GENERATIVE-DISCRIMINATIVE KNOWLEDGE DISTILLATION FOR CROSS-LINGUAL NER

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-lingual named entity recognition (NER) aims to build an NER model that generalizes to the low-resource target languages with labeled data from the high-resource source language. Existing approaches can be broadly categorized into two paradigms: discriminative methods based on small language models and generative methods based on large language models. These paradigms are complementary: discriminative models enable fast and efficient inference but may overfit source-language surface features, whereas generative models are generally more robust to unseen languages but incur substantial computational costs during decoding. To combine their complementary strengths, we propose a **Gen**erative-**Dis**criminative **K**nowledge **D**istillation **(GenDisKD)** framework, which transfers knowledge from both discriminative and generative teachers to a compact student model for cross-lingual NER. Furthermore, we introduce a candidate-set distillation loss that preserves alternative candidate labels when teachers disagree, allowing the student to resolve ambiguities through complementary supervision. Extensive experiments on three benchmarks demonstrate that GenDisKD consistently outperforms existing discriminative and generative methods, validating its effectiveness and robustness.

1 INTRODUCTION

Named Entity Recognition (NER) is a fundamental information extraction task that identifies text spans corresponding to predefined semantic categories, such as persons, organizations, and locations Tjong Kim Sang (2002); Pan et al. (2017); Fetahu et al. (2023). NER is useful for various downstream applications, such as information retrieval Zhang et al. (2023), content generation Kang et al. (2026), and complex reasoning Jeong et al. (2025). Although neural NER systems have achieved strong performance with abundant annotated data, their effectiveness remains substantially uneven across languages because high-quality labeled corpora are concentrated in a limited number of high-resource languages. For many low-resource languages, the scarcity of annotated data is further compounded by the high cost and effort required for expert annotation. Cross-lingual NER therefore seeks to address this limitation by transferring supervision from a resource-rich source language to a target language with limited or no labeled data, enabling an NER model to recognize entities across languages.

Existing cross-lingual NER methods can be broadly categorized into discriminative and generative approaches. i) *Discriminative methods* typically transfer knowledge from labeled source-language data to a target-language model through teacher-student knowledge distillation, often exploiting unlabeled target-language data for adaptation Wu et al. (2020); Li et al. (2022); Wu et al. (2021). Recent work further incorporates prototype- or contrastive-based objectives to align cross-lingual entity representations and refine noisy pseudo-labels Huang et al. (2023); Zhou et al. (2023); Ding et al. (2024); Zhu et al. (2025). ii) *Generative methods* instead formulate NER as text generation and leverage multilingual

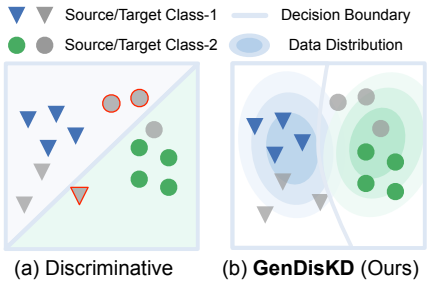

Figure 1: Comparison of supervisions under cross-lingual transfer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

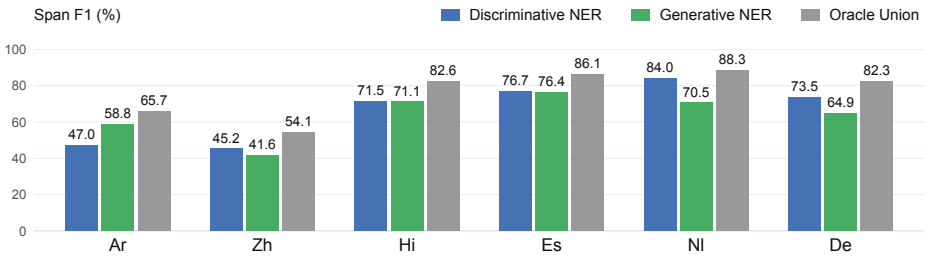

Figure 2: Span-level F1 comparison across six low-resource languages. The Oracle Union, which retains correct predictions from either Discriminative NER or Generative NER under a perfect selector, demonstrates their potential for complementary knowledge distillation.

large language models, such as BLOOM Workshop (2022) and Qwen Yang et al. (2024), either to directly produce entity annotations or to generate synthetic supervision for training smaller task-specific models Sainz et al. (2024); Cocchieri et al. (2025). In summary, discriminative methods offer fast and efficient inference but may overfit source-language surface features (as shown in Figure 1-a), whereas generative methods are generally more robust to unseen languages but incur substantially higher decoding costs.

These observations motivate us to investigate whether generative knowledge can complement discriminative supervision while maintaining an efficient inference process. To this end, we compare a discriminative NER teacher, a generative NER teacher, and their oracle union on six low-resource languages. As shown in Figure 2, the oracle union consistently outperforms both individual teachers, improving over the stronger teacher by 4.3–11.5 span-level F1 points. This result suggests that the two teachers make complementary errors. The discriminative teacher provides explicit classification features, whereas the generative teacher captures complementary features about the reconstruction of entity spans. Combining these signals therefore provides more reliable supervision than relying on either teacher alone. However, directly selecting one teacher's hard prediction when the teachers disagree may propagate specific errors and discard potentially useful alternatives.

Motivated by these findings, we propose **Gen**erative–**Dis**criminative **K**nowledge **D**istillation (**GenDisKD**), as illustrated in Figure 1(b). GenDisKD jointly transfers discriminative classification boundaries and generative entity-distributional knowledge from two teachers to a compact NER student trained on unlabeled target-language data. The generative supervision may be better at avoiding over-fitting because the objective forces them to generation the span in its entirety. GenDisKD further employs candidate-set distillation to avoid premature reliance on potentially erroneous predictions from either teacher. Under teacher disagreement, alternative predictions are retained as candidate labels rather than collapsed into a single learning target.

Our contributions are threefold:

- We propose GenDisKD, a framework that jointly distills generative and discriminative knowledge into a compact NER model for cross-lingual NER transfer.
- We introduce canditate-set distillation to mitigate teacher-specific errors propagation while preserving alternative candidate labels under teacher disagreement.
- Extensive experiments on three benchmarks demonstrate the effectiveness of GenDisKD, with ablation studies further validating the contributions of generative supervision and canditate-set distillation.

