\|\mathbf{e} - \boldsymbol{\mu}\|_2}, \quad (2)$$

where $\boldsymbol{\mu}$ denotes the global mean vector computed over all entity representations. We then construct the prototype representation of label L by averaging all normalized entity embeddings belonging to this label:

$$\mathbf{V}_L = \frac{1}{|\mathcal{E}(L)|} \sum_{e \in \mathcal{E}(L)} \tilde{\mathbf{e}}. \quad (3)$$

The semantic signal between a source label L_s and a target label L_t is computed using the cosine similarity between their prototype representations. Since negative cosine values indicate weak or opposite semantic association and should not contribute positive evidence for label alignment, we clip negative values to zero:

$$S_{sem}(L_s, L_t) = \max\left(0, \frac{\mathbf{V}_{L_s} \cdot \mathbf{V}_{L_t}}{\|\mathbf{V}_{L_s}\|_2 \|\mathbf{V}_{L_t}\|_2}\right). \quad (4)$$

Thus, $S_{sem}(L_s, L_t)$ takes values in $[0, 1]$, where a larger value indicates that the two labels are semantically closer in the contextual representation space. For example, even when two datasets have different vocabulary distributions, the aggregated prototypes of semantically related labels such as **address** and **location** may still obtain a high S_{sem} score. This semantic signal complements the empirical signal and helps alleviate the sparsity problem caused by limited annotation overlap.

3.1.3 Dual-Signal Fusion and Conflict Resolution

Empirical and semantic signals capture complementary information: empirical signal reflects the annotation norms of each dataset but may be influenced by data distribution; semantic signal captures the semantic proximity between labels but is sensitive to annotation noise. To fully leverage the advantages of both, we calculate the combined similarity $S_{merge}(L_s, L_t, \lambda)$ through linear interpolation:

$$S_{\text{merge}}(L_s, L_t, \lambda) = (1 - \lambda) \cdot S_{\text{sem}}(L_s, L_t) + \lambda \cdot S_{\text{emp}}(L_s, L_t) \quad (5)$$

where $\lambda \in [0, 1]$ is a tuning parameter used to balance the contributions of semantic signal and empirical signal. When the value of λ is low, the model leans more towards semantic signal, emphasizing the semantic closeness between labels; when the value of λ is high, the model pays more attention to empirical similarity, highlighting the commonalities in annotations. This formulation enables ESLA to balance each signal’s contribution according to the stability-constrained optimization described in Section 3.3.

3.2 Label Alignment Paths and Strategies

3.2.1 Sequential Merging Topology Search

When integrating multiple datasets $\mathcal{D} = \{D_1, D_2, \dots, D_N\}$, the number of possible merge orders grows factorially, making an exhaustive search computationally infeasible. To alleviate this complexity, we decompose the global integration task into a series of iterative pairwise dataset alignments.

To determine the sequential merging order, we propose a greedy directional path-selection strategy based on the asymmetric empirical signal. The empirical signal is used at this stage because it captures directional annotation compatibility: $S_{\text{emp}}(L_s, L_t)$ measures how well the annotation behavior learned from D_s can be expressed in the label schema of the target dataset D_t . In contrast, the semantic signal is primarily used to identify label-level semantic correspondence and provides limited information about the direction of dataset integration. This strategy prioritizes the merging of datasets with the highest label empirical signal, thereby reducing alignment uncertainty during subsequent merging at an early stage. The search process is formalized as follows:

1. Directional score initialization: For each ordered dataset combination (D_s, D_t) , we calculate the aggregated dataset-level directional score by summing the empirical signal $S_{\text{emp}}(L_s, L_t)$ within their label spaces. To ensure robustness, unaligned label pairs with zero empirical overlap will be ignored. Because the empirical signal is asymmetric, $D_s \rightarrow D_t$ and $D_t \rightarrow D_s$ are treated as two distinct candidate directions. A higher aggregated score indicates that more annotation behavior learned from D_s can be consistently expressed in the label schema of D_t .
2. Initial direction selection: The ordered dataset pair with the highest aggregated empirical signal score is selected as the first merging step. This alignment generates the intermediate merged dataset, denoted as D_{int} .
3. Iterative path expansion: In each subsequent round, we evaluate the empirical signal between the current intermediate dataset D_{int} and all other remaining datasets. The remaining dataset with the highest aggregated score to D_{int} is selected for the next pairwise alignment, and D_{int} is updated.
4. Termination condition: This incremental expansion continues until all N datasets in the collection successfully integrate into the unified pattern.

3.2.2 Deterministic Label Mapping Strategy

During pairwise label alignment, a source label may obtain high fused alignment scores with multiple candidate target labels. For example, `time` in BosonNER may align with both the `DATE` and `TIME` in OntoNotes. However, mapping one source label to multiple target labels would introduce ambiguous and conflicting supervision into the merged dataset. To prevent this, we enforce a deterministic single-target mapping constraint. Specifically, for each source label L_s , we first identify target labels whose fused alignment scores satisfy

$$S_{\text{merge}}(L_s, L_t; \lambda) \geq \tau. \quad (6)$$

If multiple target labels satisfy the threshold, L_s is mapped to the target label with the highest fused score:

$$L_t^* = \arg \max_{L_t \in \mathcal{L}_t} S_{\text{merge}}(L_s, L_t; \lambda). \quad (7)$$

If no target label satisfies the threshold, L_s remains unmerged. Each source label can therefore be mapped to at most one target label, whereas multiple source labels are allowed to map to the same target label. The strategy consequently supports many-to-one mappings while prohibiting one-to-many mappings.

3.3 Stability-Constrained Parameter Optimization

Unlike manually selecting alignment parameters, ESLA formulates parameter selection as a stability-constrained optimization problem to ensure the reliability of the merged dataset. Our objective is to maximize label coverage while constraining degradation in source-domain NER performance. We formulate this as :

$$\max_{\lambda, \tau} N_{\text{merged}} \quad \text{s.t.} \quad \Delta F_1 = F_1^{\text{merged}} - F_1^{\text{independent_training}} \geq -\delta, \quad (8)$$

where N_{merged} is the count of successfully merged label pairs under parameters (λ, τ) , and δ is the stability tolerance threshold. In this study, we set δ to 2%, providing a conservative tolerance for performance degradation while allowing additional label coverage.

The sequential merging path is fixed before the search over (λ, τ) . The path-selection stage uses only the directional empirical signal, whereas λ and τ control the fused dual-signal label mappings along the selected path. This separation avoids coupling the merging-order search with the subsequent stability-constrained parameter optimization.

The interpolation parameter λ governs the balance between empirical and semantic signal, while the threshold τ dictates whether a given label pair should be consolidated. We perform a systematic search over the hyperparameter space ($\lambda \in [0, 1]$, $\tau \in [0, 1]$). For each candidate pair (λ, τ) , we generate a candidate merged dataset and fine-tune the BERT-based NER model. If the resulting source-domain F1 decreases by more than δ , the merging configuration is rejected as unstable merging. Among the

remaining configurations, the one with the largest N_{merged} is selected. This optimization automatically filters out aggressive merging strategies such as extremely low τ that introduce excessive noise, thereby ensuring the robustness of the label alignment.

