Zero-shot Cross-lingual NER via Mitigating Language Difference: An Entity-aligned Translation Perspective

Zhihao Zhang[1], **Sophia Yat Mei Lee**[2], **Dong Zhang**[1*], **Shoushan Li**[1] **and Guodong Zhou**[1]
[1]School of Computer Science & Technology, NLP Lab, Soochow University, China
[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
dzhang@suda.edu.cn

Abstract

Cross-lingual Named Entity Recognition (CL-NER) aims to transfer knowledge from high-resource languages to low-resource languages. However, existing zero-shot CL-NER (ZCL-NER) approaches primarily focus on Latin script language (LSL), where shared linguistic features facilitate effective knowledge transfer. In contrast, for non-Latin script language (NSL), such as Chinese and Japanese, performance often degrades due to deep structural differences. To address these challenges, we propose an entity-aligned translation (EAT) approach [†]. Leveraging large language models (LLMs), EAT employs a dual-translation strategy to align entities between NSL and English. In addition, we fine-tune LLMs using multilingual Wikipedia data to enhance the entity alignment from source to target languages. Extensive experiments demonstrate that EAT outperforms prior methods on NSL by bridging language gaps through entity-aware translation.

1 Introduction

Cross-lingual Named Entity Recognition (CL-NER) aims to transfer knowledge from high-resource source languages to low-resource target languages, so as to enhance the NER performance (Li et al., 2021; Mhaske et al., 2023; Xie et al., 2023, 2024). Recently, while zero-shot CL-NER (ZCL-NER) approaches demonstrate strong performance on several low-resource languages (Zeng et al., 2022; Ge et al., 2023, 2024), we observe an interesting phenomenon in prior approaches and different language scripts.

Previous ZCL-NER approaches typically apply a teacher-student (T-S) learning framework, transferring English knowledge to the target language (Ma et al., 2022; Zeng et al., 2022). This recasts the **source language into the same space as the target language** to achieve ZCL-NER. Regarding the existing studies (Ge et al., 2023, 2024), they achieve competitive performance on the English-like target languages that have shared vocabulary origins, as well as similar grammatical and syntactic structures such as German and French (Finiasz et al., 2024). For example, as illustrated in Figure 1, the German words "Britische" and "fotografieren" are derived from the English words "British" and "photograph", respectively. We refer to such languages, which are closely related to English, as Latin script language (**LSL**). However, as shown in our pilot experiments and previous reports (Ge et al., 2023, 2024), T-S approaches work not well on non-Latin script language (**NSL**), such as Chinese and Japanese. This is mainly due to significant linguistic discrepancies between LSL and NSL, i.e., differences in scripts, grammar, and syntax (Singh et al., 2022), which are summarized in Table 1 for various languages. For example, German (DE), as LSL, shares the same Subject-Verb-Object (SVO) and fusional scripts characteristics with English

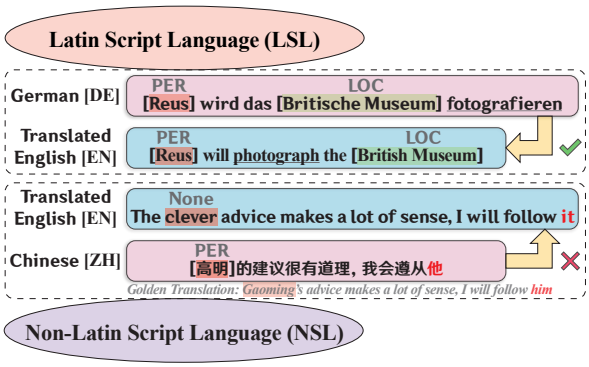

Figure 1: Two examples: German, as LSL, tends to be translated more accurately into English due to their shared lexicon, making it more suitable for NER. In contrast, Chinese, as NSL, faces inherent challenges in translation to English because of significant typological divergences. The translations are obtained by GPT-4.

*Corresponding Author
[†]Our code and dataset are available at: https://github.com/ZelateCalcite/EAT_NER

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4541–4557
November 4-9, 2025

	AR	HI	HY	JA	ES FR	EN
Language Family	Afroasiatic · Semitic	Indo-European · Indo-Iranian	Indo-European · Armenian	Japonic · Japanese	Indo-European · Romance	Indo-European · Germanic
Linguistic Type	Fusional	Fusional	Fusional	Agglutinative	Fusional	Analytic
Scripts	Arabic Abjad	Devanagari	Armenian script	Kanji & Kana	Latin Scripts[i]	Latin Scripts[i]
Word Order	VSO[ii]	SOV	SOV	SOV	SVO	SVO
	KA	**KO**	**RU**	**ZH**	**DE NL**	**EN**
Language Family	Kartvelian · Karto-Zan	Koreanic · Korean	Indo-European · Slavic	Sino-Tibetan · Sinitic	Indo-European · Germanic	Indo-European · Germanic
Linguistic Type	Agglutinative	Agglutinative	Fusional	Isolating	Fusional	Analytic
Scripts	Georgian Scripts	Hangul / Chosŏn'gŭl	Cyrillic	Chinese Characters	Latin Scripts[i]	Latin Scripts[i]
Word Order	SVO	SOV	SVO	SVO	SVO[iii]	SVO

Table 1: Comparison of linguistic typology between different languages using different scripts (Language Code follows ISO 639-1:2002 *). SVO (Subject-Verb-Object, similar with SOV / VSO) refers to the order in which the elements of a sentence typically appear in languages that follow this structure. [i]: These languages are both based on Latin Scripts with unique pronunciation differences or additional characters. [ii]: The word order of Modern Written Arabic is VSO while Modern Spoken Arabic is SVO, here we only discuss the writing systems. [iii]: The usual word order of these two languages is SVO, but in subordinate clauses the word order shifts to SOV.

(EN), while Japanese (JA), as NSL, is an agglutinative language with Subject-Object-Verb (SOV) order. Consequently, translating Japanese into English poses greater challenges compared to German due to these typological divergences.

To this end, we believe that it is essential to mitigate the language gap between NSL and English. In this way, we can utilize the abundant English resources to improve ZCL-NER performance on targeted NSL, which constitutes the core focus of this paper. As we know, translation appears to be the most intuitive approach to bridging the linguistic disparity (Li et al., 2024). However, direct translation between English and NSL typically results in the key entity omissions due to various word misalignment issues (Yang et al., 2022). For instance, in Figure 1, the Person entity "高明" in target language (Chinese) is incorrectly translated as the adjective "clever" (by GPT-4o and Deepseek in May 2025 with evidences in Appendix). Such translation inconsistencies hinder the progress of our task.

To address the above issues, we propose an Entity-Aligned Translation (EAT) approach at dual levels with large language models (LLMs) for ZCL-NER, with a focus on NSL as the target languages. Different from T-S approaches, we recast **the target NSL into the same space with English** to achieve ZCL-NER. Specifically, we first leverage the powerful reasoning and interpretation abilities of LLMs to perform target-to-English forward translation using multi-round chain-of-thought (MrCoT). Then, we extract potential entities from the translated English text using pre-trained English NER extractor. Finally, to restore these entities in the target language, we design a backward translation process with MrCoT, ensuring that this stage's translated entities correspond to the correct fragments in the original sentence. The process allows us to achieve ZCL-NER without relying on parallel cross-lingual corpora. To further refine entity alignment, we fine-tune LLMs on English-oriented entity-aligned cross-lingual corpora (EACL), sourced from multilingual Wikipedia. In general, our contributions are summarized:

• We identify the linguistic script discrepancies and entity misalignment between LSL and NSL for ZCL-NER.

• We propose an entity-aligned translation (EAT) framework from the perspective of enhancing entity-aware translation ability for various NSL as the target languages.

• We introduce two metrics (BLEU and Shannon Entropy) to quantify the correlation between translation quality and ZCL-NER performance.