2 Related Work

2.1 Cross-Lingual Named Entity Recognition

Cross-lingual named entity recognition (NER) has attracted substantial attention in recent years. Existing approaches can be broadly divided into discriminative methods based on encoder-style small language models and generative methods based on decoder-style large language models. Discrimi-

native methods often adopt teacher-student knowledge distillation to transfer complementary NER knowledge from one or more source-language teachers to a target-language student, while leveraging unlabeled target-language data to improve transfer quality and target-domain adaptation Wu et al. (2020); Li et al. (2022); Wu et al. (2021). Some studies further employ prototype or contrastive-based objectives to align cross-lingual entity representations, while refining noisy pseudo-labels and improving the discriminability of learned entity features Huang et al. (2023); Zhou et al. (2023); Ding et al. (2024); Zhu et al. (2025). Generative approaches instead formulate NER as a language-generation problem and leverage the instruction-following and knowledge-transfer capabilities of large language models (LLMs). One group directly prompts or fine-tunes multilingual LLMs to generate entity annotations. Multilingual decoder-style models such as BLOOM-7B1 Workshop (2022) and the Qwen family Yang et al. (2024) provide strong generative foundations for cross-lingual NER. Another group uses LLMs primarily as data generators or annotators, and then trains a smaller task-specific model on the resulting artificial data. GoLLie Sainz et al. (2024) instruction-tunes an LLM to follow task-specific annotation guidelines, enabling it to produce structured supervision for zero-shot information extraction. Desc Cocchieri et al. (2025) adopts a description-driven framework that combines LLM-generated supervision with entity-type descriptions to support zero-shot NER.

2.2 Generative Classfication

Generative classifiers train a class-conditional generative model to learn $p(x|y)$, and it uses Bayes' rule at inference time to compute $p(y|x)$ for classification. Discriminative classifiers estimate the label distribution conditioned on an input, whereas generative classifiers assign labels through class-conditional input likelihoods. Recent studies have revisited this distinction using autoregressive language models, masked language models, and diffusion models, examining their sample efficiency, calibration, and robustness under distribution shift Kasa et al. (2025). In particular, prior evidence suggests that class-conditional generation can exhibit less dependence on a small set of predictive features in some shortcut-learning settings because classification requires modeling the input under each candidate class Li et al. (2025); Jaini et al. (2024). We treat this behavior as motivation for introducing a complementary generative signal, rather than assuming that generation universally eliminates shortcut reliance. Existing generative-classification studies primarily consider document or image-level prediction and commonly retain the generative model at inference time Zhang et al. (2024; 2025). Cross-lingual NER instead requires labeling many potentially overlapping spans while preserving accurate entity boundaries, making direct class-conditional generation computationally expensive. GenDisKD adapts the generative-classification principle to masked span reconstruction: the generative teacher evaluates the compatibility between a candidate span, its context, and each entity type. These scores provide offline supervision and are not used for autoregressive prediction at deployment.

3 Task Formulation

Cross-lingual named entity recognition (NER) aims to identify and classify entity mentions in a target language using annotations available only in a source language. Formally, let $\mathcal{D}_s = \{(\mathbf{X}_i^s, \mathbf{Y}_i^s)\}_{i=1}^{N_s}$ denote a source-language corpus containing N_s labeled sentences, and let $\mathcal{D}_t = \{\mathbf{X}_i^t\}_{i=1}^{N_t}$ denote an unlabeled corpus in the target language. The model is trained using both corpora without access to target-language annotations.

Following prior work Zhu et al. (2025), we formulate NER as span classification. Given a sentence $\mathbf{X} = (x_1, \ldots, x_n)$, we enumerate contiguous spans up to a maximum length L:

$$\mathcal{T}(\mathbf{X}) = \{s_{j:k} = (x_j, \ldots, x_k) \mid 1 \le j \le k \le n,\ k - j + 1 \le L\}.$$

Each candidate span is assigned a label from $\mathcal{C} = \{0, 1, \ldots, K-1\}$, where labels $1, \ldots, K-1$ represent entity types and label 0 denotes a non-entity span. For source-language sentences, the span label $y_{j:k}$ is derived from the gold annotations: a span receives its entity type if it exactly matches an annotated mention, and 0 otherwise. The learning objective is to predict these span labels for target-language sentences.

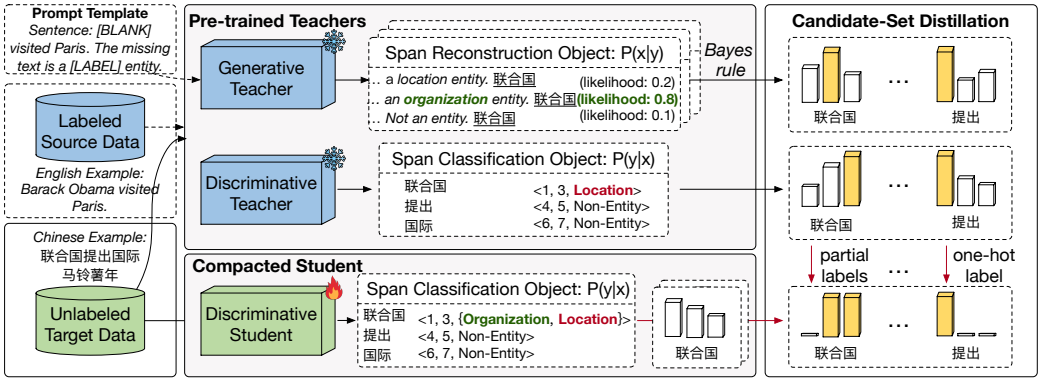

Figure 3: Overview of the proposed **GenDisKD** framework. The generative and discriminative teachers provide complementary supervision through candidate-set distillation to train a compact student model on unlabeled target-language data.

4 GENDISKD

In this section, we introduce the proposed **Gen**erative-**Dis**criminative **K**nowledge **D**istillation **(GenDisKD)**. Figure 3 illustrates the framework, GenDisKD first pre-trains a discriminative teacher that directly predicts span labels and a generative teacher that evaluates label-conditioned span likelihoods. Their predictions are then converted into canditate-set labels, allowing a compact student to learn from teacher agreement without forcing a unique pseudo-label under disagreement.