3.4 Context-Aware Label Completion

When fusing source datasets, some labels available in one dataset may be absent from another due to differences in annotation scope. For example, labels such as **CARDINAL**, **MONEY**, and **PERCENT** are defined in the OntoNotes but do not have relevant annotations in the CLUENER. As a result, mentions of these absent categories are implicitly treated as negative examples during training, leading to missing-label conflicts and introducing supervision bias after data fusion. To mitigate missing-label conflicts, we adopt a context-aware label completion based on pseudo-labeling [37].

Specifically, for labels that are absent from the source dataset, we use a pretrained NER model to generate pseudo-labels. For instance, to address the absence of the **CARDINAL** entity type in the CLUENER dataset, we fine-tune an NER model on the OntoNotes dataset and apply it to the CLUENER corpus to predict relevant entities. The same completion procedure was applied to other missing numerical and temporal entity categories.

Although pseudo-labeling inevitably introduces annotation noise, it provides positive supervision for entity types that would otherwise receive no supervision after dataset fusion. This trade-off is particularly beneficial for labels with distinctive contextual patterns, such as numeric and temporal entities, whose contextual representations remain relatively consistent across different datasets. Thus, the proposed context-based label completion strategy alleviates label conflicts arising from inconsistent annotation scopes and provides more complete supervision for multi-source NER training.

4 Experiments and Results

4.1 Experiment Setup

Our experiments consist of two stages. The first stage evaluates the quality of source dataset fusion using different label alignment strategies. The second stage evaluates cross-domain transfer by training models on individual source datasets or ESLA-Merge and testing them on FinReportNER under a target-unlabeled setting.

4.1.1 Datasets and Data Preprocessing

Datasets: We conduct the study on Chinese NER using three representative datasets as source datasets: the Chinese portion of OntoNotes 5.0, which includes news-like corpora with a relatively standardized entity label system and coarse granularity; CLUENER, derived from multiple sources such as news, encyclopedias, and social media, featuring finer label granularity, broader coverage across multiple domains, and a high-coverage label system; and BosonNER, primarily based on social media platforms like Weibo, with a relatively flat but practical label system. These datasets

differ in entity types, annotation styles, and domains, making them suitable as representative examples for integrating heterogeneous multi-source NER corpora. Specific details about the datasets are shown in Table 2. The numbers in parentheses indicate the quantity of entities for each type.

Table 2: Details of selected NER datasets

Name	Size (Sentences)	Entity types(#)
OntoNotes (Chinese)	48K	PERSON (13506), EVENT (1208), CARDINAL (8703), ORG (10363), DATE (10029), NORP (3214), GPE (19221), LOC (2565), MONEY (1452), WORK_OF_ART (1012), TIME (1847), ORDINAL (1408), QUANTITY (1058), FAC (1514), PRODUCT (375), PERCENT (1009), LANGUAGE (345), LAW (312)
BosonNER	2k	person_name (5141), location_name (4597), org_name (2689), time (4250), company_name (2374), product_name (4122)
CLUENER	13.4k	person (4112), organization (3419), position (3477), company (3263), address (3193), game (2612), government (2041), scene (1661), book (1283), movie (1259)

To evaluate cross-domain transfer, we construct FinReportNER from the Management Discussion and Analysis sections of annual reports published by Chinese A-share-listed semiconductor companies between 2020 and 2022. The corpus consists of 823 annotated sentences and nine entity categories: **RATIO**, **TIME**, **NUM**, **FTERM**, **INDUSTRY**, **ORG**, **TREND**, **PRODUCT**, **EVENT**. This dataset is not used for model training, validation, model selection, or threshold tuning. FinReportNER is used only as an evaluation benchmark, with its label definitions used to define the target-label mapping. During transfer evaluation, predictions from each source-trained model are mapped onto the FinReportNER label schema using predefined target-label projection rules. These rules are determined by label definitions and source-side alignment results and remain fixed across all experiments.

Data Preprocessing: Each source dataset (OntoNotes, CLUENER, and BosonNER) follows its official train, validation, and test split. Only the training sets are used for model training, label similarity calculation, and merging. Validation sets are used for model selection. Test sets are used for final evaluation. Furthermore, to enhance the efficiency of the grid search, we filter out invalid values when calculating S_{emp} to ensure stable signal calculations.

4.1.2 Model Architecture and Hyperparameters

We use Chinese-BERT-wwm-ext [34] with a Conditional Random Field (CRF) layer as the backbone sequence labeling architecture. While recent years have witnessed the proliferation of large generative models, encoder-only architectures remain the most stable, cost-effective baselines for classic NER tasks. We fine-tune the model with the AdamW optimizer [38, 39], using a learning rate of 3e-5 for the BERT layer and 3e-2 for the CRF layer, for up to 30 epochs. The batch size is 32, and the maximum sequence length is set to 150. All models are trained on an NVIDIA RTX 3080 Ti

GPU. After each epoch, the validation set is used to evaluate the model, and the model with the highest validation F1 score is saved as the final model.

4.1.3 Evaluation Metrics

Source-side Fusion Evaluation.

To evaluate whether source-side dataset fusion preserves the original NER capability, we report Precision, Recall, and Micro-F1 on the merged source corpus. Since this stage aims to assess the stability of the fused corpus, Micro-F1 is adopted as the primary evaluation metric.

Cross-domain Transfer Evaluation.

To evaluate transfer performance on the financial benchmark, we report Micro-F1, Macro-F1, and per-label F1 scores. Micro-F1 reflects the overall recognition performance across all entity instances, whereas Macro-F1 assigns equal importance to each target label and is therefore less affected by label imbalance.

4.1.4 Baseline Methods

To comprehensively evaluate the proposed ESLA framework, we compare it with both single-source training and representative label alignment strategies.

Single-source Models: No source dataset fusion is performed. Separate NER models are trained on CLUENER, BosonNER, and OntoNotes using identical model architectures and training settings. During the cross-domain evaluation, the predictions of each model are converted to the FinReportNER evaluation label space using the definition-based mapping protocol described in Section 4.1.5. These models serve as single-source transfer baselines and also provide no-fusion reference performance in the source-side evaluation.

Direct-Merge: The three source corpora are concatenated without label alignment, schema conflict resolution, or context-aware label completion. Labels from different datasets retain their original identities. This baseline measures the effect of naively combining heterogeneous corpora without applying any integration mechanism.