2 Related Work

2.1 Linguistic Differences Between Scripts

Previous research has explored how linguistic and cognitive processes are influenced by different writing systems, such as Latin scripts, Hanzi, and Devanāgarī. Gelman and Tardif (1998) argue that formal language properties (e.g., scripts, morphemes) impact the use of generic noun phrases, which are essential for knowledge organization and reasoning (Gelman and Markman, 1986; Finiasz et al., 2024). The Script Relativity Hypothesis (Pae,

*https://en.m.wikipedia.org/wiki/ISO_639-1

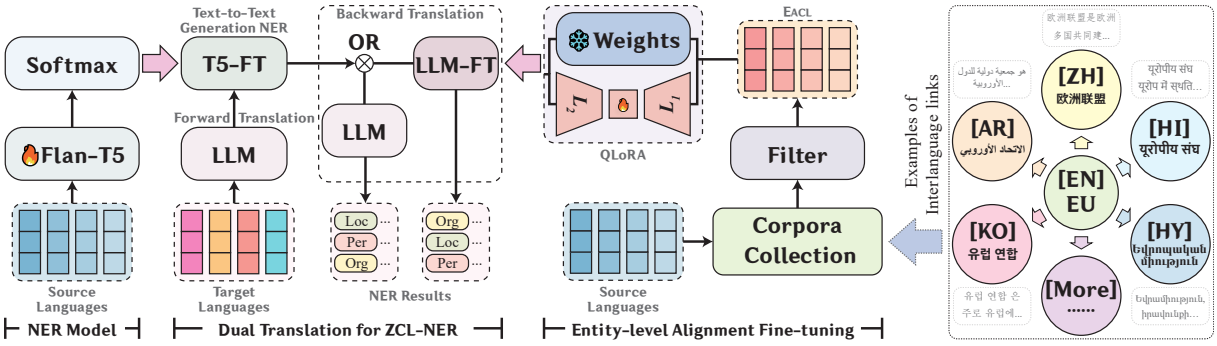

Figure 2: The overall architecture of our proposed EAT approach.

2020) suggests that script influences cognitive processes, with first-language experiences shaping how we process other languages (Li and Koda, 2022; Singh et al., 2022). Systemic functional linguistics (Halliday and Matthiessen, 1999) also highlights how information is conveyed differently across languages due to their unique semantic and syntactic structures (Yang, 2008; Chen, 2016; Arús-Hita et al., 2018). In contrast to previous studies, which generalize NER models without considering linguistic typology, our approach leverages LLMs to bridge these gaps, yielding improved performance.

2.2 Zero-shot Cross-Lingual NER

Recent zero-shot CL-NER approaches, particularly those using teacher-student (T-S) learning frameworks, have achieved promising results by distilling NER knowledge from source languages to target languages (Wu et al., 2020; Liang et al., 2021; Chen et al., 2021; Zeng et al., 2022). Some works focus on improving knowledge distillation by reducing noise (Ma et al., 2022; Ge et al., 2023, 2024). Additionally, machine translation is employed to generate pseudo-training data for CL-NER, using methods such as dictionary-based translation (Mayhew et al., 2017; Xie et al., 2018), sequence translation models (Liu et al., 2021; Yang et al., 2022) and label projections (Chen et al., 2023; Parekh et al., 2024) for data augmentation. However, these methods often underperform compared to the T-S learning framework, mainly because of their limited entity-aligned translation capabilities. While T-S frameworks achieve strong results for LSL, they struggle with NSL, such as Chinese and Japanese, where linguistic differences hinder accurate semantic and lexicogrammatical alignment.

2.3 LLMs for NER and Multilinguality

Large Language Models (LLMs) have demonstrated significant potential across NLP tasks, including NER (Xie et al., 2023, 2024; Liu et al., 2024). LLMs have also been used for data augmentation to enhance the performance of smaller models (Zhang et al., 2021, 2024; Kang et al., 2024; Ju et al., 2024). Recent studies have further improved their multilingual capabilities, enabling cross-lingual applications (Wu et al., 2025). However, due to performance disparities between LLMs' English and non-English capabilities (Zhu et al., 2024a), many studies use LLMs to translate non-English texts into English before performing downstream tasks (Zhu et al., 2024b; Chen et al., 2024). Inspired by these approaches, we utilize LLMs for entity-aligned translation, enhancing the alignment between source and target languages. The advanced semantic understanding and reasoning abilities of LLMs help capture implicit information during translation, mitigating information loss and improving overall performance.

3 Methodology

In this section, we first introduce our proposed **E**ntity-**A**ligned **T**ranslation (**EAT**) framework, as illustrated in Figure 2: dual translation, source-oriented cross-lingual corpora collection, and entity-level alignment fine-tuning. Then, we present two metrics to evaluate the correlation between translation quality and ZCL-NER performance.

3.1 Dual Translation for ZCL-NER

To minimize entity loss during translation, we propose a dual translation model (**DT**) that focuses on preserving potential entities of target language throughout the process.

Target to Source Forward Translation. LLMs

may not be able to paraphrase the words of entities in single round of inference and may directly output the raw input words. Therefore, the generated translation with raw words from languages that do not use Latin scripts will affect the models' overall understanding and may lead to failure in subsequent tasks. One possible solution is to instruct the LLMs in several rounds instead of single round in each direction, which we call *multi-round text translation with chain-of-thought* (MrCoT).

In the first round, we instruct the model to **consider the entities the target sentence $\boldsymbol{x}$ may contain** and explain those entities, when translating from target language a to source language (English) b. Formally,

$$o_1^t = \text{LM}(p_1^t, \boldsymbol{x}, a, b) \tag{1}$$

where LM denotes a large language model. p_1^t is the first prompt that instructs the model to consider the entities where $\boldsymbol{x}$ may contain and describe them. o_1^t is supposed to provide a CoT context for the next round. Then the second round output is:

$$o_2^t = \text{LM}^{o_1^t}(p_2^t, \boldsymbol{x}, a, b) \tag{2}$$

where p_2^t is the second prompt that instructs the model to translate $\boldsymbol{x}$ taking previous inference o_1^t into consideration. The translation result $\mathcal{T}_{a\to b}^t(\boldsymbol{x})$ is obtained by filtering non-relevant words:

$$\mathcal{T}_{a\to b}^t(\boldsymbol{x}) = \text{LM}^{o_2^t}(p^f, \boldsymbol{x}, a, b) \tag{3}$$

where p^f is the filter prompt. $\mathcal{T}_{a\to b}^t(\boldsymbol{x})$ denotes the complete sentence by target-to-English language translation.

Text-to-Text Generation for NER. To better utilize the semantics of the sentences, we reformulate the NER task as a text-to-text generation task following (Zhang et al., 2024). The inputs are divided as: 1) **PREFIX**(P): define the task as labeling entities of the input sentence. 2) **TAG**: the set $\mathbb{T}$ of entity tags from the dataset. 3) **SENTENCE**: the input sentence $\mathcal{T}_{a\to b}^t(\boldsymbol{x})$. Then, given the entire input $I = (P, \mathbb{T}, \mathcal{T}_{a\to b}^t(\boldsymbol{x}))$, the output entities by generation model are defined as:

$$\boldsymbol{E} = \text{EXTRACTOR}_{\boldsymbol{\rho}}(I) \tag{4}$$

where $\boldsymbol{\rho}$ denotes the trainable parameters of the text-to-text English NER model EXTRACTOR. The result $\boldsymbol{E}$ contains entity pairs as $(\mathcal{T}_{a\to b}^t(\boldsymbol{x})_{l_1:r_1}, tag)$, where $l_1 : r_1$ denotes the text boundary.

Source to Target Backward Translation. This stage plays a crucial role in ensuring entity alignment across languages. In neural machine translation, LLMs may generate tokens that are semantically related to entities, rather than directly translating them. As a result, relying solely on the LLM-generated translation may lead to the omission of possible entities that were captured in the above stage. To ensure effective cross-lingual entity alignment, we check whether the translated results of this stage for potential input entities in English **(i.e., output) closely match the corresponding segments in the real target language**.

Formally, to translate an entity $\mathcal{T}_{a\to b}^t(\boldsymbol{x})_{l_1:r_1}$ from language b to a, the first round output is:

$$o_1^e = \text{LM}(p_1^e, \mathcal{T}_{a\to b}^t(\boldsymbol{x})_{l_1:r_1}, \boldsymbol{x}, a, b) \tag{5}$$

where $l_1 : r_1$ is the text boundary, and p_1^e is the first prompt that instructs the model to translate the entity and analyze if o_1^e could appear in $\boldsymbol{x}$. Then the translation result in second round with MrCoT is obtained similarly:

$$o_2^e = \text{LM}(p_2^e, o_1^e, \boldsymbol{x}, a, b) \tag{6}$$

$$\boldsymbol{x}_{l_2:r_2} = \mathcal{T}_{b\to a}^e(\mathcal{T}_{a\to b}^t(\boldsymbol{x})_{l_1:r_1}) = \text{LM}(p^f, o_2^e) \tag{7}$$

where $l_2 : r_2$ is the text boundary, and p_2^e is the second prompt to instruct the model to check if the result appears in $\boldsymbol{x}$, and p^f is the filter prompt.

