Mitigating Entity Type Confusion in Cross-Domain NER via Multidimensional Quantification and Reasoning Enhancement

Jingyu Wang*, **Shijie Wu**, **Fusheng Jin***
Beijing Institute of Technology
jyw.bit.2003@gmail.com, wsj1595@bit.edu.cn, jfs21cn@bit.edu.cn

Abstract

Cross-domain Named Entity Recognition (CD-NER) aims to transfer the rich knowledge in the source domain to the target domain. Recent studies adopting decomposition or generation paradigms have achieved significant performance improvements, demonstrating high accuracy in entity span detection. However, during entity type classification, models severely suffer from entity type confusion, the erroneous tendency that models classify entities of one type in the text as another similar but incorrect type. To address this issue, we first propose a Multidimensional Confusion Quantification Model (MCQM) that quantifies a model's confusion extent between entity types from three dimensions: source-target hierarchy analysis, semantic similarity analysis, and explicit data evaluation. Moreover, we propose the Progressive Bidirectional Reasoning Chain (PBRC). PBRC leverages the source-target hierarchy and confusion analysis from the MCQM to prompt the LLM to generate two-stage reasoning information. The two-stage reasoning information is utilized to augment the knowledge of the model, significantly mitigating entity type confusion and improving the model's generalization performance. Experimental results demonstrate that our method achieves new state-of-the-art results on all domains of the CrossNER dataset.

1 Introduction

Named Entity Recognition (NER) is a core task in information extraction, which aims to identify specific entities in texts that belong to predefined types, such as person, location, and organization [Li *et al.*, 2022; Hu *et al.*, 2024; Esmaail *et al.*, 2024]. NER plays a critical role in various tasks, including information retrieval [Long *et al.*, 2024; Cong *et al.*, 2023; Nguyen *et al.*, 2024], knowledge graphs [Chen *et al.*, 2024; Jin *et al.*, 2022; Gao *et al.*, 2025], and recommendation [Li *et al.*, 2024; Li *et al.*, 2025]. With large-scale annotated data, neural network-based methods have performed remarkably in traditional NER [Zhu and Li, 2022; Shen *et al.*, 2023; Liu *et al.*, 2022]. However, in cross-domain NER, several challenges persist, especially in low-resource scenarios where model performance experiences a significant decline.

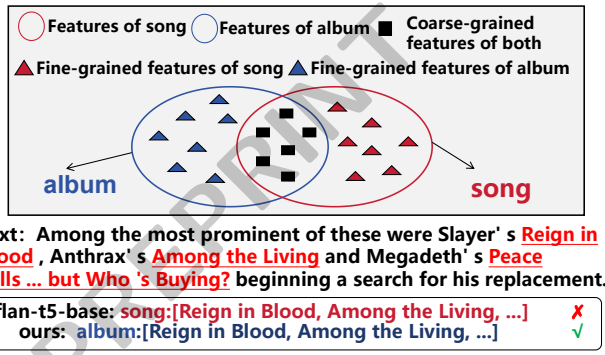

Figure 1: *song* and *album* share similar contextual features and entity forms, which can lead to confusion.

Recent studies [Xu and Cai, 2023; Zhang *et al.*, 2024a; Xu *et al.*, 2024] decompose cross-domain NER into subtasks to capture transferable patterns across domains. These methods focus on identifying shared feature patterns between domains through specialized components to strengthen source-domain knowledge transfer. Other approaches [Chen *et al.*, 2023; Zhang *et al.*, 2024b; Nandi and Agrawal, 2024] employ generative pre-trained language models, transforming NER into a generation task through the use of task-specific input prompts. These methods can effectively utilize large-scale knowledge from pretraining and exhibit high flexibility. Although both paradigms have demonstrated effectiveness in improving cross-domain NER performance, they remain susceptible to entity type confusion, leading to significant performance degradation in entity type classification.

We observe that cross-domain NER currently faces two major challenges. (1) Granularity differences between coarse-grained entity types in the source domain and fine-grained entity types in the target domain. Entity types in the source domain are coarse-grained, such as PER and LOC, whereas entity types in the target domain are more specific and domain-specialized, such as scientist and country. Compared to coarse-grained types, distinguishing fine-grained types often requires more detailed and nuanced features. Re-

*Corresponding authors.

lying solely on the coarse-grained features learned from the source domain makes it difficult for the model to accurately classify fine-grained types in the target domain, such as song and album (as shown in Figure 1). (2) The lack of contextual information in the text further limits the model's ability to classify fine-grained entity types accurately. In the target domain, some samples lack the necessary contextual cues, making it difficult for the model to effectively learn and capture the required fine-grained features for type classification. As shown in Figure 1, we selected a sample from CrossNER [Liu *et al.*, 2021], where the song and the album are two entity types that are difficult to distinguish. Entities of both types share similar forms and contextual environments. Due to insufficient information in the text, the model misclassifies these entities as the song. The lack of adequate contextual information prevents the model from effectively learning the feature differences between similar fine-grained entity types, which reduces the precision of type classification.