Rule-based Label Alignment: We implement a rule-based label alignment baseline as a simple and fully automated reference method, representing classic heuristic-based schema matching techniques [40, 41]. This approach relies solely on surface string matching, label names, and a few predefined rules, without leveraging entity context information, statistical evidence, or external knowledge. Specifically, first, we normalize label names. All label names are converted to lowercase, and abbreviations are expanded (e.g., `ORG` to organization, `GPE` to geopolitical entity). A source label L_s is merged into a target label L_t if and only if they satisfy an exact match condition or a substring inclusion constraint (e.g., `product.name` to `product`). All remaining label pairs are unmerged. This strategy is conservatively designed and has an extremely low computational cost.

Expert-driven Label Alignment: We further construct an expert-driven baseline to serve as a human reference for label alignment. In this setting, two co-authors independently analyzed the annotation guidelines and entity samples from the three

source datasets. They constructed a label mapping table based on semantic equivalence and taxonomic relationships. For example, creative works like `books` or `movies` are categorized as `WORK_OF_ART` instead of `PRODUCT`. This baseline is not intended to be automatic or scalable, but it provides a stable, interpretable human reference for evaluating automated alignment strategies.

Instruction-based LLM Label Alignment: We additionally evaluate an instruction-based LLM baseline (GPT-5), inspired by recent advances in LLM-driven data integration [42, 43]. For each candidate label pair (L_s, L_t) , we construct a label profile by collecting the top- k most frequent entities of each label from the corresponding training corpora, together with their contexts. We fix $k = 30$ to control prompt length and reduce the risk of hallucination. The prompt instructs LLMs to infer label-level alignment decisions from the provided profiles. For each label pair, the LLM needs to produce a binary alignment decision (`merge` or `not_merge`), specify the corresponding set-theoretic relation (equivalent, superset/subset, overlap, or disjoint), and provide a confidence score normalized between 0 and 1. The complete prompt template is provided in the Supplementary Note.

For all alignment methods, the resulting label mappings are used to construct a unified multi-source corpus. The same NER model and training configuration are then applied to all merged corpora to ensure a fair comparison.

4.1.5 Target-Label Mapping for Cross-Domain Evaluation

Since the source domain model and FinReportNER use different label naming or annotation schemes, we define a fixed label mapping for cross-domain evaluation. This mapping is derived from the label definitions and annotation guidelines of both the source and target domains, without relying on any entity instances, label distributions, or evaluation results from FinReportNER. The two co-authors who constructed the Expert-based baseline independently developed candidate target-label mappings based solely on label names, definitions, and annotation guidelines. Any discrepancies between the candidate mappings were resolved by re-examining the annotation definitions. The mapping is then fixed and uniformly applied to all single-source and multi-source models.

The six transferable target labels are `ORG`, `PRODUCT`, `TIME`, `EVENT`, `NUM`, and `RATIO`. Specifically, organization-related source labels are mapped to `ORG`; product-related labels are mapped to `PRODUCT`; `DATE` and `time` are mapped to `TIME`; currency label `MONEY` is mapped to `NUM`; `PERCENT` is mapped to `RATIO`; and `EVENT` retains its original label. Labels without corresponding target categories are excluded from the transfer evaluation. The complete mapping table is provided in Supplementary Table S1.

4.2 Main Results: Comparative Analysis

4.2.1 Source-side Fusion Stability

This section reports the quality of the source dataset fusion produced by different label alignment strategies.

Table 3: Source-side Evaluation of Different Label Alignment Strategies

Method	Merged Labels	Precision	Recall	Micro-F1
Independent Training	N/A	0.79	0.81	0.80
Rule-based Alignment	8	0.76	0.78	0.77
Expert-driven Alignment	13	0.77	0.80	0.79
LLM Alignment (GPT-5)	11	0.78	0.80	0.79
ESLA (Ours)	15	0.78	0.81	0.80

The sequential merging path is first determined using the asymmetric empirical signal, resulting in CLUENER $\rightarrow$ BosonNER $\rightarrow$ OntoNotes. Along this fixed path, specific label alignments are determined using the fused dual-signal score S_{merge} , with (λ, τ) selected through the stability-constrained optimization described in Section 3.3.

Overall Performance Analysis. The results on the source-side merged corpus are summarized in Table 3. Under the optimal parameter setting ($\lambda \in \{0.5, 0.6\}$, $\tau=0.3$), the ESLA framework successfully aligned the maximum number of labels while strictly maintaining the Micro-F1 score (0.80), consistent with independent training. In contrast, rule-based methods suffered lower label coverage due to rigid string-matching constraints, failing to capture implicit semantic mappings. Meanwhile, although LLM-based (GPT-5) and expert-driven approaches preserve a comparable Micro-F1 score, they achieve fewer overall aligned label pairs, demonstrating the crucial role of the dual-signal framework proposed in this paper.

Fine-grained Label Evaluation. To provide a more detailed understanding of label alignment behavior, Table 4 and 5 present a detailed comparison of label-level performance and label alignment paths. Overall, the rule-based alignment approach underperforms compared to other methods, while expert-driven, LLM-based, and ESLA alignments show generally similar performance. The rule-based approach relies on strict, coarse heuristics that often produce inaccurate or overly cautious mappings for labels like **time** and **location**. It also does not combine related categories such as **company**, **organization**, and **government** under **ORG**, leading to a fragmented label space and reduced discriminative ability. For labels with clear semantic equivalence, expert, LLM-based, and ESLA maintain a high degree of stability (such as **PERSON** and **DATE**). An exception is that LLM-based (GPT-5) fails to resolve the mapping of **time** to **DATE** or **TIME**. For subset relations, labels such as **company**, **government**, and **organization** are consistently mapped to **ORG**. However, for the **address** and **scene** labels, we observe a challenge. ESLA aligns both of them to the intermediate label **location** (from BosonNER). Subsequently, ESLA mapped the entire **location** to **GPE**. Although this strategy successfully improves consistency for **address**, it may introduce ambiguity for **scene**. This alignment choice reasonably explains the slight performance decline observed in **GPE**. We also observe significant differences in the partial overlap relations of the **PRODUCT** category. The expert-driven and LLM-based baselines classified creative works (**book**, **movie**) as **WORK_OF_ART**, whereas ESLA aligned them under **PRODUCT** based on dataset evidence, trading off label purity for broader coverage. Even if **EVENT** is not explicitly merged by any strategy, ESLA still performs better on this label. This indicates that improving the consistency of the unified label schema may also positively affect categories that are not directly merged.