4.1 PRE-TRAINED NER TEACHERS

We pre-train two teachers on $\mathcal{D}_s$ using different modeling objectives. The discriminative teacher learns to distinguish entity type labels, whereas the generative teacher learns to reconstruct span text conditioned on its entity label and surrounding context.

Discriminative NER Teacher The discriminative teacher estimates $p_{\mathrm{d}}(y \mid \mathbf{X}, s_{j:k})$ using contextual and span-level features. A pre-trained multilingual encoder E_{d} first produces contextual token representations

$$\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n) = E_{\mathrm{d}}(\mathbf{X}).$$

For each candidate span, we concatenate its boundary representations with a learnable length embedding $\boldsymbol{\ell}_{j:k}$ and a morphology embedding $\mathbf{m}_{j:k}$. The latter aggregates token-level features within the span, including capitalization and numeric patterns:

$$\mathbf{e}_{j:k} = [\mathbf{h}_j; \mathbf{h}_k; \boldsymbol{\ell}_{j:k}; \mathbf{m}_{j:k}] \,.$$

The span representation is fed into an MLP classification head followed by a softmax function:

$$p_{\mathrm{d}}(y \mid \mathbf{X}, s_{j:k}) = \mathrm{softmax}\left(f_{\mathrm{cls}}(\mathbf{e}_{j:k})\right)_y \,. \tag{1}$$

We train the teacher using cross-entropy over source-language candidate spans:

$$\mathcal{L}_{\mathrm{DIS}} = -\sum_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}_s} \sum_{s_{j:k}\in\mathcal{T}(\mathbf{X})} \log p_{\mathrm{d}}(y_{j:k} \mid \mathbf{X}, s_{j:k}). \tag{2}$$

During inference, the discriminative teacher predicts the label of each candidate span by selecting the label with the highest posterior probability:

$$\hat{y}^{\mathrm{d}}_{j:k} = \arg\max_{y\in\mathcal{C}} p_{\mathrm{d}}(y \mid \mathbf{X}, s_{j:k}). \tag{3}$$

Bayes-based Generative NER Teacher In contrast to discriminative classifiers, which tends to minimize their training objectives using just a few spurious features, generative classifiers need to model the entire span input $s_{j:k}$. This means that they cannot stop at just the spurious features, their span reconstruction objective requires them to learn both core and spurious features.

For each candidate span $s_{j:k}$ in a sentence $\mathbf{X}$, we replace the span with a special token *[BLANK]* and construct a label-conditioned prompt using the corresponding entity-type description. Let $\widetilde{\mathbf{X}}_{j:k}$ denote the resulting masked context. Given a candidate label $y \in \mathcal{C}$, an autoregressive Transformer reconstructs the original span token by token.

Specifically, the generative teacher is trained with teacher-forced cross-entropy:

$$\mathcal{L}_{\mathrm{GEN}} = -\sum_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}_s}\sum_{s_{j:k}\in\mathcal{T}(\mathbf{X})} \log p_g\left(s_{j:k} \mid \widetilde{\mathbf{X}}_{j:k}, y_{j:k}\right), \tag{4}$$

At inference time, given a candidate span $s_{j:k}$, the generative teacher evaluates every possible label $y \in \mathcal{C}$. Specifically, it performs $|\mathcal{C}|$ forward passes, for each label, the model obtains the corresponding conditional span log-likelihood:

$$\ell_{\mathrm{g}}(y \mid \mathbf{X}, s_{j:k}) = \log p_g\left(s_{j:k} \mid \widetilde{\mathbf{X}}_{j:k}, y\right). \tag{5}$$

We assume a uniform prior over candidate labels, under which the posterior is proportional to the label-conditioned span likelihood. We then use Bayes' rule to convert the label-conditioned span likelihoods into a posterior label distribution:

$$p_{\mathrm{g}}(y \mid \mathbf{X}, s_{j:k}) = \frac{p_g\left(s_{j:k} \mid \widetilde{\mathbf{X}}_{j:k}, y\right)}{\sum_{y'\in\mathcal{C}} p_g\left(s_{j:k} \mid \widetilde{\mathbf{X}}_{j:k}, y'\right)}. \tag{6}$$

The generative teacher therefore predicts the label with the highest posterior probability:

$$\hat{y}^{\mathrm{g}}_{j:k} = \arg\max_{y\in\mathcal{C}} p_{\mathrm{g}}(y \mid \mathbf{X}, s_{j:k}). \tag{7}$$

4.2 Compacted NER Student

After the two teachers have been trained, they independently predict labels for candidate spans in unlabeled target-language sentences. These predictions are then used to construct supervision for training a compact student model.

Candidate-Set Knowledge Distillation When the two teachers make different predictions for the same candidate span, there is no target-language gold annotation to determine which prediction is correct. Selecting either prediction as a hard pseudo-label may therefore force the student to learn an incorrect label and discard the possibility that the other prediction is correct. To address this issue, we propose Candidate-Set Knowledge Distillation, which preserves the predictions of both teachers as a set of candidate labels. Specifically, for each candidate span $s_{j:k}$ in a target-language sentence $\mathbf{X} \in \mathcal{D}_t$, the discriminative and generative teachers predict:

$$\hat{y}^{\mathrm{d}}_{j:k} = \arg\max_{y\in\mathcal{C}} p_{\mathrm{d}}(y \mid \mathbf{X}, s_{j:k}), \qquad \hat{y}^{\mathrm{g}}_{j:k} = \arg\max_{y\in\mathcal{C}} p_{\mathrm{g}}(y \mid \mathbf{X}, s_{j:k}). \tag{8}$$

We construct the candidate label set by taking the union of the two predictions:

$$\mathcal{A}_{j:k} = \left\{\hat{y}^{\mathrm{d}}_{j:k}, \hat{y}^{\mathrm{g}}_{j:k}\right\}. \tag{9}$$

When the teachers agree, $\mathcal{A}_{j:k}$ contains a single label and provides one-hot supervision. When they disagree, $\mathcal{A}_{j:k}$ contains two labels and provides partial-label supervision.