At this stage, we can finalize ZCL-NER, where $\boldsymbol{x}_{l_2:r_2}$ represents *the target language entity we aim to extract*.

3.2 Source-oriented Cross-lingual Corpora

To reduce the hallucination and further amplify the ability of entity alignment in the above translation process, we collect English-oriented entity-aligned cross-lingual corpora (EACL) to fine-tune the translation model. The EACL Corpora are collected from **Interlanguage Links** * provided by the Wikipedia API †.

We leverage the entities from the CoNLL2003 English dataset to construct EACL. As shown in Figure 2, for an English entity $e \in \boldsymbol{e}$, the links from the API of e are the content of target language related to e:

$$\mathcal{I}(e) = \{(u^a, \boldsymbol{v}^a)_e | a \in \mathcal{A}\} \tag{8}$$

*Detailed information in https://en.wikipedia.org/wiki/Help:Interlanguage_links

†https://api.wikimedia.org/wiki/Main_Page

	AR	HI	HY	JA
Sen.	2,746	1,195	1,568	2,890
Tok.	69,336	83,287	62,669	217,252
	KA	**KO**	**RU**	**ZH**
Sen.	1,451	2,014	796	2,382
Tok.	8,177	32,667	7,521	206,279

Table 2: Detailed information of our collected EACL Corpora, including the amounts of $(u^a, v_1^a)_e$ and tokens in v_1^a for each language a.

where u^a and $\boldsymbol{v}^a$ denote the title (entity or short description for e in target language a) and summaries of language a corresponding to e (a text to explain and describe title), and $\mathcal{A}$ denotes the set of languages. We leverage u^a as the entity, and the first sentence v_1^a of $\boldsymbol{v}^a$ as the text to construct the entity-description pair: $\mathcal{D}^a = \{\mathcal{I}_a(e)|e \in \boldsymbol{e}\} = \{(u^a, v_1^a)_e|e \in \boldsymbol{e}\}$.

However, not all entities have corresponding Wikipedia pages with interlanguage links. Moreover, not all languages have Wikipedia pages that align with English entities, especially in less-resourced languages. As a result, the corpus size varies across languages. Detailed information on the collected corpora can be found in Table 2.

3.3 Entity-level Alignment Fine-tuning

To amplify the DT model's entity-level alignment ability, we leverage EACL Corpora obtained above to fine-tune (FT) the backward translation. This is because the model struggles to identify the positions of entities. Specifically, we leverage Quantized-LoRA (QLoRA) (Dettmers et al., 2024) to accelerate model fine-tuning under constrained resources. Instead of global quantization, we use block-wise k-bit quantization (Dettmers et al., 2022; Tang et al., 2025):

$$D^2(\boldsymbol{c}, \boldsymbol{c}^d, \boldsymbol{W}^q) = D(D(\boldsymbol{c}, \boldsymbol{c}^d), \boldsymbol{W}^q) = \boldsymbol{W} \quad (9)$$

where the quantization constants $\boldsymbol{c}$ are quantized as $\boldsymbol{c}^d$. $\boldsymbol{W}$ and $\boldsymbol{W}^{\boldsymbol{q}}$ denote the model's raw and quantized weights.

We employ cross-entropy loss to train the DT model:

$$L_T(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{i=1}^{\eta} \hat{y}_i \log(y_i) \quad (10)$$

where $\boldsymbol{y} = \text{LM}(e, v_1^a)$. $\hat{y}_i \in \hat{\boldsymbol{y}}$ is u^a in $(u^a, v_1^a)_e$ of EACL. $y_i \in \boldsymbol{y}$ denotes the predicted entity. LM denotes a large language model and η denotes the max length of model output.

Lang.	Tokens Train	Tokens Valid	Tokens Test	Ratio
AR	129,184	64,291	64,347	4.96%
HI	29,443	5,808	6,005	11.05%
HY	95,614	6,214	6,220	0.08%
JA	603,301	300,844	306,959	1.60%
KA	80,402	81,159	81,922	0.05%
KO	162,031	80,786	80,841	1.06%
RU	141,529	70,279	71,288	3.33%
ZH	420,054	213,682	207,505	17.75%
EN	160,394	80,536	80,326	19.76%

Table 3: Detailed information of our selected languages in the dataset WikiANN. Ratio represents the approximate proportion of speakers to the total world population, and the statistics are referred from Wikipedia.

3.4 Evaluating Entity-aligned Translation

We use BLEU score (Papineni et al., 2002) and Information Entropy (Shannon, 1948) to measure the information loss in the translation process, so as to demonstrate the relevance between the NER results and the information loss.

Bilingual Evaluation Understudy (BLEU) is widely used as an evaluation metric in machine translation due to its fast and unified features. In our setting, the BLEU score is evaluated between the generated target sentence in backward translation and original input in forward translation.

Information Entropy (Shannon Entropy) is also commonly used to quantify the information of the sentences. For better analysis, we leverage the Bi-Gram Model to calculate the joint information entropy as: $H(\boldsymbol{s}) = \sum_{i=1}^{n} H(s_{i-1}, s_i)$.

To evaluate the information loss between raw target sentence and dual translated sentence, we define the measurement using the above Shannon Entropy as: $L_e = \frac{H(\mathcal{T}_{b\to a}^t(\boldsymbol{s}_r))}{H(\boldsymbol{s}_r)}$, where $\boldsymbol{s}_r$ denotes the sentence of target language b and a denotes the source language.

4 Experimentation

4.1 Datasets

The experiments are conducted on two public and most widely used datasets, including WikiANN (Pan et al., 2017) and CoNLL2003 (Tjong Kim Sang and De Meulder, 2003):

1) **WikiANN** involves 176 languages, and each language has balanced train, valid, and test splits. The entity categories of WikiANN are PERSON (PER), LOCATION (LOC), and ORGANIZATION

Models	Non-Latin Scripts								
	AR	**HI**	**HY**	**JA**	**KA**	**KO**	**RU**	**ZH**	**Avg.**
mBert (Wu and Dredze, 2019)	42.30	64.79	52.12	29.82	64.68	57.38	64.09	43.85	52.38
A-align(Dou and Neubig, 2021)	46.00	73.90	49.83	20.30	70.40	57.70	64.80	45.40	53.54
CROP (Yang et al., 2022)	52.44	55.55	44.49	45.37	46.02	48.93	50.73	45.33	48.61
EasyProject(Chen et al., 2023)	34.40	73.00	48.61	41.30	66.40	48.20	66.30	42.00	52.54
CLaP(Parekh et al., 2024)	48.70	73.10	51.68	45.30	70.50	60.10	68.30	49.70	58.42
TSLM (Wu et al., 2020)	43.12	65.26	53.56	31.19	66.20	58.94	66.02	45.60	53.74
RIKD (Liang et al., 2021)	45.96	65.69	55.17	31.49	66.83	58.03	65.63	47.38	54.52
AdvPicker (Chen et al., 2021)	49.16	70.00	52.49	37.62	68.37	59.25	68.28	53.02	57.27
DualNER (Zeng et al., 2022)	59.00	66.24	55.92	31.07	67.28	57.48	65.06	47.84	56.24
MSD (Ma et al., 2022)	62.88	73.43	56.22	33.34	69.23	61.44	67.71	57.06	60.16
ProKD (Ge et al., 2023)	50.91	70.72	62.58	33.72	69.07	61.31	65.59	51.80	58.21
DenKD (Ge et al., 2024)	60.01	69.76	65.20	37.90	69.30	62.51	69.35	55.62	61.21
EAT w/o FT	66.53	**76.26**	65.62	45.43	71.63	**66.03**	**71.45**	60.12	65.38
EAT	**67.29**	75.46	**68.02**	**52.26**	**73.68**	65.23	63.25	**61.27**	**65.81**

Table 4: Performance comparison of existing zero-shot CL-NER studies and our approaches. Bold represents the best result, and underlining represents the second best result.

(ORG). We select 8 non-Latin script languages as our test sets (target language), representing both widely and less widely spoken languages: Arabic (AR), Hindi (HI), Armenian (HY), Japanese (JA), Georgian (KA), Korean (KO), Russian (RU), and Chinese (ZH). The English (EN) train and valid sets (source language) are leveraged to build the NER Extractor. Detailed information and statistics about these languages are presented in Table 3.