Table 4: Label-level F1-score Comparison in the Unified Corpus

Label	Support	Independent Training	Rule-based	Expert-driven	LLM (GPT-5)	ESLA (Ours)
GPE	1778	0.85	0.84	0.85	0.84	0.84
PERSON	1261	0.92	0.91	0.92	0.92	0.93
ORG	1105	0.79	0.73	0.78	0.78	0.79
DATE	976	0.82	0.78	0.82	0.81	0.82
CARDINAL	742	0.67	0.66	0.66	0.67	0.67
LOC	268	0.72	0.64	0.72	0.70	0.72
NORP	245	0.64	0.63	0.64	0.61	0.61
PERCENT	177	0.87	0.86	0.86	0.87	0.88
TIME	160	0.67	0.61	0.67	0.66	0.67
MONEY	156	0.91	0.90	0.90	0.91	0.91
FAC	155	0.61	0.54	0.60	0.59	0.60
QUANTITY	135	0.60	0.58	0.61	0.57	0.60
ORDINAL	126	0.83	0.82	0.81	0.78	0.81
EVENT	100	0.66	0.65	0.66	0.65	0.70
WORK_OF_ART	63	0.55	0.55	0.57	0.56	0.58
PRODUCT	35	0.58	0.40	0.37	0.38	0.47
LAW	17	0.63	0.59	0.60	0.44	0.56
LANGUAGE	8	0.70	0.70	0.70	0.70	0.70
Micro-F1	7507	0.80	0.77	0.79	0.79	0.80

Table 5: Comparison of Label Alignment Decisions and Mapping Paths across Different Strategies

Source Label (L_s)	Rule-based	Expert-driven	LLM (GPT-5)	ESLA (Ours)
name	→ person_name → PERSON	→ person_name → PERSON	→ person_name → PERSON	→ person_name → PERSON
time	→ TIME	→ DATE	Conflict (DATE/TIME)	→ DATE
address/scene	N/A	address → GPE scene → FAC	Conflict (GPE/LOC/FAC)	→ location → GPE
movie/book	N/A	→ WORK_OF_ART	→ WORK_OF_ART	→ product_name → PRODUCT
game	N/A	N/A	→ PRODUCT	→ product_name → PRODUCT
company	→ company_name	→ company_name → ORG	→ company_name → ORG	→ company_name → ORG
organization	→ org_name → ORG	→ org_name → ORG	→ org_name → ORG	→ org_name → ORG
government	N/A	→ org_name → ORG	→ ORG	→ org_name → ORG

4.2.2 Transfer Performance on the Financial Domain

Table 6 compares three single-source models with five multi-source training strategies on FinReportNER. Single-source performance is uneven across target categories: CLUENER and BosonNER obtained zero F1 for NUM and RATIO, whereas OntoNotes provided the strongest results overall. Directly merging three corpora also produce weak cross-domain transfer performance. Direct-Merge achieved a Micro-F1 of 0.22 and a Macro-F1 of 0.17, compared with 0.56 and 0.57 for OntoNotes. This

Table 6: Cross-domain Transfer Performance on the Financial Benchmark (FinReport-NER)

Training Corpus	Micro-F1	Macro-F1	ORG	PRODUCT	TIME	EVENT	NUM	RATIO
CLUENER	0.15	0.11	0.65	0.01	0.00	0.00	0.00	0.00
BosonNER	0.25	0.19	0.65	0.02	0.49	0.00	0.00	0.00
OntoNotes	0.56	0.57	0.68	0.01	0.87	0.00	0.89	0.97
Direct-Merge	0.22	0.17	0.65	0.01	0.36	0.00	0.00	0.00
Rule-based	0.52	0.55	0.66	0.01	0.76	0.00	0.92	0.97
Expert-based	0.56	0.58	0.69	0.02	0.88	0.00	0.92	0.99
LLM (GPT-5)	0.56	0.58	0.69	0.05	0.83	0.00	0.93	0.98
ESLA-Merge	0.56	0.59	0.69	0.06	0.89	0.00	0.93	0.99

contrast shows that naive corpus aggregation is ineffective under unresolved schema and missing-label conflicts.

The alignment-based strategies largely recovered the performance lost through direct merging. ESLA-Merge matched the highest observed Micro-F1 of 0.56 and achieved the highest observed Macro-F1 of 0.59. Its Macro-F1 was 0.01 higher than both the Expert-based and LLM baselines. This result indicates more balanced performance across the six transferable target categories.

At the fine-grained label level, ESLA-Merge achieve the highest **PRODUCT** and **TIME** scores among all evaluated configurations. **PRODUCT** F1 reaches 0.06, three times the best single-source score of 0.02 and 0.01 above the LLM baseline. **TIME** reached 0.89, exceeding the best single-source score by 0.02 and the Expert-based baseline by 0.01. ESLA-Merge also get the highest scores for **ORG**, **NUM**, and **RATIO**. Relative to the best single-source scores, these values get a increase of 0.01, 0.04, and 0.02, respectively. The ablations in Section 4.4 separately examine label completion for numerical and temporal categories and product-related mappings for **PRODUCT**.

4.3 Analysis of the Alignment Process

4.3.1 Empirical Signal and Alignment Path

We calculated the empirical signal matrices for each pair of datasets (CLUENER, BosonNER, and OntoNotes) to provide a quantitative basis for label alignment and path selection during merging. Therefore, we conducted six experiments, considering bidirectional calculation: CLUENER $\leftrightarrow$ BosonNER, CLUENER $\leftrightarrow$ OntoNotes, and BosonNER $\leftrightarrow$ OntoNotes. For each direction $D_s \rightarrow D_t$, a NER model trained on the source training set D_s is used to predict entities on the target training set D_t , from which $S_{\text{empirical}}(L_s, L_t)$ is computed. All resulting empirical signal matrices are visualized as heatmaps in Figure 2.

The results show an advantage in label mapping between CLUENER and BosonNER. Fine-grained labels such as **book**, **movie**, and **game** in CLUENER can be merged into **product_name** in BosonNER. At the same time, **organization** and **government** highly overlap with **org_name**, indicating that the label granularity of CLUENER is finer and BosonNER is coarser-grained. Based on the aggregated directional empirical