Let $p_{\mathrm{s}}(y \mid \mathbf{X}, s_{j:k}; \Theta_{\mathrm{s}})$ denote the compact student's probability distribution over the label set $\mathcal{C}$. We train the student by maximizing the total probability assigned to the candidate label set:

$$\mathcal{L}_{\mathrm{CSKD}} = -\sum_{\mathbf{X}\in\mathcal{D}_t}\sum_{s_{j:k}\in\mathcal{T}(\mathbf{X})} \log\left(\sum_{y\in\mathcal{A}_{j:k}} p_{\mathrm{s}}\left(y \mid \mathbf{X}, s_{j:k}; \Theta_{\mathrm{s}}\right)\right). \tag{10}$$

Methods \ Datasets	CoNLL				WikiAnn			
	de	es	nl	Avg	ar	hi	zh	Avg
PRAM Huang et al. (2023)	77.64	82.06	83.15	80.95	57.44	74.67	55.46	62.52
ContProto Zhou et al. (2023)	76.41	85.02	83.69	81.71	72.20	83.45	61.47	72.37
GLoDe Ding et al. (2024)	79.15	85.29	85.76	83.40	74.35	83.61	64.81	74.26
MARAL Zhu et al. (2025)	80.28	86.12	85.92	84.11	**78.11**	84.46	66.26	76.28
Qwen2.5-14B Yang et al. (2024)	48.45	53.15	43.61	48.40	31.28	43.13	27.39	33.93
GPT4o-Zero OpenAI (2023)	22.14	50.15	33.08	35.12	37.14	56.43	46.37	46.65
GPT4o-Few OpenAI (2023)	24.02	52.66	35.21	37.30	39.80	56.90	46.56	47.75
Bloom-7B1 Workshop (2022)	54.40	73.46	68.39	65.42	55.11	66.95	38.48	53.48
GenDisKD (Ours)	**81.70**	**86.88**	**86.03**	**84.87**	77.14	**85.08**	**66.88**	**76.37**

Table 1: Performance comparison on the CoNLL and WikiAnn.

Methods \ Datasets	MultiCoNER							
	bn	fa	fr	it	pt	sv	uk	Avg
PRAM Huang et al. (2023)	57.93	61.44	70.71	77.16	73.54	73.47	66.63	68.70
ContProto Zhou et al. (2023)	75.09	68.74	75.16	83.01	81.34	79.73	75.57	76.95
GLoDe Ding et al. (2024)	76.03	69.55	76.27	82.31	80.83	79.14	74.83	76.99
MARAL Zhu et al. (2025)	78.21	71.28	78.10	84.49	**83.02**	81.23	76.85	79.03
Qwen2.5-14B Yang et al. (2024)	34.36	46.06	35.99	34.11	52.31	37.01	45.86	40.81
GPT4o-Zero OpenAI (2023)	43.02	56.23	67.55	69.52	38.68	69.33	73.85	59.74
GPT4o-Few OpenAI (2023)	43.78	56.40	68.63	70.98	38.94	70.49	74.07	60.47
Bloom-7B1 Workshop (2022)	46.44	54.42	63.62	68.05	68.46	67.71	61.25	61.42
GenDisKD (Ours)	**78.30**	**71.85**	**78.89**	**84.75**	82.75	**82.33**	**77.18**	**79.44**

Table 2: Performance comparison on MultiCoNER.

Student Inference For inference in the target language, we only use the distilled student model to predict the probability distribution over entity labels for each candidate span $s_{j:k}$ in a test sentence $\mathbf{X}$. We then select the label $y \in \mathcal{C}$ with the highest probability as the predicted label $\hat{y}_{j:k}$:

$$\hat{y}_{j:k} = \arg\max_{y \in \mathcal{C}} p_{\mathrm{s}}\left(y \mid \mathbf{X}, s_{j:k}; \Theta_{\mathrm{s}}\right). \tag{11}$$

Here, $p_{\mathrm{s}}(y \mid \mathbf{X}, s_{j:k}; \Theta_{\mathrm{s}})$ denotes the probability assigned by the student model to label y.

5 Experiments

5.1 Setup

Datasets We conduct experiments on three typical cross-lingual NER benchmarks. **(1) CoNLL** Tjong Kim Sang (2002); Tjong Kim Sang & De Meulder (2003) covers English, German, Spanish, and Dutch, and defines four entity types: ORG, PER, LOC, and MISC. **(2) WikiAnn** Pan et al. (2017) includes English, Arabic, Hindi, and Chinese, with three entity types: ORG, PER, and LOC. **(3) MultiCoNER** Fetahu et al. (2023) is a more challenging benchmark covering English, Bengali, Farsi, French, Italian, Portuguese, Swedish, and Ukrainian, with six entity types: PER, LOC, GRP, PROD, CW, and MED. For all datasets, English serves as the source language, while the remaining languages are treated as target languages. We use gold annotations only for the English source-language data and treat the target-language training data as unlabeled during training.

Implementation Details We use XLM-R-Large Conneau et al. (2020) as the discriminative teacher (or student) backbone and BLOOM-7B1 Workshop (2022) as the generative Bayesian

teacher backbone. We set the maximum input length to 128 tokens and enumerate spans up to length four. The discriminative model is trained with AdamW using a learning rate of 2×10^{-5}, a weight decay of 0.01, a batch size of 16, and training epochs of 20. For the generative teacher, we use a learning rate of 1×10^{-5}, and a batch size of 4, and training epochs of 8. The generative model is adapted with LoRA adapters of rank 8, scaling factor 16, and dropout 0.05, while the pre-trained BLOOM-7B1 parameters remain frozen. We compute the conditional span score using the mean token-level log-likelihood. All experiments are conducted on a single NVIDIA RTX PRO 6000 GPU per run. To reduce the influence of random initialization and data sampling, each experiment is repeated three times with different random seeds, and we report the average results.