2) **CoNLL2003** includes two languages: English and German. Each language has train, valid, and test splits. Due to its higher quality of manual annotation compared to WikiANN, we use the entity phrases only from the English train set to construct the EACL Corpora.

4.2 Baselines and Implementation

Baselines. To compare our approach with previous translation-based and teacher-student approaches, we report the baselines as follows:

Translation based: 1) **mBert** (Wu and Dredze, 2019). 2) **A-align** (Dou and Neubig, 2021) 3) **CROP** (Yang et al., 2022) 4) **EasyProject** (Chen et al., 2023) 5) **CLaP** (Parekh et al., 2024).

T-S based: 3) **TSLM** (Wu et al., 2020). 4) **RIKD** (Liang et al., 2021). 5) **AdvPicker** (Chen et al., 2021). 6) **DualNER** (Zeng et al., 2022). 7) **MSD** (Ma et al., 2022). 8) **ProKD** (Ge et al., 2023). 9) **DenKD** (Ge et al., 2024), the *SOTA* for ZCL-NER.

Implementation Details. The EACL Corpora are leveraged to fine-tune the model **Qwen2.5-14B-Instruct** (Yang et al., 2024; Qwen Team, 2024). We first quantize the model to 4-bits, and freeze its parameters. The rank parameter and the scale parameter of the Low-Rank Adapter are set to 64 and 16. The ratio of train and valid sets is 90: 10, and the model is trained using the divided corpora for 5 epochs with the learning rate set to $1.0e{-}4$.

We leverage English train and valid sets from the WikiANN dataset to train the **Flan-T5-Base** model (Chung et al., 2024). After 50 epochs of training with a learning rate of $1.0e{-}4$, we will obtain the final NER Extractor.

Since there are no reference translation results (entities) for the WikiANN dataset, we utilize the raw input $\boldsymbol{x}$ of target language and dual translated text $\mathcal{T}^t_{b\to a}(\mathcal{T}^t_{a\to b}(\boldsymbol{x}))$ to calculate the **BLEU** scores and **Entropy Loss**.

4.3 Main Results

We use the token-level micro F1 score to evaluate the NER results, following previous works (Ge et al., 2023, 2024; Ma et al., 2022). The baseline results of T-S based approaches are cited from their papers. Based on the performance comparison in Table 4, we primarily address the following questions:

How does EAT perform and why do we design it? Translation-based methods perform poorly in NSL. This indicates that the traditional translation schema between English and NSL is not suitable for ZCL-NER. Therefore, we need to develop a completely new translation mechanism to account for the significant differences between English and NSL, which aligns with our core motivation. Furthermore, although T-S framework (e.g., DenKD) apparently outperforms previous translation meth-

	AR		HI		JA		ZH	
	F1.↑	BLEU↑	F1.↑	BLEU↑	F1.↑	BLEU↑	F1.↑	BLEU↑
EAT-14B	66.53	12.10	76.26	10.66	45.43	1.86	60.12	9.26
EAT-7B	65.30	9.70	73.89	8.57	40.25	1.60	59.73	8.81
EAT-3B	54.93	6.63	67.87	6.03	34.25	1.46	55.62	7.18
EAT-1.5B	46.51	4.10	60.59	2.22	28.06	0.57	48.46	5.54

Table 5: Performance comparison of our EAT w/o FT using homologous LLMs with different sizes of parameters.

	AR	HI	JA	ZH
EAT-T5-NER	67.29	76.26	52.26	61.27
EAT-mBERT-NER	67.63	73.75	52.72	60.99

Table 6: Performance comparison of our approach using different NER EXTRACTOR models. LLM for DT keeps 14B for fair comparison.

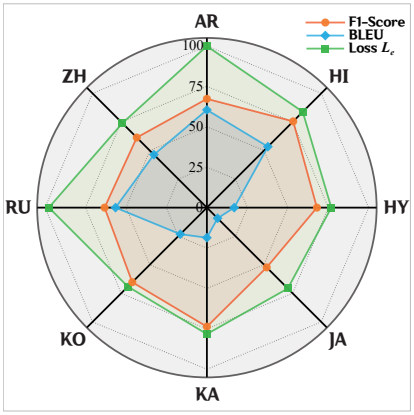

Figure 3: The relevance of BLEU scores and entropy loss compared with NER results (F1-scores). The BLEU scores are not the accurate values, we resize them to draw the plot.

ods (e.g., CLaP), the principles of the T-S series remain largely similar. As a result, its effectiveness in bridging the linguistic gap between English and NSL is limited. Consequently, performance gains over previous works from the past two years have remained modest, averaging only at 1%-2%. However, our EAT achieves a substantial improvement of 4% over SOTA. This suggests that our approach genuinely addresses the fundamental disparities between different languages.

In particular, the T-S based approaches perform particularly poorly on JA and ZH, almost at 30%-55%. However, we found that there are a significant number of people worldwide speak these languages, as the ratio in Table 3. This motivates us to enhance the performance of NER for these languages. Therefore, we develop an intuitive and effective approach that significantly outperforms SOTA in ZCL-NER.

Why doesn't EAT w/ FT always perform better than it w/o FT? Although our EAT w/ FT performs well on average, it does not take effect in certain languages. One possible reason is inductive bias. The inductive bias is the structure imposed in the FT datasets to instruct the LLMs how to think, and it may be toxic:

> [*]*Clever structures posed by human researchers typically become the bottleneck when scaled up.*

The fine-tuning on EACL allows LLMs to learn new knowledge, but the structures introduced by FT may hinder LLMs' inference ability (Kimi Team et al., 2025; DeepSeek-AI et al., 2025). In other words, FT may force LLMs to think structurally, hence hindering the inference process where LLMs can correct previously generated errors.

4.4 Analysis and Discussion

Impact of Translation Ability. We present Figure 3 and Table 5 to examine *whether translation ability directly affects NER performance by the translation quality evaluation metrics we introduced.*

Figure 3 shows that on each language, the coordinate point of the F1-Score is always between BLEU and Loss L_e. This suggests a positive correlation between NER performance and translation ability. Table 5 demonstrates that the BLEU score decreases as the size of homologous LLMs decreases, which means that the model's translation ability weakens. This, in turn, negatively impacts the performance of our EAT framework. These findings underscore the critical importance of improving translation ability to enhance ZCL-NER performance. In addition, we also evaluate our approach with heterologous LLMs (different backbones) as presented in Appendix D.

Impact of NER Extractor Ability. To better compare our approach with T-S based approaches, we leverage **mBERT** (Pires et al., 2019) with sequence-labeling as the NER model. As shown in Table 6, there is a little difference between mBERT and T5 as both models are well-trained for the English NER task. In general, our EAT with T5 performs slightly better than it with mBERT. This suggests the robustness of our EAT.

Impact of Direct Using LLMs for ZCL-NER. We leverage **both LSL and NSL LLMs** (GPT and

[*]Hyung Won Chung (OpenAI). 2024-05. Don't teach. Incentivize: Scale-first view of Large Language Models. MIT EI seminar.

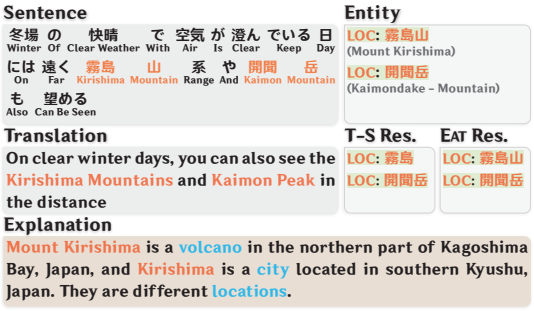

Figure 4: Comparison of the NER results for T-S based approaches (T-S Res.) and our approach (EAT Res.).

	AR	HI	JA	ZH
EAT w/o FT	**66.53**	**76.26**	**45.43**	**60.12**
GPT-4	60.60	56.14	34.44	52.35
Qwen-14B-ICL	45.71	61.18	41.95	52.89
Qwen-14B-DA	29.82	72.27	41.34	49.05

Table 7: Performance comparison of our approach and CoT-based LLMs for ZCL-NER. ICL denotes in-context learning for Qwen with source labeled data. DA denotes data augmentation for NER model with translated source labeled data. Details are described in Appendix Eand F.

