# Bridging the Gap: Efficient Cross-Lingual NER in Low-Resource Financial Domain

**Sunisth Kumar[1], Mohammed ElKholy[1], Davide Liu[1], Alexandre Boulenger[1],**

[1]Genify,

ksunisth@gmail.com, mo.u.kholy@gmail.com, davide@genify.ai, alex@genify.ai

## Abstract

We present an innovative and efficient modeling framework for cross-lingual named entity recognition (NER), leveraging the strengths of knowledge distillation and consistency training. Our approach distills knowledge from an XLM-RoBERTa model pre-trained on a high-resource source language (English) to a student model, which then undergoes semi-supervised consistency training with KL divergence loss on a low-resource target language (Arabic). We focus our application on the financial domain, using a small, sourced dataset of financial transactions as seen in SMS messages.

Using datasets comprising SMS messages in English and Arabic containing financial transaction information, we aim to transfer NER capabilities from English to Arabic with minimal labeled Arabic samples. The framework generalizes named entity recognition from English to Arabic, achieving F1 scores of 0.74 on the Arabic financial transaction dataset and 0.61 on the WikiANN dataset, surpassing or closely competing with models that have 1.7× and 5.3× more parameters, respectively, while efficiently training it on a single T4 GPU.

Our experiments show that using a small number of labeled data for low-resource cross-lingual NER applications is a wiser choice than utilizing zero-shot techniques while also using up fewer resources. This framework holds significant potential for developing multilingual applications, particularly in regions where digital interactions span English and low-resource languages.

## 1 Introduction

Named Entity Recognition (NER) has become pivotal in Natural Language Processing (NLP) within finance, driven by the vast amount of digital content and the need to extract meaningful insights from financial texts. NER involves identifying and classifying named entities such as organizations, currencies, financial instruments, and monetary values, critical for applications like sentiment analysis, risk assessment, and investment recommendation systems. However, developing accurate and efficient cross-lingual NER models remains challenging due to sparse labeled data in low-resource languages and the complexities of capturing cross-lingual variations.

Cross-lingual NER is particularly crucial for industries operating globally, where analyzing customer feedback, social media posts, and other unstructured data in multiple languages can provide valuable insights for strategic decision-making. Achieving accurate entity recognition in diverse languages supports trend identification, enhances intelligence extraction, and improves operational efficiency. However, creating robust cross-lingual NER models requires substantial resources, including annotated data, multilingual expertise, and computational infrastructure (Nasar et al., 2021; Raiaan et al., 2024; Magueresse et al., 2020). Despite the major improvements in SOTA performance across a range of NLP tasks achieved by Large Language Models (LLMs), several challenges remain for adapting them to cross-lingual NER and NER in general. Supervised NER Baselines have been shown to outperform LLMs on NER tasks up until very recently (Wang et al., 2023; Zhou et al., 2023). Despite that, many of these LLMs adapted for NER require specific prompting techniques with no guarantees that these prompts will work across different domains. Additionally, these LLMs are extremely costly to train, fine-tune, and efficiently deploy due to their substantial parameter sizes, especially in low-resource cross-lingual settings. Thus, there is a growing need for efficient, scalable cross-lingual NER models capable of transferring knowledge from resource-rich to resource-scarce languages.

In this paper, we propose a novel framework to enhance cross-lingual named entity recognition. We designate English as a resource-rich language and Arabic as a low-resource language (Almansor

*Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal*, pages 54–62
January 19–20, 2025. ©2025 Association for Computational Linguistics

et al., 2020). Our framework leverages knowledge distillation to transfer insights from a teacher model pre-trained on a resource-rich language (English) to a compact student model. Subsequently, we employ consistency training to fine-tune the student model specifically for Arabic. Our experiments focus on identifying entities in semi-structured SMS texts containing financial transaction information in both English and Arabic, with access restricted to a small number of labeled examples in the target language. Additionally, we evaluate our model's performance using the WikiAnn dataset (Pan et al., 2017), comparing it against state-of-the-art models in cross-lingual knowledge transfer. Our findings demonstrate that our model achieves superior or comparable results while using significantly less training data than existing benchmarks.

The proposed approach showcases robust cross-lingual learning capabilities at a fraction of the data labeling costs typically associated with such tasks. Our contributions include:

- Leveraging knowledge distillation and consistency training to enhance cross-lingual NER.
- Demonstrating efficiency and effectiveness by requiring only a small number of labeled examples in the target language.
- Establishing competitive performance against state-of-the-art models, underscoring our model's ability to generalize effectively across different linguistic contexts.

This paper contributes to advancing the field of cross-lingual NER by proposing a pragmatic and scalable solution that addresses the challenges of language resource disparity in real-world applications.