Baselines **(a) Discriminative:** (1) PRAM Huang et al. (2023) uses prototypes as cross-lingual bridges and an attribution consistency objective to align entity representations with predictions. (2) ConProto Zhou et al. (2023) combines contrastive self-training with prototype-based pseudo-labeling to learn class-discriminative for cross-lingual NER. (3) Glode Ding et al. (2024) progressively refines noisy pseudo-labels using global and local semantic distributions. (4) MARAL Zhu et al. (2025) combines adaptively reweighted contrastive learning with progressive cross-lingual adaptation to address class imbalance and pseudo-label bias. **(b) Generative:** (5) Bloom-7B1 Workshop (2022) is a 7B-parameter open-source multilingual large language model. We fine-tune it to the NER task using LoRA while keeping its original parameters frozen. (6) Qwen2.5-14B-Instruct Yang et al. (2024) is an open-source multilingual instruction-tuned causal language model with 14B parameters. We prompt it with NER task-specific instructions to produce entity annotations. (7) GPT4o-Zero/Few OpenAI (2023) are used as closed-source generative baselines, where GPT-4o-Zero performs NER without demonstrations, while GPT-4o-Few uses labeled source-language examples as in-context demonstrations.

5.2 Main Results

Table 1 and Table 2 report the cross-lingual NER results of GenDisKD on the CoNLL, WikiAnn, and MultiCoNER datasets, alongside those of prior strong methods and several re-implemented baselines, including BLOOM-7B1 and Qwen2.5-14B-Instruct. GenDisKD achieves the best average F1-score on all three datasets, obtaining 84.87 on CoNLL, 76.37 on WikiAnn, and 79.44 on MultiCoNER. Compared with the strongest prior method, MARAL, GenDisKD improves the average F1-score by 0.76, 0.09, and 0.41 points on CoNLL, WikiAnn, and MultiCoNER, respectively. Moreover, GenDisKD consistently outperforms BLOOM-7B1, with gains of 19.45, 22.89, and 18.02 F1 points on the three datasets, respectively. At the language level, GenDisKD achieves the best performance on all three CoNLL languages, two of the three WikiAnn languages (Hindi and Chinese), and six of the seven MultiCoNER languages (Bengali, Farsi, French, Italian, Swedish, and Ukrainian). These results demonstrate the effectiveness and robustness of GenDisKD for zero-resource cross-lingual NER across diverse datasets and languages.

5.3 Ablation Study

To demonstrate the effectiveness of our approach, we designed the following ablation studies. Table 3 presents the results.

w/o Generative NER Teacher, which removes the generative teacher and leaves the student without supervision from entity-generation distributions, reduces the average F1-score to 82.29 on CoNLL and 69.97 on WikiAnn, corresponding to drops of 2.58 and 6.40 points, respectively. This degradation suggests that the generative teacher provides complementary information about entity characteristics that discriminative supervision alone may overlook. Such information can help the student rely less on source-language-specific classification cues and generalize better to target languages.

w/o Discriminative NER Teacher, which removes the discriminative teacher, produces the largest performance degradation, lowering the average F1-score to 71.06 on CoNLL and 57.14 on WikiAnn, corresponding to drops of 13.81 and 19.23 points, respectively. This result indicates that discriminative supervision provides an essential foundation for accurate entity classification. Although generative distributions offer useful complementary signals, they may not sufficiently distinguish competing entity types and non-entity spans on their own. The substantial decline therefore suggests that generative knowledge is most effective when combined with explicit discriminative guidance.

Methods \ Datasets	CoNLL				WikiAnn			
	de	es	nl	Avg	ar	hi	zh	Avg
GenDisKD (Ours)	**81.70**	**86.88**	**86.03**	**84.87**	**77.14**	**85.08**	**66.88**	**76.37**
w/o Gen. NER Teacher	77.52	84.94	84.39	82.29	67.17	81.71	61.05	69.97
w/o Dis. NER Teacher	64.94	77.34	70.90	71.06	58.77	71.07	41.59	57.14
w/o Knowledge Distillation	76.59	82.80	82.13	80.51	61.77	79.61	59.67	67.02

Table 3: Ablations on the CoNLL and WikiAnn datasets.

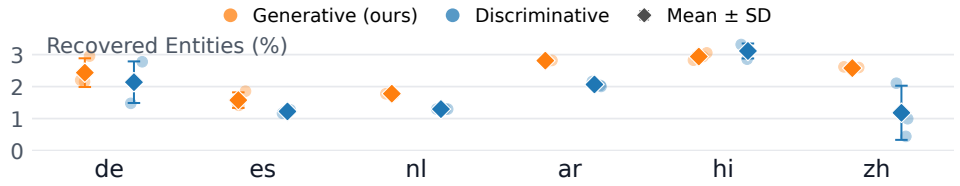

Figure 4: Recovered entities comparison of different auxiliary teachers.

w/o Knowledge Distillation, which disables knowledge distillation, reduces the average F1-score to 80.51 on CoNLL and 67.02 on WikiAnn, falling below the full model by 4.36 and 9.35 points, respectively. This result highlights the importance of transferring teacher knowledge to the compact student. Distillation exposes the student to teacher-derived supervision on unlabeled target-language data, helping it generalize beyond the source language. The larger decline on WikiAnn suggests that this transfer contributes more strongly to performance on its target languages.

5.4 Further Analysis

We aim to further analyze GenDisKD by addressing the following questions: (1) Do generative teachers provide greater complementary supervision beyond discriminative teachers? (see § 5.4.1) (2) What benefits does the canditate-set distillation provide for GenDisKD? (see § 5.4.2) (3) What trade-off does GenDisKD achieve between inference efficiency and performance? (see § 5.4.3)

5.4.1 Complementary Supervision from Different Teachers

Our intuition is that label-conditioned generation provides complementary evidence for entity recognition beyond that obtained from an additional discriminative prediction. We test this by fixing a discriminative prediction as an anchor and measuring how many gold entities missed by this anchor are recovered by an additional generative or discriminative prediction. We normalize both counts by the total number of gold entities, using the same anchor for both comparisons. Figure 4 shows that the generative predictions provide greater complementary coverage in five of the six target languages, with particularly pronounced gains in Dutch, Arabic, and Chinese. In Chinese, the generative predictions recover approximately 2.6% of all gold entities beyond the anchor, compared with 1.0% for the additional discriminative predictions. The contributions are similar in German, whereas the additional discriminative predictions provide greater coverage in Hindi. These results suggest that incorporating a generative view can recover entities that are less readily recovered by another discriminative prediction, supporting its use as a complementary source of supervision. This advantage is language-dependent, however, and additional coverage alone does not establish the precision of the added predictions or whether their information is successfully transferred to the student.