To address the challenges in cross-domain NER, we propose the Multidimensional Confusion Quantification Model (MCQM) and the Progressive Bidirectional Reasoning Chain (PBRC). (1) We propose the MCQM that quantifies the model's confusion extent from three dimensions: source-target hierarchy analysis, semantic similarity analysis, and explicit data evaluation. MCQM enables in-depth analysis and quantification of the model's confusion extent between fine-grained entity types in the target domain, providing a foundation for addressing the entity type confusion. (2) Furthermore, we innovatively leverage the source-target hierarchy, along with the confusion analysis from the MCQM, to propose the PBRC, which employs a two-stage reasoning strategy. Specifically, PBRC first instructs the LLM to perform initial reasoning based on contextual information using coarse-grained entity types from the source domain. Then, it instructs the LLM to perform bidirectional reasoning using fine-grained entity types from the target domain and confusion analysis from the MCQM. Finally, the LLM generates two-stage reasoning information for knowledge augmentation. The two-stage reasoning strengthens the fine-grained features and alleviates the granularity gap between the source and target domains, while external knowledge provided by the LLM solves the issue of insufficient contextual information. Our method significantly mitigates entity type confusion. Experimental results demonstrate that our method effectively improves the model's generalization ability, with the average F1 score increasing by more than 10.00%. In summary, our contributions are as follows:

- We systematically analyze the entity type confusion in cross-domain NER. Then, we propose the MCQM to quantify the model's confusion extent effectively.
- We propose the PBRC based on the source-target hierarchy and confusion analysis from the MCQM, which can leverage the LLM to mitigate entity type confusion in cross-domain NER significantly.
- Extensive experiments show that our method significantly improves the model's generalization ability, achieving new state-of-the-art performance across all domains of the CrossNER dataset.

2 Related Work

Named Entity Recognition (NER). Traditional methods rely on handcrafted features or rules combined with machine learning models like CRF [Lafferty *et al.*, 2001] for entity recognition, which are labor-intensive and lack adaptability to complex contexts. With the development of deep learning, neural network-based methods, such as LSTM [Huang *et al.*, 2015] and transformer [Vaswani *et al.*, 2017] models, have become mainstream. Models [Raffel *et al.*, 2020; Radford *et al.*, 2019; Devlin *et al.*, 2019] based on these architectures capture sequential relationships and contextual dependencies, offering improved performance and flexibility. However, existing methods still face challenges in cross-domain NER due to the variations across different domains.

Cross-Domain NER. Current studies can be broadly categorized into two paradigms: decomposition-based and generation-based. Decomposition-based methods [Xu and Cai, 2023; Xu *et al.*, 2024; Zhang *et al.*, 2024a; Zhang *et al.*, 2022] reformulate cross-domain NER into multiple subtasks. These methods improve knowledge transfer by learning shared cross-domain patterns through specialized subtask modules. Generation-based methods [Bao and Yang, 2024; Chen *et al.*, 2023; Zhang *et al.*, 2024b; Nandi and Agrawal, 2024] transform the NER into a generation task. These methods employ prompts containing task descriptions and auxiliary information, achieving robust performance through instruction fine-tuning. Although both paradigms effectively improve cross-domain performance, particularly in entity span detection, they still suffer from significant accuracy degradation in entity type classification.

LLMs for NER. With the emergence of LLMs, their reasoning capabilities have been leveraged to improve NER performance. Recent studies [Ashok and Lipton, 2023; Wang *et al.*, 2025; Zhu *et al.*, 2024; Ji, 2023] employ prompt learning for NER, where task-specific templates are augmented with demonstration examples and knowledge. LLMs are also used for cross-domain NER. Nandi and Agrawal [Nandi and Agrawal, 2024] perform instruction fine-tuning of LLMs by retrieving and leveraging similar examples. Zhang et al. [Zhang *et al.*, 2024b] utilize the LLM to generate task-oriented knowledge, used to conduct additional task-oriented pre-training of the backbone model for domain adaptation. These methods demonstrate strong few-shot performance.

Compared to fine-tuning LLMs [Nandi and Agrawal, 2024] and additional pre-training corpora [Zhang *et al.*, 2024b], our method reduces training time for domain adaptation.

3 Methodology

3.1 Task Description

Given a sentence $X = \{x_1, x_2, \ldots, x_N\}$, where w_i represents the i-th token in the sentence and N is the sentence length, the goal of NER is to identify all entities E within the sentence and classify each entity into a specific type. The set of entity types is denoted as L. Each entity $e_i \in E$ can be represented as $e_i = (y, x_{l:r})$, where $y \in L$ indicates the entity type, and l and r represent the start and end boundary indexes of the entity within the sentence, respectively. In