Qwen) directly on the ZCL-NER task to explore whether LLMs could perform better without our proposed Dual-Translation mechanism. As shown in Table 7, our approach performs better than GPT-4 (OpenAI et al., 2024), which is recognized as one of the best LLMs. Even directly using source-labeled data for Qwen in-context learning (**ICL**) or translated into target language (**DA**) both fail to improve its poor NER performance, which aligns with insights from existing studies (Xie et al., 2023, 2024; Li et al., 2025a,b).

Generalization Ability. As shown in Table 8, our approach also outperforms the SOTA teacher-student model and GPT-4 significantly on MultiCoNER-1 (Malmasi et al., 2022). This suggests the excellent generalization ability of our proposed **EAT**.

Case Study. From **NER Results** aspect, as shown in Figure 4, our approach EAT correctly comprehends the semantics of the given sentence due to accurate translation and explanation. Hence, EAT accurately identifies the entities and their corresponding tags. However, T-S based approaches struggle to grasp the meaning of all phrases. As a result, they recognize a wrong entity that mismatches the intended semantics.

From **Entity Alignment** aspect, as shown in Figure 5, EAT model w/ FT seems to become in-

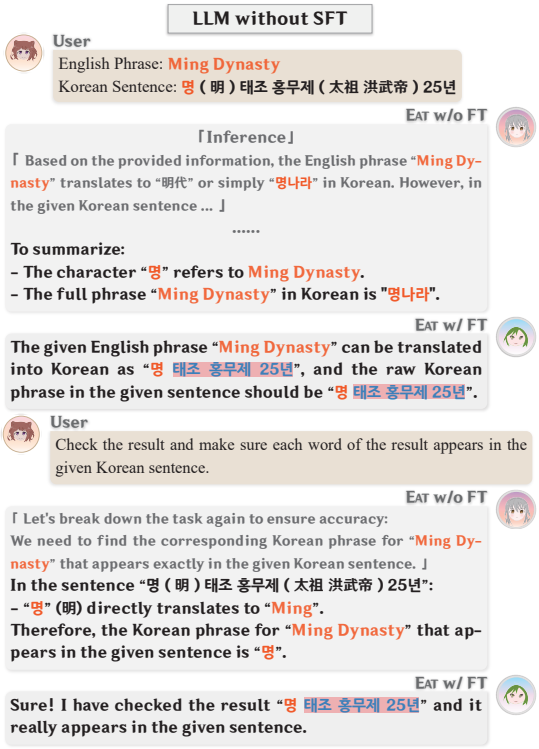

Figure 5: Comparison of the entity alignment process for LLM with and without FT. Texts in pink background denote the wrong alignments, w/o pink denotes right.

	HI	KO	RU	ZH
EAT w/o FT (Ours)	**47.38**	**64.12**	**52.66**	**63.00**
GPT-4	47.02	49.45	37.72	48.58
DenKD (Ge et al., 2024)	33.67	44.61	45.26	41.48

Table 8: Performance comparison on MultiCoNER-1 (Malmasi et al., 2022).

flexible. It loses the ability of inference and thus outputs the wrong entity alignment. However, EAT model w/o FT can infer multiple alignment results for entities from the source language to the target language. This not only presents error analysis of our EAT w/ FT, but also indicates that the inference progress will enhance alignment results and makes the NER results more accurate. This finding also matches recent studies (DeepSeek-AI et al., 2025; Kimi Team et al., 2025): reasoning progress can improve LLMs' performance.

5 Conclusion

We introduce a novel entity-aligned translation (EAT) approach with LLMs for zero-shot cross-lingual NER (ZCL-NER) approach to mitigate the linguistic differences between non-Latin script language (NSL) and English, so as to better leverage the rich English NER resources for multilingual

NER tasks. In addition, we fine-tune the LLM using collected source-oriented cross-lingual corpora to enhance entity alignments for better NER. Furthermore, we employ BLEU and information entropy to analyze the correlation between NER performance and translation ability.

Acknowledgements

This work was supported by NSFC grants (No. 62206193 and No. 62376178) and General Research Fund (GRF) project sponsored by the Research Grants Council Hong Kong (Project No.15611021).

Limitations

Although our approach has achieved impressive results on zero-shot cross-lingual NER, there are still limitations. The LLMs are pre-trained on uneven corpora in different languages, making their multilingual ability differs. However, multilingual semantic understanding is crucial in our approach, as it directly affect translation ability. If the LLM fails to operate dual translation, our approach will also fail. In addition, the amount of knowledge in different languages correlates positively with the size of LLMs, and the model size affects computational resources usage. For instance, Qwen2.5 supports 29 languages while GPT-4 supports over 80, with significantly different resource requirements. Therefore, we must strike a balance between performance and universality, where the LLM is large enough to undertake the multilingual dual translation while remaining as small as possible to maximize the inference speed and minimize the energy usage and carbon emissions.

References

Jorge Arús-Hita, Kazuhiro Teruya, Mohamed Ali Bardi, Abhishek Kumar Kashyap, and Isaac N. Mwinlaaru. 2018. Quoting and reporting across languages: A system-based and text-based typology. *WORD*, 64(2):69–102.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Shu-Kun Chen. 2016. Circumstantiation of projection: Functional syntax of angle in english and chinese. *Ampersand*, 3:71–82.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. AdvPicker: Effectively Leveraging Unlabeled Data via Adversarial Discriminator for Cross-Lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 743–753, Online. Association for Computational Linguistics.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

DeepSeek-AI, Daya Guo, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Zoe Finiasz, Susan A. Gelman, and Tamar Kushnir. 2024. Testimony and observation of statistical evidence interact in adults' and children's category-based induction. *Cognition*, 244:105707.

Ling Ge, Chunming Hu, Guanghui Ma, Jihong Liu, and Hong Zhang. 2024. Discrepancy and uncertainty aware denoising knowledge distillation for zero-shot cross-lingual named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18056–18064.

Ling Ge, Chunming Hu, Guanghui Ma, Hong Zhang, and Jihong Liu. 2023. Prokd: An unsupervised prototypical knowledge distillation network for zero-resource cross-lingual named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12818–12826.

Susan A. Gelman and Ellen M. Markman. 1986. Categories and induction in young children. *Cognition*, 23(3):183–209.

Susan A Gelman and Twila Tardif. 1998. A cross-linguistic comparison of generic noun phrases in english and mandarin. *Cognition*, 66(3):215–248.

Michael Halliday and Christian Matthiessen. 1999. *Construing Experience Through Meaning: A Language Based Approach to Cognition*. Cassell, London.

Xincheng Ju, Dong Zhang, Suyang Zhu, Junhui Li, Shoushan Li, and Guodong Zhou. 2024. ECFCON: emotion consequence forecasting in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 2233–2241. ACM.

Hyeonseok Kang, Hyein Seo, Jeesu Jung, Sangkeun Jung, Du-Seong Chang, and Riwoo Chung. 2024. Guidance-based prompt data augmentation in specialized domains for named entity recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 665–672, Bangkok, Thailand. Association for Computational Linguistics.

Kimi Team, Angang Du, and et al. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, Jing Xu, Guoxiu He, and Meihuizi Jia. 2021. Modularized interaction network for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 200–209, Online. Association for Computational Linguistics.

Xiaomeng Li and Keiko Koda. 2022. Linguistic constraints on the cross-linguistic variations in l2 word recognition. *Reading and Writing*, 35(6):1401–1424.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. 2025a. M^{2}iv: Towards efficient and fine-grained multimodal in-context learning via representation engineering. *Preprint*, arXiv:2504.04633.

Yanshu Li, Tian Yun, Jianjiang Yang, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025b. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. *arXiv preprint arXiv:2505.17098*.

Zhuoran Li, Chunming Hu, Junfan Chen, Zhijun Chen, Xiaohui Guo, and Richong Zhang. 2024. Improving zero-shot cross-lingual transfer via progressive code-switching. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6388–6396. ijcai.org.

Shining Liang, Ming Gong, Jian Pei, Linjun Shou, Wanli Zuo, Xianglin Zuo, and Daxin Jiang. 2021. Reinforced iterative knowledge distillation for cross-lingual named entity recognition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 3231–3239, New York, NY, USA. Association for Computing Machinery.