5.4.2 Impacts of Canditate-Set Distillation Loss

We investigate the contribution of the introduced candidate-set distillation loss in our framework. To isolate its effect, we keep the teacher predictions and training setup fixed, and evaluate the student models on instances where the two teachers produce inconsistent predictions. We compare three loss functions: (i) Our canditate-set distillation loss, which applies the partial loss to teacher-

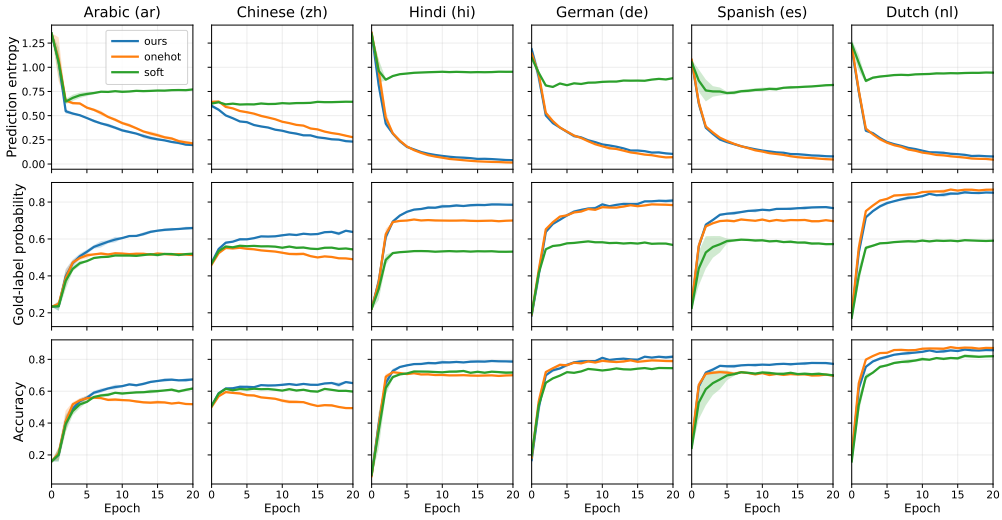

Figure 5: Analysis results of the canditate-set knowledge distillation loss.

disagreement cases and the one-hot loss when the teachers agree. (ii) One-hot labels loss, which uses one-hot supervision for all cases. (iii) Soft labels loss, which applies soft supervision throughout. As shown in Figure 5, the mixed objective generally achieves lower prediction entropy and higher gold-label probabilities than the soft objective, while yielding the best final accuracy in Arabic, Chinese, Hindi, German, and Spanish. Compared with the one-hot variant, mixed further improves accuracy in most languages, although one-hot performs slightly better in Dutch. These controlled comparisons indicate that the canditate-set distillation loss makes a substantive contribution to our framework, especially for examples on which the teachers disagree, by improving the student's ability to learn from non-identical teacher predictions without relying exclusively on hard targets.

5.4.3 Inference Efficiency

We compare the inference efficiency and performance of GenDisKD with discriminative (XLM-R-Large) and generative (BLOOM-7B1) approaches on CoNLL. As shown in Table 4, GenDisKD uses only 0.35B parameters and requires 243 seconds of inference time, while achieving an F1 score of 84.87. Compared with the generative approach, which uses 7B parameters and requires 1,566 seconds, GenDisKD reduces the model size by 20× and inference time by 1,323 seconds (84.5%), while improving the F1 score by 19.45 points. Compared with the discriminative approach, GenDisKD has the same parameter size but requires 21 additional seconds of inference time, while achieving a 4.36-point improvement in F1. These results demonstrate that GenDisKD provides a favorable trade-off between inference efficiency and task performance.

Method	**Size**	**Time**	**F1**
Discriminative	0.35B	222s	80.51
Generative	7B	1,566s	65.42
GenDisKD	0.35B	243s	84.87

Table 4: Inference Efficiency Comparison.

6 Conclusion

This work presents GenDisKD, a generative–discriminative knowledge distillation framework for cross-lingual named entity recognition (NER). To address teacher disagreement, GenDisKD avoids forcing a single pseudo-label and instead preserves the teachers' predictions as partial supervision, while applying one-hot supervision when the teachers agree. Experiments across multiple target languages demonstrate that the generative perspective can identify entities missed by the discriminative teacher and that disagreement-aware distillation improves the student's performance by preserving complementary teacher predictions under disagreement. At inference time, the student requires only a single discriminative model, making GenDisKD suitable for deployment scenarios in which target-language annotations are unavailable.

AI USE STATEMENT

In this work, large language models (LLMs) were used solely during the final stages of the writing process, specifically for tasks such as proofreading and improving linguistic clarity. Their use was limited to enhancing the readability, fluency, and grammatical correctness of the manuscript, ensuring the effective communication of our ideas. Crucially, LLMs were not involved in any core aspects of the research itself, including the formulation of methodology, experimental design, or result interpretation. We take full responsibility for the content presented in this paper.

REPRODUCIBILITY STATEMENT

We provide a comprehensive description of the proposed method in section 4, with detailed implementation and datasets specifics provided in section 5.1 and A.1.

REFERENCES

Alessio Cocchieri, Marcos Martínez Galindo, Giacomo Frisoni, Gianluca Moro, Claudio Sartori, and Giuseppe Tagliavini. ZeroNER: Fueling zero-shot named entity recognition via entity type descriptions. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15594–15616, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.805. URL https://aclanthology.org/2025.findings-acl.805/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Zhuojun Ding, Wei Wei, Xiaoye Qu, and Dangyang Chen. Improving pseudo labels with global-local denoising framework for cross-lingual named entity recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/691. URL https://doi.org/10.24963/ijcai.2024/691.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2027–2051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.134. URL https://aclanthology.org/2023.findings-emnlp.134/.

Yucheng Huang, Wenqiang Liu, Xianli Zhang, Jun Lang, Tieliang Gong, and Chen Li. PRAM: An end-to-end prototype-based representation alignment model for zero-resource cross-lingual named entity recognition. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3220–3233, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.201. URL https://aclanthology.org/2023.findings-acl.201/.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=rmg0qMKYRQ.