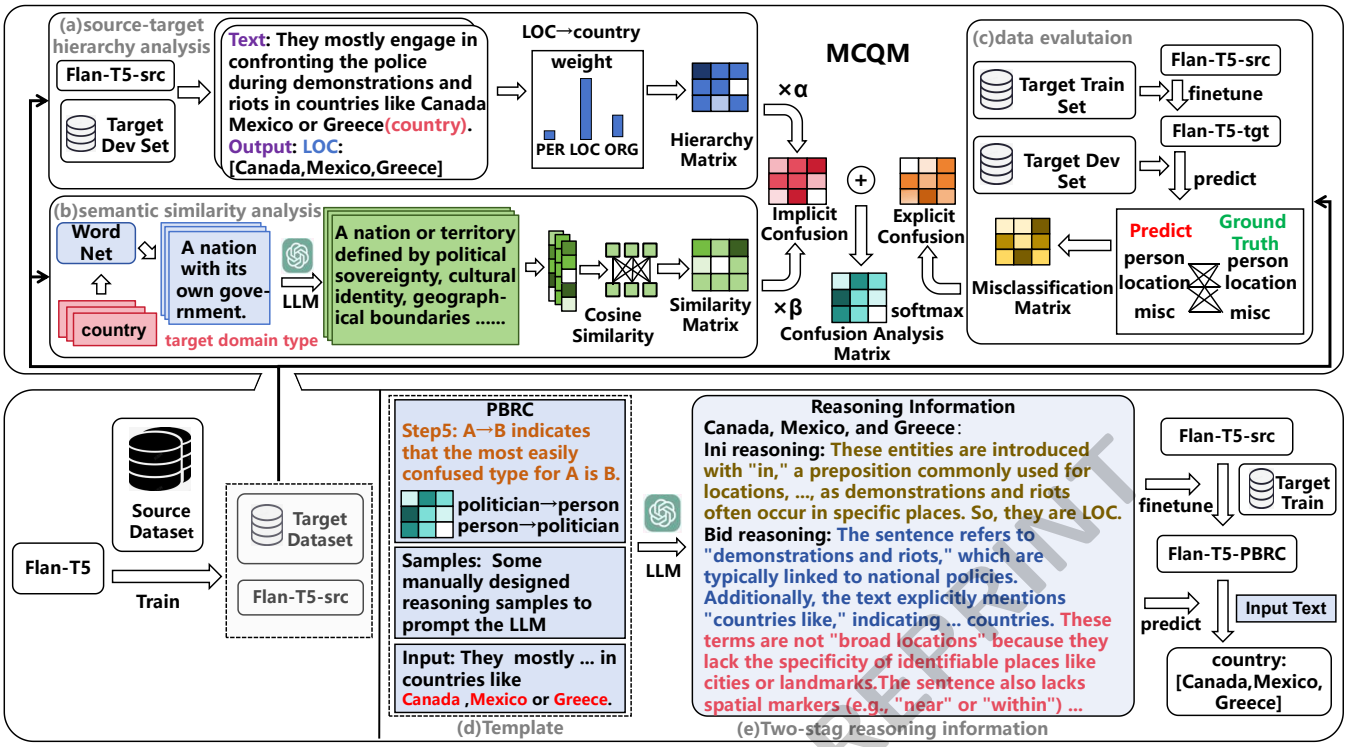

Figure 2: Overview of our proposed method, including MCQM and PBRC.

the cross-domain NER, two distinct datasets are considered: the source domain dataset $\mathcal{D}_{src}$ and the target domain dataset $\mathcal{D}_{tgt}$. The set of entity types in the source domain is denoted as $L_s = \{s_1, s_2, \ldots, s_n\}$, and the set of entity types in the target domain is denoted as $L_t = \{t_1, t_2, \ldots, t_m\}$. The goal is to leverage the knowledge learned from $\mathcal{D}_{src}$ to improve recognition performance on $\mathcal{D}_{tgt}$. Specifically, we focus on low-resource scenarios, where the training data in the target domain is significantly smaller than that in the source domain, i.e., $|\mathcal{D}_{tgt}| \ll |\mathcal{D}_{src}|$.

3.2 Multidimensional Confusion Quantification Model

Entities of the same type share certain common features, referred to as entity type features, and entity types with similar features are more prone to confusion. We quantify entity type confusion from three dimensions in Figure 2.

Source-Target Hierarchy Analysis

Entity types in the source domain are conceptually coarse-grained, whereas those in the target domain are fine-grained. In the feature space, the fine-grained entity types in the target domain can be viewed as extensions of the entity types in the source domain. For example, PER in the source domain can be expanded into fine-grained entity types in the target domain, such as politician and scientist. The expanded entity types retain the coarse-grained features in the source domain. Therefore, fine-grained entity types derived from the same entity type in the source domain exhibit similar features, meaning closer in the feature space and more prone to confusion among these entity types.

To quantify the model's confusion arising from the source-target hierarchy between the source and target domains, inspired by [Bao and Yang, 2024], we propose a reliable quantification method. For each entity in the target domain, its feature vector $\boldsymbol{\mathcal{H}}_e$ can be considered as a combination of coarse-grained features $\boldsymbol{H}_{coarse}$ and fine-grained features $\boldsymbol{H}_{fine}$.

We only focus on extracting $\boldsymbol{\mathcal{H}}_{coarse}$. We utilize the model $\mathcal{M}_{src}$, trained on $\mathcal{D}_{src}$, to extract $\boldsymbol{\mathcal{H}}_{coarse}$ of entities in the target domain and perform predictions based on the source domain type set L_s. We calculate the proportion of entities corresponding to each fine-grained type t_i in the target domain $\mathcal{D}_{tgt}$ that are classified as the coarse-grained type s_j in the source domain $\mathcal{D}_{src}$:

$$R_H\left(t_i \rightarrow s_j\right) = \frac{P\left(l_{tgt} = t_i \wedge l_{pred} = s_j\right)}{P\left(l_{tgt} = t_i\right)} \tag{1}$$

where $P(l_{tgt} = t_i)$ is the proportion of entities in $\mathcal{D}_{tgt}$ that belong to type t_i, $P(l_{tgt} = t_i \wedge l_{pred} = s_j)$ is the proportion of entities in $\mathcal{D}_{tgt}$ that belong to type t_i and are classified as s_j, and the proportion of entities of type t_i classified as s_j is denoted as $R_H(t_i \rightarrow s_j)$.

We use the source domain entity type l_{src} with the highest proportion as the prefix for the target domain entity type l_{tgt}, denoted as $l_{src} \rightarrow l_{tgt}$. The target domain entity type l_{tgt} is considered a fine-grained extension of the source domain entity type l_{src}:

$$prefix\left(t_i\right) = \arg\max_{s_j \in L_s}\left(R_H\left(t_i \rightarrow s_j\right)\right) \tag{2}$$

where $prefix(t_i)$ is the prefix of t_i, and L_s is the set of entity types in the source domain.

$R_H(t_i \rightarrow s_j)$ measures how much of the coarse-grained features of s_j are contained in t_i. The more entities corresponding to t_i are classified as s_j, the more coarse-grained features of s_j are contained in t_i. Therefore, for two fine-grained entity types, t_a and t_b, if they share the same prefix, it implies that t_a and t_b exhibit a significant overlap in their coarse-grained features, which originate from the same source domain entity type. This overlap increases the likelihood of confusion between t_a and t_b. We quantify the confusion arising from the source-target hierarchy using $R_H(t_i \rightarrow s_j)$, based on the prefixes of fine-grained entity types:

$$u_{hie}^{(t_i,t_j)} = \delta_{hie} \cdot R_H\left(t_i \rightarrow prefix\left(t_j\right)\right) \tag{3}$$

where $u_{hie}^{(t_i,t_j)}$ is the extent of hierarchical confusion from the fine-grained entity type t_i to t_j, and δ_{hie} is the scaling factor. This method quantifies the model's confusion arising from the source-target hierarchy, serving as an important metric in the MCQM for quantifying entity type confusion.