## 2 Related Works

Cross-lingual Named Entity Recognition (NER) in low-resource languages poses significant challenges and has been a focal point of recent research efforts. Various approaches have emerged, leveraging transfer learning techniques to enhance NER performance across different languages. Transfer learning, particularly using pre-trained models like BERT and XLM-RoBERTa, has proven effective in adapting NER models to cross-lingual scenarios (Ma et al., 2022; Wu et al., 2020). These models capitalize on large-scale pre-training on diverse datasets, enabling them to learn robust representations that generalize well across languages.

Knowledge Distillation has emerged as a powerful technique in cross-lingual NER (Zhou et al., 2022; Wang and Henao, 2021). By transferring knowledge from a pre-trained teacher model to a smaller student model, knowledge distillation facilitates efficient knowledge transfer while maintaining performance. This approach is particularly advantageous in scenarios with limited labeled data, such as low-resource languages, where it enhances the student model's ability to capture complex linguistic patterns.

In addition to knowledge distillation, incorporating consistency training further enhances cross-lingual NER performance (Zhou et al., 2022; Wang and Henao, 2021). Consistency training uses unsupervised learning principles to enforce consistent predictions under small perturbations of the input data. This regularization technique improves model robustness and generalization, crucial for adapting NER models to diverse linguistic contexts.

Our proposed approach integrates both knowledge distillation and consistency training to address cross-lingual NER challenges in semi-structured financial text data in low-resource languages. By combining these techniques, we aim to leverage the strengths of pre-trained models while enhancing model adaptability to specific target languages, such as Arabic.

## 3 Methods

In this section, we present the methodology employed for the problem of cross-lingual NER for semi-structured financial text data in low-resource languages. This section is structured into three sub-sections: Problem Formulation, Model Architecture, and Training.

### 3.1 Problem Formulation

We formulate the task of cross-lingual NER as follows:

Let an input text sequence $X = \{x_1, x_2, ..., x_n\}$, where $n$ is the length of the sequence. Each token $x_i$ is associated with a label $y_i$, representing its NER tag. The set of possible NER tags is denoted as $Y = \{y_1, y_2, ..., y_k\}$, where $k$ is the total number of entity types.

Given a set of labeled data $D = \{(X_1, Y_1), (X_2, Y_2), ..., (X_m, Y_m)\}$, where each $(X_i, Y_i)$ pair represents an input text sequence and its corresponding NER tag, the objective is to learn a model $M$ that can accurately predict the NER tag $Y_i$ for an unseen input sequence $X_i$ in different languages. In this paper, we are using the Arabic language as a low-resource language.

The cross-lingual aspect of the problem arises from the scarcity of labeled data in low-resource languages. Therefore, the model $M$ should be capable of transferring knowledge from high-resource languages such as English, to low-resource languages, such as Arabic, to improve the performance of NER in those languages.

## 3.2 Model Architecture

To address these challenges of the cross-lingual NER task, we propose a novel framework based on knowledge distillation and consistency training. The model architecture is designed to leverage the benefits of both student-teacher knowledge distillation and consistency training.

The overall architecture consists of two components, namely (1) knowledge distillation with supervised cross-entropy loss and (2) consistency training.

### 3.2.1 Teacher Model

The teacher model $T$ is a pre-trained XLM-RoBERTa model (Conneau et al., 2020), fine-tuned on the source language (English) dataset. It serves as the source of knowledge transfer and provides soft target distributions for the student model during training. The teacher model takes input tokens $X$ and produces token-level predictions $P^T = \{p_1^T, p_2^T, ..., p_n^T\}$, where $p_i^T$ represents the probability distribution over the NER tags for the token $x_i$.

### 3.2.2 Student Model

The student model $S$ is a DistilBERT model (Sanh et al., 2019). It consists of a multi-layer transformer encoder, similar to the teacher model but with fewer layers and smaller hidden dimensions. The student model takes input tokens $X$ and produces token-level predictions $P^S = \{p_1^S, p_2^S, ..., p_n^S\}$, where $p_i^S$ represents the probability distribution over the NER tags for the token $x_i$.

### 3.2.3 Knowledge Distillation

We use knowledge distillation to transfer knowledge from the teacher model to the student model, to reduce the model size. The distillation loss combined with supervised cross-entropy loss (i.e., $\mathcal{L}_{CE}$) is defined as:

$$\mathcal{L}_{distill} = \alpha \mathcal{L}_{CE} + (1-\alpha)\mathrm{KL}\left(P^T \| P^S\right) \quad (1)$$

where $\alpha$ is the weight coefficient, and $P^T$ and $P^S$ represent the softened probability distributions obtained by applying the softmax function to the logits of the teacher model and the student model, respectively.