Hongjun Jeong, Minji Kim, Heesoo Jung, Ko Keun Kim, and Hogun Park. Enhancing complex reasoning in knowledge graph question answering through query graph approximation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.),

Findings of the Association for Computational Linguistics: ACL 2025, pp. 27038–27056, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1387. URL https://aclanthology.org/2025.findings-acl.1387/.

Myunghoon Kang, Dahyun Jung, Suhyune Son, Seonmin Koo, Changwoo Chun, Daniel Rim, Haeyoung Kwon, Yuna Hur, and Heuiseok Lim. EASE: Entity-aware sub-table generation for real-world multi-table QA. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 277–302, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.10. URL https://aclanthology.org/2026.acl-long.10/.

Siva Rajesh Kasa, Karan Gupta, Sumegh Roychowdhury, Ashutosh Kumar, Yaswanth Biruduraju, Santhosh Kumar Kasa, Nikhil Priyatam Pattisapu, Arindam Bhattacharya, Shailendra Agarwal, and Vijay Huddar. Generative or discriminative? revisiting text classification in the era of transformers. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 9604–9626. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.486. URL https://doi.org/10.18653/v1/2025.emnlp-main.486.

Alexander Cong Li, Ananya Kumar, and Deepak Pathak. Generative classifiers avoid shortcut solutions. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=oCUYc7BzXQ.

Zhuoran Li, Chunming Hu, Xiaohui Guo, Junfan Chen, Wenyi Qin, and Richong Zhang. An unsupervised multiple-task and multiple-teacher model for cross-lingual named entity recognition. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 170–179, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.14. URL https://aclanthology.org/2022.acl-long.14/.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL https://aclanthology.org/P17-1178/.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Y3wpuxd7u9.

Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. URL https://aclanthology.org/W02-2024/.

Erik F. 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*, pp. 142–147, 2003. URL https://aclanthology.org/W03-0419/.

BigScience Workshop. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/ARXIV.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Jian-Guang Lou, and Biqing Huang. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6505–6514, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.581. URL https://aclanthology.org/2020.acl-main.581/.

Qianhui Wu, Zijia Lin, Börje F. Karlsson, Biqing Huang, and Jian-Guang Lou. Unitrans: unifying model transfer and data transfer for cross-lingual named entity recognition with unlabeled data. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Fengda Zhang, Qianpei He, Kun Kuang, Jiashuo Liu, Long Chen, Chao Wu, Jun Xiao, and Hanwang Zhang. Distributionally generative augmentation for fair facial attribute classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22797–22808. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02151. URL https://doi.org/10.1109/CVPR52733.2024.02151.

Shizhou Zhang, Xueqiang Lv, Yinghui Xing, Qirui Wu, Di Xu, and Yanning Zhang. Revisiting generative replay for class incremental object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20340–20349. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01894. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Revisiting_Generative_Replay_for_Class_Incremental_Object_Detection_CVPR_2025_paper.html.

Zhongping Zhang, Yiwen Gu, and Bryan Plummer. Show, write, and retrieve: Entity-aware article generation and retrieval. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8684–8704, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.581. URL https://aclanthology.org/2023.findings-emnlp.581/.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, and Chunyan Miao. Improving self-training for cross-lingual named entity recognition with contrastive and prototype learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4018–4031, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.222. URL https://aclanthology.org/2023.acl-long.222/.

Guangcheng Zhu, Ruixuan Xiao, Haobo Wang, Zhen Zhu, Gengyu Lyu, and Junbo Zhao. Large margin representation learning for robust cross-lingual named entity recognition. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4270–4291, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.215. URL https://aclanthology.org/2025.acl-long.215/.

Lang	Type	Train	Dev	Test
English en	Sentence	14,987	3,466	3,684
	Entity	23,499	5,942	5,648
German de	Sentence	12,705	3,068	3,160
	Entity	11,851	4,833	3,673
Spanish es	Sentence	8,323	1,915	1,517
	Entity	18,798	4,351	3,558
Dutch nl	Sentence	15,806	2,895	5,195
	Entity	13,344	2,616	3,941

Table 5: Statistics of CoNLL-2003.

Lang	Type	Train	Dev	Test
English en	Sentence	20,000	10,000	10,000
	Entity	27,931	14,146	13,958
Arabic ar	Sentence	20,000	10,000	10,000
	Entity	22,500	11,266	11,259
Hindi hi	Sentence	5,000	1,000	1,000
	Entity	6,124	1,226	1,228
Chinese zh	Sentence	20,000	10,000	10,000
	Entity	25,031	12,493	12,532

Table 6: Statistics of WikiAnn.

Lang	Type	Train	Dev	Test
Bangla bn	Sentence	9,708	507	19,859
	Entity	13,223	676	25,013
Farsi fa	Sentence	16,321	855	219,168
	Entity	23,662	1,236	312,115
French fr	Sentence	16,548	857	249,786
	Entity	26,377	1,352	398,197
Italian it	Sentence	16,579	858	247,881
	Entity	26,443	1,398	397,222
Portuguese pt	Sentence	16,469	854	229,490
	Entity	24,439	1,292	340,753
Swedish sv	Sentence	16,363	856	231,190
	Entity	25,414	1,391	361,159
Ukrainian uk	Sentence	16,429	851	238,296
	Entity	21,956	1,135	315,374

Table 7: Statistics of MultiCoNER.

A Appendix

A.1 Datasets Statistic

Table 5, 6 and 7 shows the statistics of all three datasets. CoNLL-2003 Tjong Kim Sang (2002); Tjong Kim Sang & De Meulder (2003) is a widely used multilingual NER benchmark containing newswire data in English, German, Spanish, and Dutch. We use English as the source language and German, Spanish, and Dutch as target languages. The dataset defines four entity categories: person (PER), organization (ORG), location (LOC), and miscellaneous (MISC), in addition to the outside label O.

WikiAnn Pan et al. (2017) is a multilingual NER benchmark constructed from Wikipedia and covering a broad range of languages. In our cross-lingual setting, we use English as the source language and Arabic, Hindi, and Chinese as target languages. The label space contains three entity categories: person (PER), organization (ORG), and location (LOC), together with the outside label O.