Semantic Similarity Analysis

The semantics of entity types reflect the common features of their entities. When the model maps the embedding vectors of the semantics of two entity types to closer positions in the semantic space, it considers the entity features of the two types as more similar, which increases the probability of confusing the two types in classification.

Since entity type labels are short words, they cannot sufficiently reflect the features of the entity types. To enrich the semantics of each target domain entity type $t_i \in L_t$, we utilize the WordNet [Fellbaum, 2010] to obtain a brief description C_s of t_i. Then, C_s is further refined into a final description C_e of t_i by leveraging the LLM. Then, C_e is fed into the encoder of the model, from which the hidden layer feature sequence $\boldsymbol{\mathcal{H}} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_Z] \in \mathbb{R}^{Z \times d}$ can be extracted:

$$\boldsymbol{\mathcal{H}} = \text{Encoder}\left(C_e\right) \tag{4}$$

where $\boldsymbol{h}_i$ denotes the feature vector from the final hidden layer for the i-th token, Z is the number of tokens, and d represents the dimension of the final hidden layer of the encoder. This study [Reimers and Gurevych, 2019] demonstrates that Mean Pooling outperforms both [CLS] token and Max Pooling strategies in semantic textual similarity tasks. Therefore, we use Mean Pooling to calculate the global feature vector $\overline{\boldsymbol{\mathcal{H}}}$, which better integrates the semantic information of all tokens and provides a more comprehensive representation of the entity type features:

$$\overline{\boldsymbol{\mathcal{H}}} = \frac{\sum_{i=1}^{Z} m_i \cdot \boldsymbol{h}_i}{\sum_{i=1}^{Z} m_i} \tag{5}$$

where m_i is the mask value of the i-th token, used to ignore the influence of padding.

In high-dimensional space, the features of entity types are amplified, and the mapping rules for similar features are also similar. Therefore, we calculate the semantic similarity between entity types using $\overline{\boldsymbol{\mathcal{H}}}$. We quantify the confusion arising from the semantic similarity by computing the cosine similarity between the global feature vectors $\overline{\boldsymbol{\mathcal{H}}}$ of entity types:

$$u_{\text{sim}}^{(t_i,t_j)} = \frac{\overline{\boldsymbol{\mathcal{H}}}_i \cdot \overline{\boldsymbol{\mathcal{H}}}_j}{\left\|\overline{\boldsymbol{\mathcal{H}}}_i\right\| \cdot \left\|\overline{\boldsymbol{\mathcal{H}}}_j\right\|} \tag{6}$$

where $u_{sim}^{(t_i,t_j)}$ is the extent of semantic confusion from the fine-grained entity type t_i to t_j. This method effectively quantifies the model's confusion arising from the semantic similarity between entity types, serving as an important metric in the MCQM for quantifying entity type confusion.

Fusion of Implicit Confusion Factors

The quantification of confusion arising from the source-target hierarchy and the semantic similarity of entity types is considered implicit confusion analysis. We use two fixed mixing ratios, α and β, to balance the weights of these two factors:

$$u_{iml}^{(t_i,t_j)} = \alpha \cdot u_{hie}^{(t_i,t_j)} + \beta \cdot u_{sim}^{(t_i,t_j)} \tag{7}$$

Then, we use the softmax function to model the implicit confusion and define the implicit confusion distribution among fine-grained entity types in the target domain:

$$v_{iml}(t_i \rightarrow t_j) = \frac{\exp(w_0 \cdot u_{iml}^{(t_i,t_j)})}{\sum\limits_{\substack{t_k \in L_t \\ t_k \neq t_i}} \exp(w_0 \cdot u_{iml}^{(t_i,t_k)})} \tag{8}$$

where $w_0 \in \mathbb{R}^+$ is a temperature coefficient.

Explicit Data Evaluation

We denote the model obtained by fine-tuning $\mathcal{M}_{src}$ on the target domain dataset as $\mathcal{M}_{tgt}$. We use the dev set of the target domain to perform a statistical analysis of the misclassification proportions for all entity types. Evaluating $\mathcal{M}_{tgt}$ reflects the model's ability to distinguish between different fine-grained entity types and explicitly reveals the model's confusion extent between entity types:

$$R_M^{(t_i,t_j)} = \frac{P(l_{true} = t_i \wedge l_{pred} = t_j)}{P(l_{true} = t_i)} \tag{9}$$

where $P(l_{true} = t_i)$ is the proportion of entities that truly belong to type t_i, $P(l_{true} = t_i \wedge l_{pred} = t_j)$ is the proportion of entities that truly belong to type t_i but are misclassified as t_j, and the proportion of entities of type t_i misclassified as t_j is denoted as $R_M^{(t_i,t_j)}$.

The value of $R_M^{(t_i,t_j)}$ explicitly reflects the model's learning effectiveness in distinguishing the fine-grained features of t_i and t_j. A larger value indicates that t_i and t_j are more prone to confusion. We use $R_M^{(t_i,t_j)}$ as an important metric for confusion analysis in the MCQM. Then, we use the softmax function to model explicit confusion and define the explicit confusion distribution among fine-grained entity types in the target domain:

$$v_{exl}(t_i \rightarrow t_j) = \frac{\exp(w_0 \cdot R_M^{(t_i,t_j)})}{\sum\limits_{\substack{t_k \in L_t \\ t_k \neq t_i}} \exp(w_0 \cdot R_M^{(t_i,t_k)})} \tag{10}$$

where $w_0 \in \mathbb{R}^+$ is a temperature coefficient.