Liang Liu, Dong Zhang, Shoushan Li, Guodong Zhou, and Erik Cambria. 2024. Two heads are better than one: Zero-shot cognitive reasoning via multi-llm knowledge fusion. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 1462–1472. ACM.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Llama team, Aaron Grattafiori, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jun-Yu Ma, Beiduo Chen, Jia-Chen Gu, Zhenhua Ling, Wu Guo, Quan Liu, Zhigang Chen, and Cong Liu. 2022. Wider & closer: Mixture of short-channel distillers for zero-shot cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5171–5183, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2536–2545, Copenhagen, Denmark. Association for Computational Linguistics.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A

large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Hye K Pae. 2020. *Script effects as the hidden drive of the mind, cognition, and culture*. Springer Cham.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual label projection for cross-lingual structured prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Anisha Singh, Min Wang, and Yasmeen Faroqi-Shah. 2022. The influence of romanizing a non-alphabetic l1 on l2 reading: the case of hindi-english visual word recognition. *Reading and Writing*, 35(6):1475–1496.

Quanwei Tang, Sophia Yat Mei Lee, Junshuang Wu, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. 2025. A comprehensive graph framework for question answering with mode-seeking preference alignment. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21504–21523. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. 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.

Dan Wu, Xincheng Ju, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. 2025. Emotion across modalities and cultures: Multilingual multimodal emotion-cause analysis with memory-inspired framework. In *Proceedings of ACM MM 2025*.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical study of zero-shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7935–7956, Singapore. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593, Mexico City, Mexico. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan

Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. CROP: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 486–496, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yanning Yang. 2008. Typological interpretation of differences between chinese and english in grammatical metaphor. *Language Sciences*, 30(4):450–478.

Jiali Zeng, Yufan Jiang, Yongjing Yin, Xu Wang, Binghuai Lin, and Yunbo Cao. 2022. DualNER: A dual-teaching framework for zero-shot cross-lingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1837–1843, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14347–14355. AAAI Press.

Zhihao Zhang, Sophia Yat Mei Lee, Junshuang Wu, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. 2024. 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, Miami, Florida, USA. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

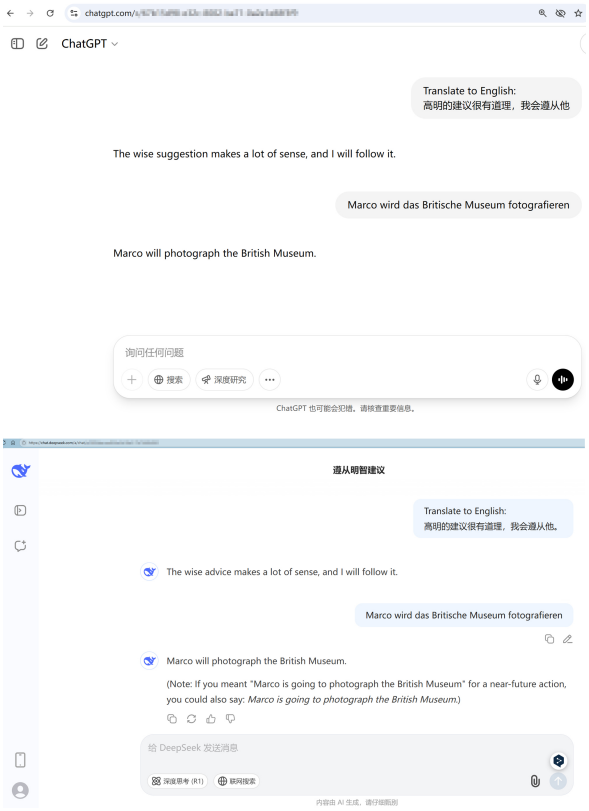

Figure 6: The screenshots of the examples in Figure 1. The left is ChatGPT-4o and the right is Deepseek-V3.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024a. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *Preprint*, arXiv:2405.01345.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A Evidences of our Motivation

As the case in Figure 1, the Person entity "高明" in target language (Chinese) is incorrectly translated as the adjective "clever" by GPT-4o and Deepseek **tested in May 2025**. This can be seen from Figure 6 as evidences. Such translation inconsistencies make it impossible to proceed our task, naturally motivating this work.

B Details of Method

B.1 Task Formulation

Given an n-token sentence $\boldsymbol{x} =< x_1, \cdots, x_n >$ and k-type entity set $\mathbb{T} =< t_1, \cdots, t_k >$, the ob-

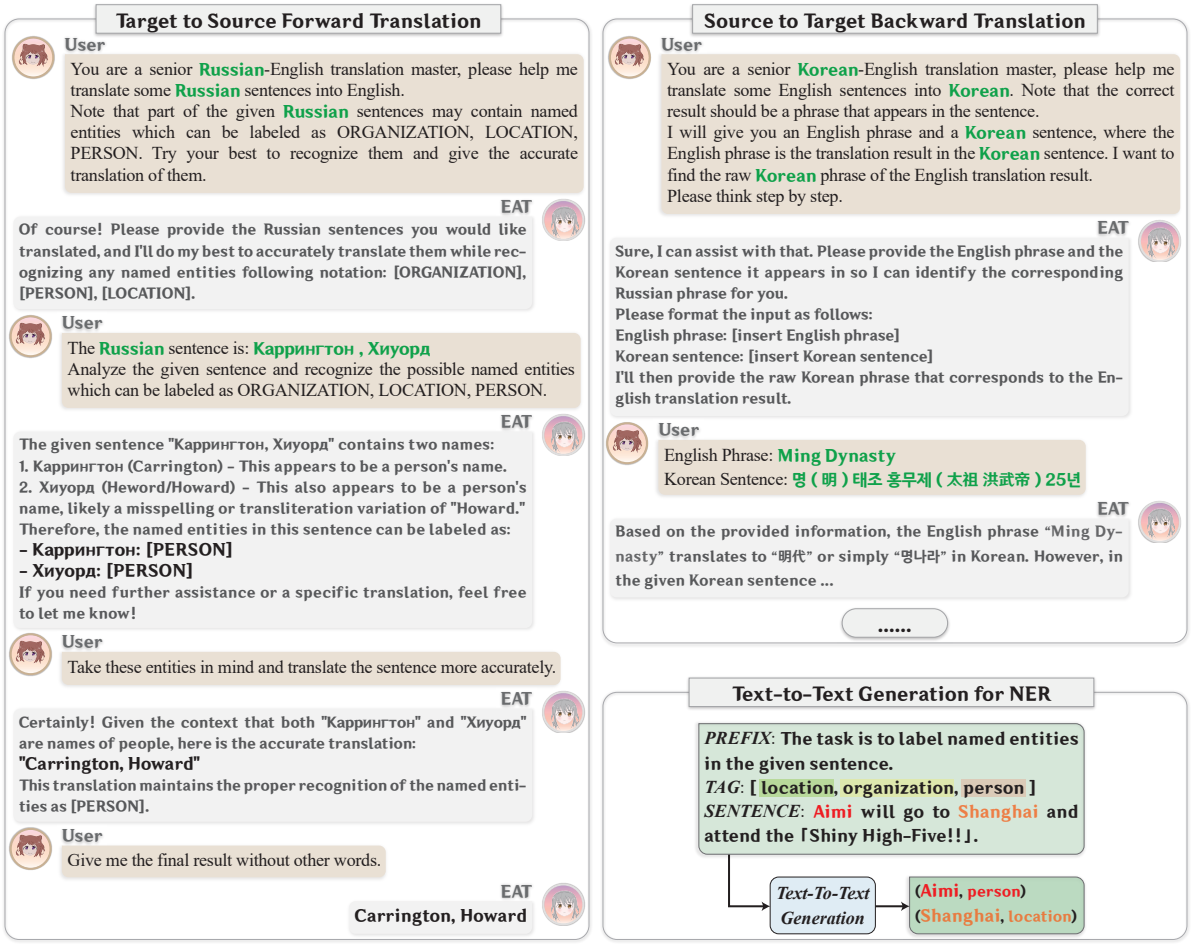

Figure 7: Full process of Dual Translation for ZCL-NER. Texts in green denote the interchangeable templates. The portion which has been appeared in Figure 5 is omitted in Source to Target Backward Translation. *

ject of NER task is to extract all entities $\boldsymbol{e}_i \in \boldsymbol{E}$ from $\boldsymbol{x}$ and assign one of the types in $\mathbb{T}$ to each entity, where $\boldsymbol{e}_i = (\boldsymbol{x}_{start:end}, t)$ denotes the i-th entity of $\boldsymbol{x}$ and $t \in \mathbb{T}$ refers to the type of the entity. $\boldsymbol{x}_{start:end}$ refers to a continues word span $< x_{start}, \cdots, x_{end} >$ in $\boldsymbol{x}$, where $start$ and end refers to the entity boundary indexes respectively. Given dataset $\mathcal{D}_s$ of the source language (i.e., English in our setting) and dataset $\mathcal{D}_t$ of the target language, the objective of the ZCL-NER task is to acquire target-related knowledge from $\mathcal{D}_s$ to enhance model's performance on $\mathcal{D}_t$.