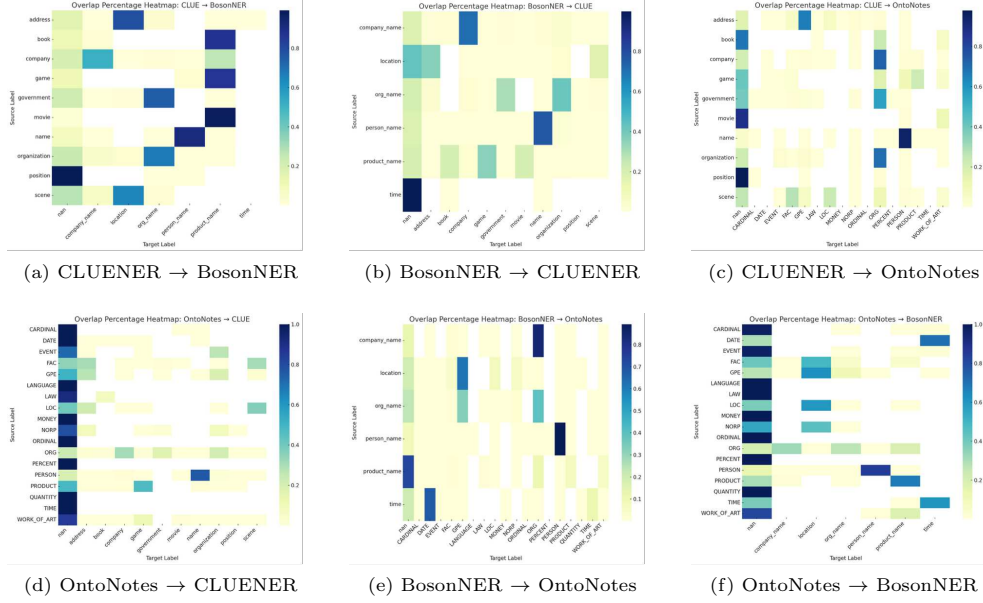


Fig. 2: Empirical signal heatmaps across dataset pairs

scores across the ordered dataset pairs, CLUENER $\rightarrow$ BosonNER obtains the highest value. This indicates that the annotation behavior learned from CLUENER can be most consistently expressed in the BosonNER label schema. Therefore, the initial merging path is selected as CLUENER $\rightarrow$ BosonNER, forming an intermediate dataset (denoted as BosonM). Further analysis found that labels such as **ORG**, **GPE**, and **PRODUCT** in OntoNotes have high correspondence with those in BosonM. Among the remaining candidate directions, BosonM $\rightarrow$ OntoNotes achieves the highest aggregated directional empirical score and is therefore selected as the second merging step. The final unified corpus consequently adopts the OntoNotes label schema.

4.3.2 Semantic Signal

To further analyze the semantic signal of named entity labels across datasets and facilitate label alignment, we conducted two experiments. These experiments used the alignment path selected in Section 4.3.1 to compute and visualize a semantic signal matrix and a 2D embedding graph. The process employed the Chinese-BERT-wwm-ext model for entity representation, combined with mean pooling, centralization, and normalization. Ultimately, cosine similarity was used to measure the semantic signal between labels.

In the first experiment, we analyzed the semantic signal between CLUENER and BosonNER labels. Figure 3 shows that CLUENER’s fine-grained labels, such as **book**, **movie**, and **game**, are semantically close to **product_name** in BosonNER. This reflects the ability of **product_name** in BosonNER to act as a semantically inclusive category.

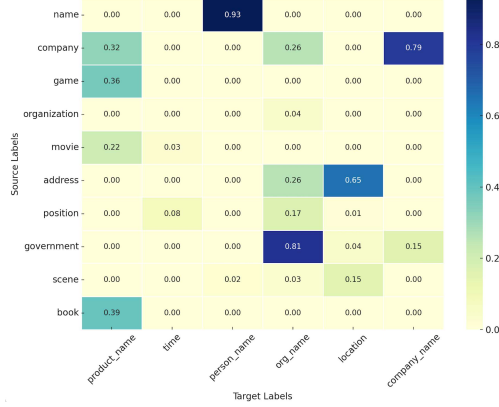


Fig. 3: Semantic signal between CLUENER and BosonNER

Furthermore, **government** aligns closely with **org_name** (0.81), while **organization** (0.04) exhibits negligible correlation, indicating that CLUENER’s **government** tends to converge semantically in BosonNER’s organizational category. Similarly, **address** in CLUENER matches well with **location** (0.65) in BosonNER, indicating overlap in spatial references. Notably, **person_name** in BosonNER displays a very high similarity with **name** (0.93) in CLUENER. They can be regarded as the same category. Overall, the heatmap shows that most BosonNER labels are semantically similar to CLUENER labels. This indicates a very high possibility for merging.

In the second experiment, we computed the semantic signal between the intermediate dataset BosonM and OntoNotes. As shown in Figure 4, **location** in BosonM shows strong similarity with **GPE** (0.80), **LOC** (0.46), and **NORP** (0.48), indicating that this label covers geopolitical regions, general places, and demographic groups. Similarly, **org_name** closely aligns with **ORG** (0.83), confirming their shared focus on organizations. Interestingly, **company_name** in BosonM shows moderate semantic signal with both **ORG** (0.48) and **PRODUCT** (0.45), reflecting its relevance across business and product contexts. Similarly, **product_name** in BosonM correlates strongly with **PRODUCT** (0.68) and moderately with **WORK_OF_ART** (0.51), supporting its coverage of both normal goods and cultural products. **time** in BosonM maps effectively to **DATE** (0.87) and **TIME** (0.54), indicating that it tends to mark more date-related entities. Likewise, **person_name** in BosonM demonstrates high similarity with **PERSON** (0.93) in OntoNotes, reinforcing its robustness in personal entity alignment. Overall, the semantic signal matrix in Figure 4 confirms that BosonM can be effectively mapped onto OntoNotes, providing support for continued label merging and unified training.

To further verify the overall semantic relationships among entity labels across datasets, we visualized the average embedding vectors of all labels using t-distributed Stochastic Neighbor Embedding (t-SNE) [44], as shown in Figure 5. The resulting distribution reveals several semantically coherent local regions across the three datasets:

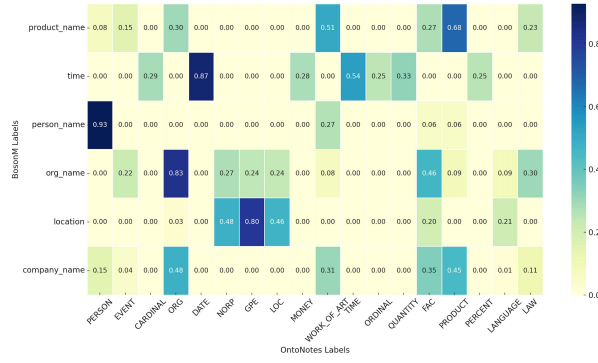


Fig. 4: Semantic signal between BosonM and OntoNotes

- Person-related labels (**PERSON**, **person_name**, **name**) are located in a similar region, suggesting that these variants share similar contextual representations and may be considered potential alignment candidates.
- Organization-related labels (**ORG**, **company_name**, **org_name**, **company**, **government**, and **organization**) are distributed within a broader central region, indicating a degree of semantic overlap among organization- and company-related entities.
- Location-related labels (**LOC**, **GPE**, **location**, **FAC**, **address**, and **scene**) are mainly distributed in the upper region, suggesting that their contextual representations capture related location semantics despite differences in annotation granularity.
- Product-related and work-related labels exhibit two related clusters. In particular, **PRODUCT** and **product_name** are positioned relatively close to each other, whereas **book**, **movie**, and **game** are more closely associated with **WORK_OF_ART**. This pattern reflects the distinction between general product entities and creative works in different annotation schemas.
- Numeric labels, including **QUANTITY**, **MONEY**, **PERCENT**, **CARDINAL**, and **ORDINAL**, form a relatively coherent region, indicating shared contextual characteristics while retaining differences among their finer-grained semantic categories.