Confusion Analysis Results

We integrate the implicit and explicit evaluations to obtain the final confusion analysis results for fine-grained entity types in the target domain:

$$v_{conf}(t_i \rightarrow t_j) = v_{iml}(t_i \rightarrow t_j) + v_{exl}(t_i \rightarrow t_j) \tag{11}$$

Step1: Identify all potential named entities from the text.

Step2: For each named entity, analyze the contextual information and classify it using the following coarse-grained types.
1.PER: Refers to individuals, including their full names, nicknames, titles, or any personal identifiers.
2.LOC: Refers to geographical locations such as cities, countries, landmarks, or any specific places.

Step3: For each named entity, based on external knowledge and contextual information of the entity, explains why the entity is classified as the corresponding coarse-grained type.

Step4: For each named entity, based on its coarse-grained type, further classify it using following fine-grained entity types.
1.PER→scientist: Refers to an individual engaged in scientific research or experimentation, often working in fields to advance knowledge.....
2.LOC→country: Refers to a geographic and political entity with sovereignty, defined borders, and a governing body.

Step5: Below is the table confusion mapping table, A→B indicates that the most easily confused type for A is B. For each named entity, based on external knowledge and contextual information of the entity, explain why the entity is classified as type A rather than type B.
1.person→scientist
2.chemicalelement→chemicalpound
3.enzyme→protein

Last, You only need to output the analysis for Step3 and Step5, without including the final classification results.

Figure 3: Contents of PBRC.

where $v_{conf}(t_i \rightarrow t_j)$ is the model's confusion extent from t_i to t_j. For each type t_i, we select the entity type with the highest confusion extent as its most easily confused type:

$$T_{conf}^{(t_i)} = \underset{t_j \in L_t, t_j \neq t_i}{\arg\max} \ (v_{conf}(t_i \rightarrow t_j)) \tag{12}$$

where $T_{conf}^{(t_i)}$ is the most easily confused entity type for the entity type t_i. MCQM provides a comprehensive analysis of the model's confusion extent between fine-grained entity types, playing a critical role in the bidirectional reasoning of PBRC.

3.3 Progressive Bidirectional Reasoning Chain

Contents of PBRC

We propose PBRC based on the source-target hierarchy and the confusion analysis from MCQM. As shown in Figure 2, PBRC prompts the LLM to output two-stage reasoning information. Figure 3 shows the contents of PBRC.

Step1: Identifying Named Entities. This step instructs the LLM to identify all potential named entities in the text. With pre-trained multi-domain knowledge and contextual understanding, the LLM is capable of recognizing potential entities in texts across various domains.

Step2 and Step3: Coarse-grained Classification and Initial Reasoning. We provide all source domain coarse-grained entity types and explanations. For each entity in Step1, we instruct the LLM to conduct coarse-grained classification and provide explanations, leveraging both its prior knowledge and an initial analysis of the context. This step unifies entity types across different domains, serving as an initial classification and reasoning process for entities.

Step4 and Step5: Fine-grained Classification and Bidirectional Reasoning. We provide all fine-grained entity types of the target domain along with their corresponding source domain prefixes and explanations. For each entity, we instruct the LLM to conduct fine-grained classification based on the initial classification and reasoning information from Step2 and Step3, and then conduct bidirectional reasoning through a deep analysis of contextual information and the LLM's rich external knowledge. As shown in Figure 3, we present the confusion analysis result of the MCQM, where $A \rightarrow B$ indicates that A's most easily confused type is B. Bidirectional reasoning consists of two parts: (1) **forward reasoning**, which infers that the entity type is A; (2) **backward reasoning**, which infers that the entity type is not B.

Two-stage Reasoning Information. The LLM only outputs the initial reasoning from Step3 and the bidirectional reasoning from Step5. The two-stage reasoning information is fed into the backbone for knowledge augmentation: the initial reasoning guides the backbone to leverage the knowledge from the source domain, strengthening knowledge transfer; the bidirectional reasoning guides the backbone to differentiate easily confused entity types, mitigating the entity type confusion. Besides, external knowledge mitigates the insufficiency of contextual information.

Knowledge Augmentation

The two-stage reasoning information generated by the LLM is utilized to augment the knowledge of $\mathcal{M}_{src}$. Through supervised fine-tuning, the backbone learns to leverage this reasoning information, yielding the final backbone model $\mathcal{M}_{tgt}^{\dagger}$.

- **Input:** Find all entities of types {politician, person, country ...} in {text}.
  Reasoning information: {...}.
  Output format: {"type 1": ["entity 1", "entity 2"], "type 2": ["entity 3"], ...}.
- **Gold Sequence:** {"country": ["Afghanistan"], "politician": ["Barack Obama"], ...}.

Domain	Dataset	Type	Train	Dev	Test
Source	CoNLL2003	4	14987	-	-
	Twitter	4	4290	-	-
Target	politics	9	200	541	651
	science	17	200	450	543
	music	13	100	380	465
	literature	12	100	400	416
	AI	14	100	350	431

Table 1: The statistics of all datasets.