B.2 Detailed Process of Dual Translation for ZCL-NER

To better describe our proposed EAT framework, we give an example in Figure 7 of the key process: Target to Source Forward Translation, Text-to-Text Generation for NER, and Source to Target Backward Translation.

Prompts Since our approach is not focusing on prompt engineering, we just use the only prompts shown in Figure 5 and 7.

B.3 Details of Entity-level Alignment Fine-tuning

To amplify the DT model's entity-level alignment ability, we leverage EACL Corpora obtained above to fine-tune the model. Specifically, we leverage Quantized-LoRA(QLoRA) (Dettmers et al., 2024) to accelerate model fine-tuning and reduce memory usage under constrained resources. Instead of global quantization, we use block-wise k-bit quantization (Dettmers et al., 2022) for its improvements in terms of accuracy, efficiency, and flexibility. More formally, for a given tensor $\boldsymbol{T}$, it is chunked into n contiguous blocks and each block is flattened to $[-1, 1]$:

$$Q_b(\boldsymbol{T}) = [\boldsymbol{T}_1^q, \cdots, \boldsymbol{T}_n^q] \tag{11}$$

*Part of the images are provided by Gu: https://b23.tv/RAKSxQg

The i-th block is independently quantized as:

$$\begin{aligned} \boldsymbol{T}_i^q = Q(\boldsymbol{T}_i) &= round(\frac{2^{k-1}-1}{absmax(\boldsymbol{T}_i)}\boldsymbol{T}_i) \\ &= round(c_i \cdot \boldsymbol{T}_i) \end{aligned} \tag{12}$$

where $c_i \in \boldsymbol{c}$ is the quantization constant for each block. And the dequantization is:

$$\boldsymbol{T}_i = D(c_i, \boldsymbol{T}_i^q) = \frac{\boldsymbol{T}_i^q}{c_i} \tag{13}$$

The model is quantified in 4-bit NormalFloat (NF4) as described above, and the quantization constants $\boldsymbol{c}$ are quantized as $\boldsymbol{c}^d$ to further reduce memory usage. For a single layer's parameters of the model with LoRA adapter is:

$$\boldsymbol{Y} = \boldsymbol{X} D^2(\boldsymbol{c}, \boldsymbol{c}^d, \boldsymbol{W}^q) + \boldsymbol{X}\boldsymbol{L}_1\boldsymbol{L}_2 \tag{14}$$

$$D^2(\boldsymbol{c}, \boldsymbol{c}^d, \boldsymbol{W}^q) = D(D(\boldsymbol{c}, \boldsymbol{c}^d), \boldsymbol{W}^q) = \boldsymbol{W} \tag{15}$$

where $\boldsymbol{W}$ and $\boldsymbol{W}^{\boldsymbol{q}}$ denote the model's raw and quantized weights, and $\boldsymbol{L}_1\boldsymbol{L}_2$ denotes the trainable parameters of the LoRA adapters.

Cross-entropy loss is optimized to train the DT model:

$$\boldsymbol{y} = \textsc{Lm}(e, v_1^a) \tag{16}$$

$$L_T(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum_{i=1}^{\eta} \hat{y}_i \log(y_i) \tag{17}$$

where $\hat{y}_i \in \hat{\boldsymbol{y}}$ is u^a in $(u^a, v_1^a)_e$ of EACL. $y_i \in \boldsymbol{y}$ denotes the predicted entity. LM denotes a large language model and η denotes the max length of model output.

B.4 Details of Entity-aligned Translation Ability Evaluation

We use BLEU score (Papineni et al., 2002) and Information Entropy (Shannon, 1948) to measure the information loss in the translation process, so as to demonstrate the relevance between the NER results and the information loss.

Bilingual Evaluation Understudy (BLEU) is widely used as an evaluation metric in machine translation. BLEU is a fast and unified metric, and it can evaluate all languages effectively. Since there are no reference translations of source language a for target language b, we obtain the candidate as:

Given a sentence $\boldsymbol{s}_r$ of target language b, the candidate sentence is:

$$\boldsymbol{s}_c = \mathcal{T}_{a\to b}^t(\mathcal{T}_{b\to a}^t(\boldsymbol{s}_r)) \tag{18}$$

And the n-gram precision is:

$$\mathcal{P}_n = \frac{\sum_k^{K_n} \min(h_k(\boldsymbol{s}_c), h_k(\boldsymbol{s}_r))}{\sum_k^{K_n} h_k(\boldsymbol{s}_c)} \tag{19}$$

where K_n denotes the n-gram divided sequence, and $h_k(\cdot)$ denotes the counts of k-th n-gram.

The **Brevity Penalty** is introduced to avoid the scoring bias:

$$BP = \begin{cases} 1 & \text{if} \quad l_c > l_r \\ e^{1-l_r/l_c} & \text{if} \quad l_c \le l_r \end{cases} \tag{20}$$

where l_c and l_r are the lengths of $\boldsymbol{s}_c$ and $\boldsymbol{s}_r$.

We adopt $n = 4$, and the BLEU score is finally calculated by the geometric mean of the n-gram precision:

$$BLEU = BP \times \exp(\frac{1}{4}\sum_{n=1}^{4} \log(\mathcal{P}_n)) \tag{21}$$

Information Entropy (Shannon Entropy) is also commonly used to quantify the information of the sentences. We leverage the Bi-Gram Model to calculate the joint information entropy as:

$$\begin{aligned} H(\boldsymbol{s}) &= \sum_{i=1}^{n} H(s_{i-1}, s_i) \\ &= \sum_{i=1}^{n} P(s_{i-1}, s_i)(-\log P(s_{i-1}|s_i)) \end{aligned} \tag{22}$$

where $P(s_{i-1}, s_i)$ denotes the joint probability of s_{i-1}, s_i appearing in the n-length text $\boldsymbol{s}$ with s_{i-1} exactly before s_i, and $P(s_{i-1}|s_i)$ denotes the conditional probability of s_{i-1} appearing before s_i.

We demonstrate the information loss is relevant to the entropy loss. The entropy loss L_e is defined using the above Shannon Entropy $H(\cdot)$ as:

$$L_e = \frac{H(\mathcal{T}_{b\to a}^t(\boldsymbol{s}_r))}{H(\boldsymbol{s}_r)} \tag{23}$$

where $\boldsymbol{s}_r$ denotes the sentence of target language b and a denotes the source language.

C More Implementation Details

As shown in Figure 2, we first collect the EACL Corpora as described in Section 3.2. The number of original English entities we used in the CoNLL2003 train set is 8,082. After obtaining all entity-text pairs, we remove those pairs where the entity does not appear in the corresponding text. Then the texts and entities are constructed in ShareGPT format following previous work (Zheng et al., 2024). Detailed statistics of our collected corpora are listed in Table 2.

Lang.	Tokens			Ratio
	Train	Valid	Test	
DE	195,387	97,805	97,646	1.75%
ES	129,283	64,329	64,728	7.30%
FR	136,788	68,220	68,754	4.07%
NL	169,449	84,146	85,122	0.39%
EN	160,394	80,536	80,326	19.76%

Table 9: Detailed information of our selected languages in the dataset. Ratio represents the approximate proportion of speakers to the total world population, and the statistics are referenced from Wikipedia.

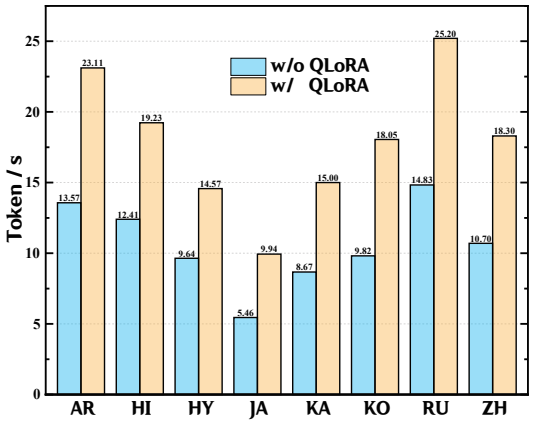

Figure 8: Speed of token generation on different languages.