One notable observation is that some labels do not align as closely as theoretically expected. For example, organization-related labels are located in the same broad semantic region but do not form a compact cluster. Despite **organization** having a semantically similar name to **org_name** and **ORG**, it is clearly separated from them in the projected embedding space. Similarly, semantically related labels may still be separated because their representations are influenced by dataset-specific contexts, annotation granularity, and entity distributions. This observation highlights that semantic signal alone may not be sufficient to determine whether two labels should be aligned. Therefore, incorporating empirical signal based on actual prediction behavior is essential for achieving more robust and reliable label alignment.

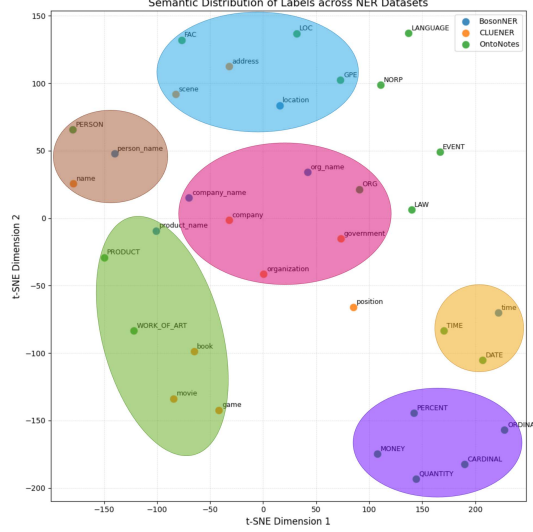


Fig. 5: Semantic distribution of labels across NER datasets

4.4 Ablation Study and Parameter Sensitivity

To evaluate the contribution of each component in ESLA, we conduct ablation experiments at two levels: first, source-side label alignment behavior (Table 7), and cross-domain transfer performance on FinReportNER (Table 8). We further perform an additional analysis of mapping decisions for **PRODUCT** (Table 9). All ablation settings use the same downstream training configuration as the full model. The merging threshold is fixed at $\tau = 0.3$, as identified as the optimal setting in Section 4.2.1. Specifically, w/o semantic signal setting sets $\lambda = 1$, retaining only the empirical signal; w/o empirical signal setting sets $\lambda = 0$, retaining only the semantic signal; w/o stability constraint removes the $\Delta F1 \geq -\delta$ condition in equation 8 and directly selects the configuration with the maximum number of aligned labels; and w/o label completion disables context-aware label completion.

Source-side Alignment Behavior. As shown in Table 7, the full ESLA aligns 15 labels while maintaining a Micro-F1 score of 0.80. Removing the semantic signal, the number of aligned labels drops to 11, with unaligned labels including **book**, **movie**, **game**, and **product_name**. These labels differ from target labels across granularity or annotation schema, making it difficult to match them solely based on empirical signal. When the empirical signal is removed, **organization** and **scene** fail to align successfully, further fragmenting the label space. Without the stability constraint, the number of aligned labels increases to 17, but incorrectly aligns **position** into **org_name**, dropping Micro-F1 to 0.77, exceeding the tolerance of $\delta = 2\%$. This confirms that the stability constraint prevents overly aggressive configurations rather than maximizing the number of alignments. Disabling label completion does not change the number of aligned labels but reduces Micro-F1 to 0.76.

Table 7: Ablation study of ESLA under different component configurations (fixing $\tau = 0.3$)

Method Variant	Aligned Labels	Micro-F1	Alignment Changes (vs. Full)
ESLA (Full)	15	0.80	None
w/o Semantic Signal	11	0.78	– <code>book</code> , <code>movie</code> , <code>game</code> , <code>product_name</code>
w/o Empirical Signal	13	0.79	– <code>organization</code> , <code>scene</code>
w/o stability constraint	17	0.77	+ <code>position</code>
w/o Label Completion	15	0.76	None

Table 8: Ablation study of ESLA components on FinReportNER.

Method Variant	Micro-F1	Macro-F1	ORG	PRODUCT	TIME	EVENT	NUM	RATIO
ESLA (full)	0.56	0.59	0.69	0.06	0.89	0.00	0.93	0.99
w/o semantic signal	0.56	0.58	0.69	0.01	0.89	0.00	0.93	0.99
w/o empirical signal	0.55	0.58	0.63	0.06	0.89	0.00	0.93	0.99
w/o stability constraint	0.56	0.58	0.65	0.06	0.89	0.00	0.93	0.99
w/o label completion	0.21	0.16	0.65	0.02	0.28	0.00	0.00	0.00

Downstream Transfer Performance. Table 8 shows clearer differences on FinReportNER. Without the semantic signal, the F1 score for `PRODUCT` drops from 0.06 to 0.01, while other categories remain nearly unchanged. This corresponds to the loss of product-related mappings such as `book`, `movie`, `game`, and `product_name`. Without the empirical signal, `ORG` decreases from 0.69 to 0.63, suggesting that the consolidation of organization labels relies more heavily on cross-dataset empirical statistical evidence. When the stability constraint is removed, `ORG` score drops to 0.65, demonstrating that incorrect alignment such as `position`→`org_name` introduces noise, meaning that larger fusion scales do not necessarily lead to better transfer performance. Label completion has a significant impact. After removing this module, the Micro-F1 drops from 0.56 to 0.21, while `NUM` and `RATIO` fall to zero, and `TIME` decreases from 0.89 to 0.28. Numerical and temporal entities are unannotated in some source corpora and may therefore become implicit negative examples after direct merging. Although label completion produces only a 0.04 decrease in Micro-F1 on the source side, it causes a 0.35 gap in the target domain. This result indicates the importance of evaluating both source-side alignment quality and downstream transfer performance.

Analysis of Product-related Mapping Decisions. To further identify the sources of `PRODUCT`’s performance improvement, we examined the impact of aligning `book`, `movie`, and `game` to `PRODUCT` on downstream results. As shown in Table 9, after incorporating these three alignment relations into the expert mapping, the F1 score for `PRODUCT` increased from 0.02 to 0.06, and the Macro-F1 also rose from 0.58 to 0.59—consistent with the results of full ESLA. Conversely, when these three alignment relations were removed from ESLA, the `PRODUCT` F1 dropped to 0.01, matching the performance after removing semantic signals. This indicates that ESLA’s advantage on the `PRODUCT` category primarily stems from the cross-granularity alignment

Table 9: Effect of mapping `book`, `movie`, and `game` to the target label `PRODUCT`. A $\checkmark$ indicates that the source label is mapped to `PRODUCT`, whereas $\times$ indicates that it is excluded from the target schema.