4 Experiments

4.1 Experimental Setup

As shown in Figure 1, we evaluate on two source domains (CoNLL2003 [Sang and De Meulder, 2003] and Twitter [Lu *et al.*, 2018]) and five target domains (CrossNER [Liu *et al.*, 2021]: politics, science, music, literature, and AI). We employ Flan-T5 [Chung *et al.*, 2024] as the backbone model and GPT-4o-mini as the LLM. To evaluate the performance of our proposed method, we compare it with the following baselines: (1) **MTD** [Zhang *et al.*, 2022]: A modular learning-based method that decomposes NER into span detection and

Method	CoNLL2003						Twitter					
	sci.	pol.	mus.	lit.	AI	Avg.	sci.	pol.	mus.	lit.	AI	Avg.
MTD	72.35	76.70	76.10	69.22	68.93	72.66	71.37	74.62	74.41	69.67	64.55	70.92
CP-NER	75.82	74.25	79.10	72.17	67.95	73.86	-	-	-	-	-	-
DH-GAT	74.21	77.06	78.77	72.51	69.30	74.37	74.55	76.46	77.33	71.52	67.65	73.50
PromptNER	72.59	78.61	84.26	74.44	64.83	74.95	-	-	-	-	-	-
Dual-CL	74.17	77.56	78.57	72.43	69.49	74.44	74.07	77.53	77.21	71.72	68.71	73.85
DT-MPrompt	73.06	80.54	79.54	73.51	70.13	75.36	73.18	79.86	77.93	72.74	69.13	74.57
IF-WRANER-7B	75.31	79.80	85.43	75.52	68.81	76.97	-	-	-	-	-	-
TOPT	80.16	81.55	82.03	77.85	72.34	78.78	-	-	-	-	-	-
Flan-T5-base + MCQM & PBRC	81.43	83.53	83.69	79.03	74.29	80.39	80.32	82.43	81.93	77.62	73.97	79.25
Flan-T5-large + MCQM & PBRC	**82.34**	**85.03**	**85.62**	**79.96**	**75.99**	**81.79**	**81.62**	**84.64**	**84.02**	**78.61**	**75.50**	**80.88**

Table 2: F1 scores on CrossNER: CoNLL2003 and Twitter as source domains, respectively. Bold marks the highest.

Domain	Flan-T5-base (w/o MCQM & PBRC)			Flan-T5-base (+ MCQM & PBRC)		
	Precision	Recall	F1 score	Precision	Recall	F1 score
science	71.31 (83.11)	68.66 (81.10)	69.96 (82.09)	83.05 (88.12)	79.88 (84.49)	81.43 (86.27)
politics	71.30 (87.45)	69.29 (86.99)	70.28 (87.22)	84.62 (91.15)	82.47 (88.01)	83.53 (89.55)
music	77.07 (88.46)	73.65 (82.99)	75.32 (85.64)	85.34 (89.45)	82.10 (87.85)	83.69 (88.64)
literature	68.28 (83.67)	67.47 (83.98)	67.87 (83.82)	80.48 (87.67)	77.64 (84.12)	79.03 (85.86)
AI	63.42 (81.15)	63.70 (81.64)	63.56 (81.39)	75.74 (86.07)	72.90 (83.01)	74.29 (84.51)

Table 3: Entity Type Confusion Analysis. CoNLL2003 as the source domain.

type classification. (2) **CP-NER** [Chen *et al.*, 2023]: Transforms NER as text-to-text task and introduces collaborative domain-prefix tuning based T5 as well. (3) **DH-GAT** [Xu and Cai, 2023]: Applies Graph Attention Networks to encode syntactic and semantic information while embedding words into hyperbolic space. (4) **PromptNER** [Ashok and Lipton, 2023]: Uses GPT4 for NER through prompt templates. (5) **Dual-CL** [Xu *et al.*, 2024]: Uses dual contrastive learning to refine ambiguous representations and learn generalizable features. (6) **DT-MPrompt** [Zhang *et al.*, 2024a]: Splits the cross-domain NER task into subtasks and uses separate functional modules for learning and knowledge transfer. (7) **IF-WRANER** [Nandi and Agrawal, 2024]: Fine-tunes 7B LLaMA with instruction fine-tuning and employs word embeddings to retrieve examples for in-context learning. (8) **TOPT (SOTA)** [Zhang *et al.*, 2024b]: Utilizes LLaMA to generate task-oriented knowledge for flan-T5 and adopts task-oriented pre-training for domain adaptation.

4.2 Main Results

The main results are shown in Table 2.

CoNLL2003 as the Source Domain. It is observed that the F1 score improves as the parameter scale of the model increases. On average, our method achieves consistent F1 score improvements of +1.61% (Flan-T5-base) and +3.01% (Flan-T5-large). Our method surpasses prior SOTA results in science, politics, literature, and AI domains (effective for both Flan-T5-base and Flan-T5-large), achieving improvements exceeding 2.10% in these domains. These results demonstrate the strong effectiveness of two-stage information generated by the LLM in cross-domain NER. We observe that **IF-WARNER-7B** (85.43%), which employs a fine-tuned 7B LLaMA, and **PromptNER** (84.26%), which utilizes GPT-4, both achieve significantly higher F1 scores in the music domain compared to other baselines using smaller models. This demonstrates the advantage of LLMs in the music domain, ensuring the quality of the information generated by our two-stage process using the LLM. We observe that Flan-T5-large achieves only a marginal improvement of 0.19% over prior SOTA in the music domain, while Flan-T5-base fails to surpass the prior SOTA. Through careful analysis, we have identified the main reasons. First, the high overlap between the domain knowledge of the LLM and Flan-T5 in the music domain diminished their complementary effects, resulting in limited benefits from two-stage reasoning. Then, though the two-stage information significantly improves F1 scores for certain entity types, their limited entity count results in a negligible impact on the overall F1 metric.

Twitter as the Source Domain. Our method consistently achieves significant improvements across all domains. Since **TOPT** is not tested with Twitter as the source domain, we cannot compare with it directly. Compared to CoNLL2003, the performance of both Flan-T5-base and Flan-T5-large shows a slight decline. We attribute this to the smaller dataset size of Twitter, which limits the model's ability to learn coarse-grained features of entities. This highlights the importance of learning the coarse-grained features of entities and further demonstrates that our method enables effective knowledge transfer from the source to the target domain.