MultiCoNER Fetahu et al. (2023) is a fine-grained multilingual named entity recognition benchmark designed to evaluate recognition of complex and long-tail entities. We use the seven target languages available in our cross-lingual setting: Bangla, Farsi, French, Italian, Portuguese, Swedish, and Ukrainian. The label space contains six entity categories, namely person (PER), location (LOC), group (GRP), product (PROD), creative work (CW), and medical entity (MED), in addition to the outside label O.

Table 8: Candidate span filtering analysis on CoNLL. Dis. Pseudo-label F1 denotes the span-level F1 obtained from the discriminative teacher's hard predictions. Oracle F1 denotes the maximum recoverable span-level F1 when predictions are restricted to the candidate spans selected by the discriminative teacher, using exact span boundaries.

Metric	Spanish	Dutch	German	Average
Dis. Pseudo-label F1	76.59	82.80	82.13	80.51
Oracle F1	91.30	93.57	88.51	**91.13**

Methods	Runs	CoNLL				WikiAnn			
		de	es	nl	Avg	ar	hi	zh	Avg
MARAL	Run 1	81.00	85.63	84.79	83.81	76.14	84.84	66.06	75.68
	Run 2	81.17	85.68	84.84	83.90	76.72	84.35	65.86	75.64
	Run 3	81.11	85.84	84.79	83.91	76.62	84.61	65.98	75.74
	Mean	81.09	85.72	84.81	83.87	76.49	84.60	65.97	75.69
	Std.	(0.08)	(0.11)	(0.03)	(0.06)	(0.31)	(0.25)	(0.10)	(0.05)
GenDisKD	Run 1	81.66	87.07	85.79	84.84	77.14	84.84	66.80	76.26
	Run 2	81.65	86.74	86.49	84.96	76.72	85.31	66.86	76.30
	Run 3	81.80	86.84	85.80	84.81	77.55	85.11	66.98	76.55
	Mean	**81.70**	**86.88**	**86.03**	**84.87**	**77.14**	**85.08**	**66.88**	**76.37**
	Std.	(0.08)	(0.17)	(0.40)	(0.08)	(0.41)	(0.23)	(0.09)	(0.16)

Table 9: Comparison between GenDisKD and the re-implemented MARAL baseline on the CoNLL and WikiAnn datasets. Each run reports the language-level F1-score, while the Mean and Std. rows report the mean and sample standard deviation over three runs, respectively. Values in parentheses denote standard deviations.

A.2 Instruction for LLM Baseline

Following prior studies Zhu et al. (2025), we design the instruction prompt for Qwen2.5-14B-Instruct as follows: *Task Description: You are working as a named entity recognition expert and your task is to identify any named entities present inthe text. The named entity types include six categories: PER (person), LOC (location), GRP (group), PROD(product), CW (creative work), and MED (medical). Specifically, you will be given a sentence and you shouldoutput a list of tuples, where each tuple consists of a span from the input text and its corresponding namedentity type. Ensure that you reply with briet, to-the-point answers with no elaboration as truthfully as possible.Moreover, before giving an answer, you need to reflect on whether ceach named entity is recognized correctly(check one by one and think step by step without additional explanations).*

A.3 Candidate Span Filtering in Generative Teacher

Enumerating all possible spans yields $O(nL)$ candidate spans for a sentence of length n with a maximum span length of L. If the generative teacher evaluates every candidate under all labels in $\mathcal{C}$, the resulting computational cost is $O(nL|\mathcal{C}|)$. This cost can be particularly substantial for long sentences or datasets with a large number of non-entity spans.

To evaluate whether this filtering step discards potentially useful entity spans, we measure the oracle span-level F1 score of the retained candidate set. Specifically, the oracle score represents the maximum achievable entity-span F1 when predictions are restricted to the spans selected by the discriminative teacher. Thus, it measures candidate coverage rather than the actual performance of the generative teacher. As shown in Table 8, the discriminative teacher achieves an average pseudo-label F1 of 80.51%, whereas the retained candidate spans yield an average strict oracle F1 of 91.13%. The substantial gap between these two values indicates that the discriminative teacher's predicted boundaries contain considerably more potentially useful entity spans than its hard predictions alone would suggest. Therefore, although the generative teacher operates only on the filtered candidates, it still

has ample opportunity to provide complementary supervision by re-evaluating these spans under alternative entity labels.

A.4 Implementation Detail of the GenerativeTeacher

The generative teacher uses a class-conditional span-reconstruction objective. For a candidate span s in a sentence, we first normalize the sentence, optionally crop a span s based on the span filtering results, and replace the span with the literal token `[BLANK]`. Under our designed prompt, the resulting causal-LM input is:

$$\begin{aligned} \texttt{Sentence:} \quad & \mathcal{C}_{\setminus s} \\ \texttt{The missing text is a} \quad & d(y)\texttt{.} \\ \texttt{Missing text:} \quad & \end{aligned}$$

where $\mathcal{C}_{\setminus s}$ denotes the context with span s replaced by `[BLANK]`, y is a candidate label, and $d(y)$ is its natural-language description (e.g., `organization name`, `person name`, `location name`, or `non-entity phrase` for `O`). The model is trained to generate only the original span after the final prompt prefix, prompt tokens are excluded from the token-level negative log-likelihood by a target mask.

A.5 Stability of Results

Table 9 compares GenDisKD with MARAL, which we re-implemented using the publicly available source code under the same experimental environment [1]. Each method was run three times with different random seeds. GenDisKD achieves higher mean F1-scores than MARAL on all six target languages, although the improvement on Hindi is relatively small. On average, GenDisKD improves over MARAL by 1.00 F1 point on CoNLL and by 0.68 F1 points on WikiAnn. The largest gains are observed on Arabic ($+0.65$ F1 points), Dutch ($+1.22$ F1 points), and Spanish ($+1.17$ F1 points). Across all runs, GenDisKD yields an average improvement of 0.84 F1 points over MARAL, and a paired t-test indicates that this overall improvement is statistically significant ($t = 8.52$, $p < 0.001$). These results demonstrate that GenDisKD provides consistent performance gains over a strong re-implemented baseline under the same experimental setting.

[1] https://github.com/gczhu/MARAL