Alignment Configuration	book	movie	game	PRODUCT	Macro-F1
Expert-based	$\times$	$\times$	$\times$	0.02	0.58
Expert-based (modified)	$\checkmark$	$\checkmark$	$\checkmark$	0.06	0.59
LLM (GPT-5)	$\times$	$\times$	$\checkmark$	0.05	0.58
LLM (GPT-5) (modified)	$\checkmark$	$\checkmark$	$\checkmark$	0.06	0.59
ESLA (modified)	$\times$	$\times$	$\times$	0.01	0.58
ESLA	$\checkmark$	$\checkmark$	$\checkmark$	0.06	0.59

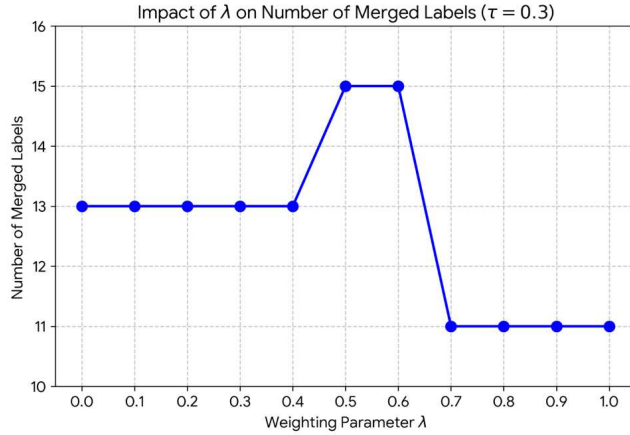


Fig. 6: Impact of the weighting parameter λ on the number of merged labels ($\tau = 0.3$)

relations identified by semantic signals. Notably, the LLM baseline aligned only `game` to `PRODUCT`, achieving an F1 score of 0.05; however, adding both `book` and `movie` further improved it to 0.06. This result demonstrates that introducing semantically related source labels can significantly alleviate the issue of insufficient supervision for the target category.

Parameter Sensitivity. We analyze the sensitivity of ESLA to the weight parameter λ . As shown in Figure 6, the number of merged labels remains stable at 13 for $\lambda \in [0.0, 0.4]$, reaches a maximum of 15 at $\lambda \in \{0.5, 0.6\}$, and then decreases to 11 when $\lambda \geq 0.7$. This pattern indicates that balancing empirical and semantic signals provides the strongest alignment coverage, whereas excessive reliance on the empirical signal reduces the number of valid alignments. Overall, with $\tau = 0.3$, $\lambda \in \{0.5, 0.6\}$ provides the most balanced alignment performance.

5 Conclusion and Future Work

This paper presents ESLA, an automated label alignment framework for multi-source cross-domain NER transfer under a target-unlabeled setting. ESLA combines empirical prediction consistency and contextual semantic signal to estimate cross-dataset label correspondences, employs sequential merging path search with stability-constrained optimization to construct a unified source corpus, and introduces label completion to mitigate missing-label conflicts caused by heterogeneous annotation schemes. Experimental results on three public Chinese NER datasets and the financial benchmark FinReportNER demonstrate that ESLA constructs a stable unified source corpus while improving cross-domain transfer performance. Compared with single-source training, ESLA-Merge achieves the best Macro-F1 and provides broader transferable label coverage, particularly for target categories that receive limited supervision from individual source datasets.

Although ESLA demonstrates effective label alignment across heterogeneous NER datasets, it currently adopts a deterministic single-target mapping strategy. This design is suitable for fine-to-coarse or many-to-one alignment, where several fine-grained labels can be merged into a broader category. However, it cannot explicitly model coarse-to-fine or one-to-many correspondences, which may occur when a broad label in one dataset corresponds to several finer-grained labels in another dataset. Future work will therefore focus on two directions. First, we plan to investigate more expressive label relationship modeling, such as hierarchical representations and graph-based reasoning, to better capture subset and partial-overlap relations between heterogeneous label schemas. Second, we will extend ESLA to broader application domains and multilingual scenarios to further evaluate its applicability across diverse NER resources.

Declarations

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 programme under grant agreement ID 101007666 (ESPERANTO).

Competing Interests

The authors declare no competing interests.

Data Availability Statement

CLUENER2020 is available from <https://github.com/CLUEbenchmark/CLUENER2020>. OntoNotes 5.0 is available from the Linguistic Data Consortium under license (LDC2013T19; <https://catalog.ldc.upenn.edu/LDC2013T19>). BosonNER is available from <https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/boson/>. FinReportNER, which was generated for the target-domain evaluation in this study, is available from the corresponding author on reasonable request.

Ethics statement

This study did not involve human research participants; therefore, ethics approval and informed consent were not applicable.

Author contributions

Xiaobo Zhang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Congqing He: Data curation, Investigation, Software, Visualization, Writing – review & editing. Ying He: Data curation, Investigation, Writing – review & editing. Jian Peng: Data curation, Investigation. Dajie Fu: Data curation, Investigation. Tien-Ping Tan: Supervision, Writing – review & editing. All authors reviewed and approved the final manuscript and agreed to be accountable for their respective contributions.

References

1. Qu, X. *et al.* A survey on Arabic named entity recognition: past, recent advances, and future trends. *IEEE Transactions on Knowledge and Data Engineering* **36**, 943–959 (2023).
2. Bade, G. Y., Kolesnikova, O. & Oropeza, J. L. The role of named entity recognition (NER): survey. *International Journal of Computer & Organization Trends* **14**, 1–7 (2024).
3. Hu, Z., Hou, W. & Liu, X. Deep learning for named entity recognition: a survey. *Neural Computing and Applications* **36**, 8995–9022 (2024).
4. Kaur, N., Saha, A., Swami, M., Singh, M. & Dalal, R. *BERT-NER: a transformer-based approach for named entity recognition. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–7 (2024).
5. Kumar, S., ElKholy, M., Liu, D. & Boulenger, A. *Bridging the gap: efficient cross-lingual NER in low-resource financial domain. Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, 54–62 (Association for Computational Linguistics, 2025).
6. Singh, D., Barve, S. & Dwivedi, A. K. OptiASAR: optimized aspect sentiment analysis with BiLSTM-GRU and NER-BERT in healthcare decision-making. *IEEE Access* **13**, 47459–47474 (2025).
7. Huang, X., Dong, L., Boschee, E. & Peng, N. *Learning a unified named entity tagger from multiple partially annotated corpora for efficient adaptation. Proceedings*