4.3 Entity Type Confusion Analysis

To verify that the model's performance improvement is attributed to enhanced entity type classification capabilities, we evaluate its performance in both entity span detection and overall performance using precision, recall, and F1 score. The results are shown in Table 3, where the values in parentheses indicate the performance on entity span detection. Before incorporating MCQM and PBRC, we observe a substan-

Setting	sci.	pol.	mus.	lit.	AI
w/o	78.06	79.47	81.44	76.12	71.42
+ SS	79.01	82.62	82.54	77.34	73.05
+ STH	79.98	82.67	82.09	77.86	73.59
+ SS&STH	80.93	83.05	83.16	78.41	73.66
+ DE	79.47	82.89	82.41	77.49	72.18
+ ALL	**81.43**	**83.53**	**83.69**	**79.03**	**74.29**

Table 4: Ablation Study on MCQM (F1 scores).

Setting	sci.	pol.	mus.	lit.	AI	Avg.
w/o	69.96	70.28	75.32	67.87	63.56	69.40
+ FR	76.58	78.26	80.09	74.77	70.03	75.95
+ IR&FR	78.06	79.47	81.44	76.12	71.42	77.30
+ BR&FR	80.62	82.76	82.49	78.06	73.52	79.49
+ ALL	**81.43**	**83.53**	**83.69**	**79.03**	**74.29**	**80.39**

Table 5: Ablation Study on PBRC (F1 scores).

tial performance gap between the entity span detection and overall performance, which indicates that the model's ability of entity type classification is the primary bottleneck, suffering from significant entity type confusion. After incorporating MCQM and PBRC, we observe that our method improves the F1 score for entity span detection by 2%–4% across the five domains, with slight improvements also observed in precision and recall. More notably, the overall performance increases by over 10% on average. These results demonstrate that our method significantly enhances the model's ability in entity type classification and effectively mitigates the model's entity type confusion.

4.4 Ablation Study

Ablation Study on MCQM

Table 4 presents the effects of MCQM components: semantic similarity (SS), source–target hierarchy (STH), and data evaluation (DE), on model performance (F1 score). By comparing the results of completely removing all components (w/o) with those of introducing only SS, STH, or DE, we observe that, across all target domains, incorporating SS, STH, or DE into confusion analysis and subsequently introducing backward reasoning based on the corresponding analysis improves model performance. This indicates that semantic similarity analysis, source–target hierarchy analysis, and explicit data evaluation can all effectively quantify the entity type confusion extent. Moreover, simultaneously introducing SS and STH (+SS&STH) or incorporating all components (+ALL) into confusion analysis further improves model performance, suggesting that leveraging multiple analytical dimensions can reduce the errors associated with a single dimension and enhance the quantification accuracy of MCQM.

Ablation Study on PBRC

We conduct ablation studies to evaluate the effectiveness of each reasoning component in PBRC: forward (FR), backward (BR), and initial reasoning (IR). As shown in Table 5, adding forward reasoning (FR) alone yields an average performance gain of 6.55%. This demonstrates that the reasoning information provided by the LLM can effectively support the backbone model in performing NER. Further incorporating initial reasoning (IR) brings an additional average improvement of 1.35%, demonstrating that the two-stage reasoning from source to target domain can enhance the backbone's cross-domain transferability. Moreover, when backward reasoning (BR) is incorporated, the average performance rises by 3.54% compared to using FR alone. This indicates that the MCQM effectively analyzes and quantifies the model's confusion among entity types, while backward reasoning can alleviate entity type confusion.

Model	Training	Inference	Dev F1
GPT-4o-mini	-	3.22s	74.78
Flan-T5-base	28.62 min	0.32s	70.25
GPT-4o-mini + base	23.60 min	3.56s	80.69
Flan-T5-large	46.28 min	0.51s	71.98
GPT-4o-mini + large	37.35 min	3.84s	82.55

Table 6: The runtime statistics of different methods.

4.5 Runtime Analysis

We report the average training time of different methods across the five target domains, the average F1 score on the dev set, and the inference time per text (in seconds), as shown in Table 6. First, although our method requires the LLM to generate reasoning information for the training set in advance, the overall training time is still lower than that of using Flan-T5 alone. We attribute this to the fact that, once reasoning information is introduced, Flan-T5 converges faster and reaches the best F1 score of the dev set with fewer epochs. Besides, for the inference time of each text, our method first leverages the LLM to generate reasoning information before feeding it into Flan-T5 for NER. Since Flan-T5 has a very short inference time, the final inference time of our method is comparable to that of the LLM. Our method maintains training time comparable to that of using Flan-T5 alone, attains inference efficiency that is nearly equivalent to the LLM, and achieves performance that substantially exceeds both Flan-T5 and the LLM when employed alone.

5 Conclusion

In this paper, we propose the MCQM, which effectively quantifies the model's confusion extent among fine-grained entity types from three dimensions: source-target hierarchy analysis, semantic similarity analysis, and explicit data evaluation. Furthermore, we propose the PBRC based on the source-target hierarchy and the MCQM, which can significantly mitigate entity type confusion and improve the model's generalization ability. Our method achieves SOTA results on all domains of the CrossNER dataset.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No. 62272045).

References

[Ashok and Lipton, 2023] Dhananjay Ashok and Zachary C Lipton. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*, 2023.

[Bao and Yang, 2024] Ke Bao and Chonghuan Yang. Label alignment and reassignment with generalist large language model for enhanced cross-domain named entity recognition. *arXiv preprint arXiv:2407.17344*, 2024.