C.1 Prompts

All prompts used in our proposed **EAT** are listed in Figure 5 and 7. It is worth noting that these prompts are only normal descriptions and instructions of what we want LLMs to do, which do not require special design or strict screening. Therefore, we believe that there is no need to conduct experiments based on the same meaning but different forms of prompts.

C.2 Baselines

To compare our approach with previous translation-based and T-S framework based approaches, we report the baselines as follows:

Translation based:

1) **mBert** (Wu and Dredze, 2019) leverages a pre-trained model to directly transfer from source languages to target languages.

2) **Awesome-align** (Dou and Neubig, 2021) fine-tunes PLMs with paralleled data on source and target languages to extract label alignments.

3) **CROP** (Yang et al., 2022) leverages a sequence translation model to operate the ZCL-NER task with a cross-lingual entity projection framework.

4) **EasyProject** (Chen et al., 2023) improves mark-then-translate method to better perform translation and label projection.

5) **CLaP** (Parekh et al., 2024) proposes contextual translation to better translate the labels to the target languages.

Teacher-Student Framework based:

3) **TSLM** (Wu et al., 2020) proposes vanilla teacher-student learning to distill knowledge for cross-lingual NER.

4) **RIKD** (Liang et al., 2021) proposes a teacher-student learning approach with reinforcement-learning-based knowledge distillation.

5) **AdvPicker** (Chen et al., 2021) introduces adversarial learning in the training process of the teacher model to denoise in knowledge distillation.

6) **DualNER** (Zeng et al., 2022) proposes a unified framework that combines NER learning paradigms and applies multi-task learning for knowledge distillation.

7) **MSD** (Ma et al., 2022) designs a multichannel distillation framework with a parallel domain adaptation to efficiently transfer information.

8) **ProKD** (Ge et al., 2023) proposes prototypical alignment with prototypical self-training for knowledge distillation to better acquire knowledge.

9) **DenKD** (Ge et al., 2024) proposes a denoising approach using uncertainty- and discrepancy-awareness to reduce the noise in the knowledge distillation process, which is the *SOTA model*.

D Additional Results

D.1 Results on Latin Script Languages

We also conduct experiments on LSL for better comparison, including German (DE), Spanish (ES), French (FR), and Dutch (NL). As shown in Table 10, our approach achieves similar performance compared with previous teacher-student learning models. This illustrates our approach works on not only NSL but also LSL.

D.2 Results on CoNLL dataset

To better evaluate our approach, we conduct experiments on the CoNLL dataset and use **Llama3-Instruct** (Llama team et al., 2024) to compare. As shown in Table 11, Qwen2.5 defeats Llama3 with the same level parameters on all languages. This result matches the translation ability difference evaluated in Qwen Team (2024).

	Latin Scripts			
	DE	**ES**	**FR**	**NL**
mBert (Wu and Dredze, 2019)	78.64	74.55	80.20	82.55
CROP (Yang et al., 2022)	60.87	62.05	55.79	54.77
TSLM (Wu et al., 2020)	79.96	77.18	80.38	83.80
RIKD (Liang et al., 2021)	80.20	77.79	81.20	84.65
AdvPicker (Chen et al., 2021)	79.72	77.81	79.91	84.27
DualNER (Zeng et al., 2022)	80.17	78.42	80.92	84.36
MSD (Ma et al., 2022)	80.62	75.75	81.16	84.23
ProKD (Ge et al., 2023)	79.74	79.19	81.45	<u>84.73</u>
DenKD (Ge et al., 2024)	<u>82.50</u>	**84.68**	82.34	**85.69**
EAT	**83.91**	83.43	<u>84.50</u>	81.79
EAT w/ SFT	81.69	<u>84.35</u>	**85.22**	81.69

Table 10: Performance comparison of our approaches on languages using Latin scripts. Bold represents the best result, and underlining represents the second best result.

	DE		ES		NL	
	Qwen	Llama	Qwen	Llama	Qwen	Llama
EAT-14B	77.51	-	82.84	-	78.76	-
EAT-7B*	71.39	60.16	78.36	62.06	76.24	-
EAT-3B	56.05	48.20	56.90	49.33	58.46	-
EAT-1B*	40.31	19.16	40.60	15.71	40.39	-

Table 11: Performance comparison on CoNLL dataset using different models as backbones. '-' represents no result since Dutch is not officially supported by Llama3 and Llama3 does not have an official 14B version.

	HI	**KO**	**RU**	**ZH**
EAT w/o FT (Ours)	**47.38**	**64.12**	**52.66**	**63.00**
GPT-4	47.02	49.45	37.72	48.58
DenKD (Ge et al., 2024)	33.67	44.61	45.26	41.48

Table 12: Performance comparison of our approach with previous SOTA approaches on MultiCoNER-1. Due to the limit of computational resources and time, the test set is downsized through random sampling.

D.3 Results on MultiCoNER-1

To better evaluate the generalization ability of our approch, we conduct experiments on MultiCoNER-1 (Malmasi et al., 2022). MultiCoNER-1 is a larger dataset and all the annotations are made by human. We mainly select four most representative NSLs and two SOTA approaches for fair comparison, as shown in Table 12. From this table, we can see that our approach obviously outperforms the SOTA teacher-student model and GPT-4. This reveals the excellent generalization ability of our proposed approach **EAT**.

D.4 Ablation Study on CoT Rounds

To measure round differences, we add ablation studies about the round depth. As shown in Table 13, the results of round 1 were not good, while the results of round 5 were too time-consuming and unstable. Therefore, 3 rounds should be a realistic choice..

	ZH	**AR**
5 rounds	61.33	66.58
3 rounds	61.27	67.29
1 round	59.62	64.17

Table 13: Performance comparison of our approach with different CoT rounds.

D.5 Inference Acceleration of QLoRA

We evaluate the efficiency of QLoRA described in Section 3.3. As shown in Figure 8, the token generation speed increases after using QLoRA.

E Comparison with In Context Learning

We conduct experiments on in context learning (ICL) for ZCL-NER. We use English NER examples and ask the model to do NER task following the given examples.

F Comparison with Data Augmentation

We conduct experiments on data augmentation (DA) for ZCL-NER. We translate the English (EN) train set into the target languages, and use translated train set to train the model for NER task.

The EN train set is directly translated using Qwen2.5-14B-Instruct without fine-tuning. Since T5-base can not directly take non-Latin scripts such as Devanagari and Hanzi as input, we use **mT5-base** (Xue et al., 2021) as an alternative. The model is trained for 50 epoch with a learning rate of $1.0e{-}4$.

	ZH	AR
Full size	61.27	67.29
1k samples	61.20	67.10

Table 14: Performance comparison of our approach with different EACL size. 1k samples are randomly selected from the full corpus.

	EAT	DenKD
Avg. second per iterator	5.85	1.86
Avg. second per token	0.026	0.74

Table 15: Average time comsumption per iterator or token for EAT and DenKD (Ge et al., 2024).

The NER results are evaluated as described in Section 4. As shown in Table 7, the DA approach performs worse than our approach. Furthermore, it also performs worse than the best T-S based approach.

In addition to the poor performance, no matter how many sentences are recognized, it is necessary for the DA approach to translate the full EN train set into one specific language for the NER model training. In other words, even recognizing just one sentence also requires complete translation and training. Translation of the full EN train set and training the NER model require large amounts of computational resources, as well as time consumption. However, the minimum train requirement of our approach is just the English NER model. Our approach is more flexible and efficient than DA approach, and is more practical for real-world applications.

G Fine-tuning Trade-offs

In our preliminary experiments, the loss reaches a stable convergence state after about 5 epochs of fine-tuning. Therefore, we set fine-tuning epochs to 5. To this end, we add further ablation studies on performance with different EACL corpus sizes.

As shown in Table 14, smaller size of EACL corpus leads to the slight drop on performance. This suggests that the data scale and fune-tuning epochs we are currently using are reasonable.

H Inference Cost and Computational Overhead Comparison

We add the average time consumption of a single sentence or token inference for assessment of practical feasibility.

Table 15 indicates that our EAT consumes slightly more time than traditional best-performed method. But this is due to the time-consuming inference process of using LLMs, and our EAT performs much better than DenKD. Following the description in the paper of DenKD, the algorithm complexity of DenKD is $O(n \log n)$ (ignoring MLPs). During training strategy, for each input token, DenKD needs to calculate the Prediction Discrepancy Loss with a double circulation. However, the algorithm complexity of EAT is $O(n)$, as there is no additional calculations for loss.