- of the 23rd Conference on Computational Natural Language Learning (CoNLL), 515–527 (Association for Computational Linguistics, Hong Kong, China, 2019). URL <https://aclanthology.org/K19-1048/>.
8. Palen-Michel, C., Pickering, M., Kruse, M., Sälevä, J. & Lignos, C. *OpenNER 1.0: standardized open-access named entity recognition datasets in 50+ languages. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 33649–33674 (Association for Computational Linguistics, Suzhou, China, 2025). URL <https://aclanthology.org/2025.emnlp-main.1708/>.
 9. Oi, T. & Miwa, M. *Enhancing NER by harnessing multiple datasets with conditional variational autoencoders. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1107–1117 (Association for Computational Linguistics, Vienna, Austria, 2025). URL <https://aclanthology.org/2025.acl-short.87/>.
 10. Weischedel, R. *et al.* OntoNotes release 5.0 (2013). URL <https://catalog.ldc.upenn.edu/LDC2013T19>. LDC2013T19, Linguistic Data Consortium, Philadelphia.
 11. Min, K., Ma, C., Zhao, T. & Li, H. *BosonNLP: an ensemble approach for word segmentation and POS tagging. Natural Language Processing and Chinese Computing*, 520–526 (Springer International Publishing, Cham, 2015).
 12. Zhao, J. *et al.* *Joining datasets via data augmentation in the label space for neural networks. Proceedings of the 38th International Conference on Machine Learning*, Vol. 139, 12686–12696 (PMLR, 2021). URL <https://proceedings.mlr.press/v139/zhao21b.html>.
 13. Arazo, E., Ortego, D., Albert, P., O’Connor, N. E. & McGuinness, K. *Pseudo-labeling and confirmation bias in deep semi-supervised learning. 2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8 (IEEE, 2020).
 14. Lu, J., Zhao, R., Mac Namee, B. & Tan, F. PUnifiedNER: a prompting-based unified NER system for diverse datasets. *Proceedings of the AAAI Conference on Artificial Intelligence* **37**, 13327–13335 (2023).
 15. Ivanova, T. I. *New perspectives of ontology alignment using large language models. 2025 International Conference on Information Technologies (InfoTech)*, 1–4 (IEEE, 2025).
 16. Musa, M. H. *et al.* Integrating and retrieving learning analytics data from heterogeneous platforms using ontology alignment: graph-based approach. *MethodsX* **14**, 103092 (2025).
 17. Macilenti, G., Stellato, A. & Fiorelli, M. SIS: leveraging semantically informed similarity of text embeddings for enhanced ontology alignment. *Procedia Computer*

Science **270**, 505–514 (2025).

18. Cui, L., Wu, Y., Liu, J., Yang, S. & Zhang, Y. *Template-based named entity recognition using BART. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1835–1845 (Association for Computational Linguistics, Online, 2021). URL <https://aclanthology.org/2021.findings-acl.161/>.
19. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022). URL <https://arxiv.org/abs/2203.02155>.
20. Zhu, Y. & Liu, Y. *LLM-NER: advancing named entity recognition with LoRA+ fine-tuned large language models. 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI)*, 364–368 (IEEE, 2025).
21. Wang, S. *et al.* *GPT-NER: named entity recognition via large language models. Findings of the Association for Computational Linguistics: NAACL 2025*, 4257–4275 (Association for Computational Linguistics, 2025).
22. Xu, W., Dang, R. & Huang, S. *LLM’s weakness in NER doesn’t stop it from enhancing a stronger SLM. Proceedings of the Second Workshop on Ancient Language Processing*, 170–175 (Association for Computational Linguistics, The Albuquerque Convention Center, Laguna, 2025). URL <https://aclanthology.org/2025.alp-1.21/>.
23. Tjong Kim Sang, E. F. & De Meulder, F. *Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147 (Association for Computational Linguistics, 2003).
24. Levow, G.-A. *The third international Chinese language processing bakeoff: word segmentation and named entity recognition. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108–117 (2006).
25. Zhang, J. Y., He, G., Dai, Z. & Liu, Y. Named entity recognition of enterprise annual reports integrated with BERT. *Journal of Shanghai Jiao Tong University* **55**, 117–123 (2021).
26. Loukas, L. *et al.* *FiNER: financial numeric entity recognition for XBRL tagging. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4419–4431 (Association for Computational Linguistics, Dublin, Ireland, 2022). URL <https://aclanthology.org/2022.acl-long.303/>.
27. Shah, A., Vithani, R., Gullapalli, A. & Chava, S. FiNER: financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157* (2023). URL <https://arxiv.org/abs/2302.11157>.

28. He, Y., Chen, J., Antonyrajah, D. & Horrocks, I. *BERTMap: a BERT-based ontology alignment system*. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 5684–5691 (2022).
29. Zhong, Z. *et al.* *A new ontology matching method based on large language models and graph structure learning*. *Neural Computing for Advanced Applications*, Vol. 2665 of *Communications in Computer and Information Science*, 131–145 (Springer, 2025).
30. Amini, R., Saki Norouzi, S., Hitzler, P. & Amini, R. *Towards complex ontology alignment using large language models*. *Iberoamerican Conference on Knowledge Graphs and Semantic Web* (2024).
31. Ushio, A. & Camacho-Collados, J. T-NER: an all-round Python library for transformer-based named entity recognition. *arXiv preprint arXiv:2209.12616* (2022). URL <https://arxiv.org/abs/2209.12616>.
32. Ma, J. *et al.* *Label semantics for few-shot named entity recognition*. *Findings of the Association for Computational Linguistics: ACL 2022*, 1956–1971 (Association for Computational Linguistics, Dublin, Ireland, 2022). URL <https://aclanthology.org/2022.findings-acl.155/>.
33. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *BERT: pre-training of deep bidirectional transformers for language understanding*. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019). URL <https://aclanthology.org/N19-1423/>.
34. Cui, Y., Che, W., Liu, T., Qin, B. & Yang, Z. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **29**, 3504–3514 (2021).
35. Ethayarajh, K. *How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings*. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65 (Association for Computational Linguistics, Hong Kong, China, 2019).
36. Gao, S., Dou, S., Zhang, Q. & Huang, X. *Kernel-whitening: overcome dataset bias with isotropic sentence embedding*. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 4112–4122 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
37. Thiam, P. I., Chasseray, Y., Mothe, J., Roche, M. & Teisseire, M. *Enhancing domain-specific named entity recognition via segmentation and pseudo-labeled*

- annotation. *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, 552–559 (2025).
38. Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization*. *International Conference on Learning Representations* (2019). URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
 39. He, C., Tan, T.-P., Xue, S. & Tan, Y. Simulating judicial trial logic: dual residual cross-attention learning for predicting legal judgment in long documents. *Expert Systems with Applications* **261**, 125462 (2025).
 40. Rahm, E. & Bernstein, P. A. A survey of approaches to automatic schema matching. *VLDB Journal* **10**, 334–350 (2001).
 41. Shvaiko, P. & Euzenat, J. Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering* **25**, 158–176 (2013).
 42. Wang, S. *et al.* *LLMATCH: a unified schema matching framework with large language models*. *Web and Big Data*, Vol. 16116 of *Lecture Notes in Computer Science*, 343–356 (Springer, 2026).
 43. Iovine, A., Huang, Y., Monteiro, M., Yakout, M. & Gokalp, S. *Effective product schema matching and duplicate detection with large language models*. *Proceedings of the LLM4ECommerce Workshop at the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2025). URL <https://www.amazon.science/publications/effective-product-schema-matching-and-duplicate-detection-with-large-language-models>. 9 pages.
 44. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ESLASupplementaryInformation.pdf](#)