[Chen *et al.*, 2023] Xiang Chen, Lei Li, Shuofei Qiao, Ningyu Zhang, Chuanqi Tan, Yong Jiang, Fei Huang, and Huajun Chen. One model for all domains: collaborative domain-prefix tuning for cross-domain ner. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

[Chen *et al.*, 2024] Zhongwu Chen, Long Bai, Zixuan Li, Zhen Huang, Xiaolong Jin, and Yong Dou. A new pipeline for knowledge graph reasoning enhanced by large language models without fine-tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1366–1381, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Chung *et al.*, 2024] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, and et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[Cong *et al.*, 2023] Xin Cong, Bowen Yu, Mengcheng Fang, Tingwen Liu, Haiyang Yu, Zhongkai Hu, Fei Huang, Yongbin Li, and Bin Wang. Universal information extraction with meta-pretrained self-retrieval. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4084–4100, Toronto, Canada, July 2023.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Esmaail *et al.*, 2024] Naji Esmaail, Nazlia Omar, Masnizah Mohd, Fariza Fauzi, and Zainab Mansur. Named entity recognition in user-generated text: A systematic literature review. *IEEE Access*, 12:136330–136353, 2024.

[Fellbaum, 2010] Christiane Fellbaum. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer, 2010.

[Gao *et al.*, 2025] Hepeng Gao, Funing Yang, Yongjian Yang, and Ying Wang. Hierarchy knowledge graph for parameter-efficient entity embedding. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2811–2819, 2025.

[Hu *et al.*, 2024] Zhentao Hu, Wei Hou, and Xianxing Liu. Deep learning for named entity recognition: a survey. *Neural Comput. Appl.*, 36(16):8995–9022, March 2024.

[Huang *et al.*, 2015] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[Ji, 2023] Bin Ji. Vicunaner: Zero/few-shot named entity recognition using vicuna. *arXiv preprint arXiv:2305.03253*, 2023.

[Jin *et al.*, 2022] Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. A good neighbor, a found treasure: Mining treasured neighbors for knowledge graph entity typing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 480–490, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[Lafferty *et al.*, 2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[Li *et al.*, 2022] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70, 2022.

[Li *et al.*, 2024] Xinhang Li, Xiangyu Zhao, Yejing Wang, Yu Liu, Chong Chen, Cheng Long, Yong Zhang, and Chunxiao Xing. Opensiterec: An open dataset for site recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1483–1493, New York, NY, USA, 2024.

[Li *et al.*, 2025] Haodong Li, Lianyong Qi, Weiming Liu, Xiaolong Xu, Wanchun Dou, Yang Cao, Xuyun Zhang, Amin Beheshti, and Xiaokang Zhou. Balancing user-item structure and interaction with large language models and optimal transport for multimedia recommendation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 3027–3035, 2025.

[Liu *et al.*, 2021] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13452–13460, 2021.

[Liu *et al.*, 2022] Jian Liu, Yufeng Chen, and Jinan Xu. Low-resource ner by data augmentation with prompting. In *IJCAI*, pages 4252–4258, 2022.

[Long *et al.*, 2024] Zijun Long, Xuri Ge, Richard McCreadie, and Joemon M. Jose. Cfir: Fast and effective long-text to image retrieval for large corpora. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2188–2198, New York, NY, USA, 2024. Association for Computing Machinery.

[Lu *et al.*, 2018] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1990–1999, 2018.

[Nandi and Agrawal, 2024] Subhadip Nandi and Neeraj Agrawal. Improving few-shot cross-domain named entity recognition by instruction tuning a word-embedding based retrieval augmented large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 686–696, 2024.

[Nguyen *et al.*, 2024] Thong Nguyen, Shubham Chatterjee, Sean MacAvaney, Iain Mackie, Jeff Dalton, and Andrew Yates. DyVo: Dynamic vocabularies for learned sparse retrieval with entities. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 767–783, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[Sang and De Meulder, 2003] Erik Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 142–147, 2003.

[Shen *et al.*, 2023] Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. Promptner: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12492–12507, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2025] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. Gpt-ner: Named entity recognition via large language models. In *Findings of the association for computational linguistics: NAACL 2025*, pages 4257–4275, 2025.

[Xu and Cai, 2023] Jingyun Xu and Yi Cai. Decoupled hyperbolic graph attention network for cross-domain named entity recognition. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, New York, NY, USA, 2023. Association for Computing Machinery.

[Xu *et al.*, 2024] Jingyun Xu, Junnan Yu, Yi Cai, and Tat-Seng Chua. Dual contrastive learning for cross-domain named entity recognition. *ACM Trans. Inf. Syst.*, 42(6), October 2024.

[Zhang *et al.*, 2022] Xinghua Zhang, Bowen Yu, Yubin Wang, Tingwen Liu, Taoyu Su, and Hongbo Xu. Exploring modular task decomposition in cross-domain named entity recognition. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, New York, NY, USA, 2022.

[Zhang *et al.*, 2024a] Xinghua Zhang, Bowen Yu, Xin Cong, Taoyu Su, Quangang Li, Tingwen Liu, and Hongbo Xu. Cross-domain ner under a divide-and-transfer paradigm. *ACM Transactions on Information Systems*, 42(5):1–32, 2024.

[Zhang *et al.*, 2024b] Zhihao Zhang, Sophia Yat Mei Lee, Junshuang Wu, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. Cross-domain ner with generated task-oriented knowledge: an empirical study from information density perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1595–1609, 2024.

[Zhu and Li, 2022] Enwei Zhu and Jinpeng Li. Boundary smoothing for named entity recognition. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Zhu *et al.*, 2024] Xingyu Zhu, Feifei Dai, Xiaoyan Gu, Bo Li, Meiou Zhang, and Weiping Wang. Gl-ner: Generation-aware large language models fornbsp;few-shot named entity recognition. In *Artificial Neural Networks and Machine Learning – ICANN 2024.*, 2024.