### 3.2.4 Consistency Training

After the knowledge distillation training, the student model is fine-tuned in the target language (Arabic) using consistency training. Consistency training encourages the model to produce consistent predictions when given different perturbations of the same input. We use a combination of supervised cross-entropy loss (i.e., $\mathcal{L}_{CE}$) and the unsupervised KL divergence as the consistency loss ($\mathcal{L}_{CT}$), comparing the predictions of the augmented data and the original data:

$$\mathcal{L}_{CT} = \alpha \mathcal{L}_{CE} + (1-\alpha)\mathrm{KL}\left(P^{aug} \| P^{orig}\right) \quad (2)$$

where $\alpha$ is the weight coefficient, and $P^{aug}$ and $P^{orig}$ represent the softmax probabilities obtained from the augmented data and the original data, respectively.

During consistency training, we generate augmented versions of the target language data (Xie et al., 2019) using back translation, RandAugment, and TF-IDF word replacement. These augmented examples are used to compute the unsupervised consistency loss and update the student model parameters accordingly.

## 3.3 Dataset

We conduct our experiments on a financial transactions dataset consisting of semi-structured SMS data in English and Arabic. The dataset is sourced from Egypt. The English language dataset consists of 1730 sentences along with associated annotated NER tags. The Arabic language dataset consists of 30 sentences. The limited size of the Arabic dataset is primarily due to challenges in acquiring larger datasets in the financial domain specific to Arabic-speaking regions. Despite its size, this dataset provides a valuable opportunity to explore cross-lingual NER in a low-resource setting. Figure 3 shows examples of the Arabic samples in the dataset. Both language datasets were preprocessed to hide sensitive information and converted to the standard IOB format for NER before training. In

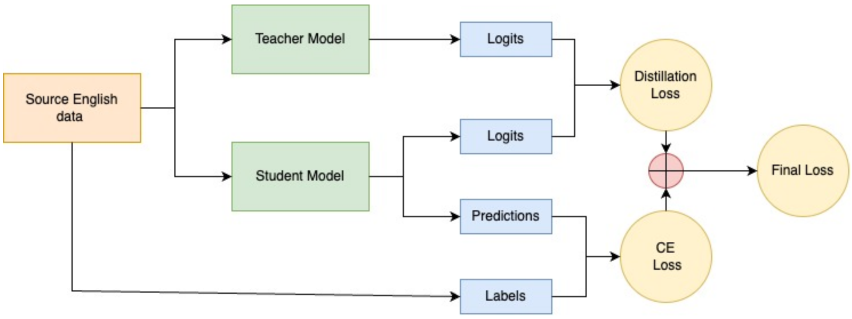


Figure 1: Overview of the student-teacher training framework (KD) with knowledge distillation and cross-entropy loss training on English data.

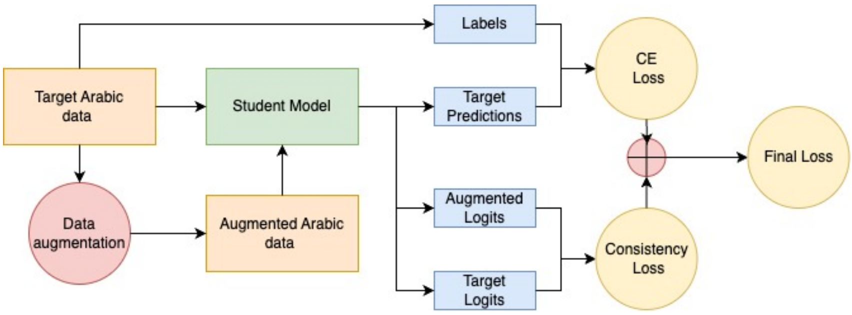


Figure 2: Overview of the knowledge distillation and consistency training framework (KD+CT) for training on Arabic data with consistency-training and cross-entropy loss.

تم شحن $\underbrace{7.00}_{amount}$ $\underbrace{\text{جنيه}}_{currency}$ مع العلم انه تم خصم $\underbrace{1.36}_{amount}$ $\underbrace{\text{جنيه}}_{currency}$ لسداد خدمة سوبر سلفني. رصيدك الحالي $\underbrace{5.82}_{balance}$ $\underbrace{\text{جنيه}}_{currency}$

*English Translation: 7.00 EGP have been added to your balance, with a deduction of 1.36 EGP for the Super Salefni service. Your current balance is 5.82 EGP.*

---

تم اضافة $\underbrace{620}_{amount}$ $\underbrace{EGP}_{currency}$ من بطاقة الخصم المباشر رقم $\underbrace{5791}_{card-number}$ عند $\underbrace{BERKET-AL-SABA-AL-RAAESY}_{supplier}$

يوم $\underbrace{25-10}_{date}$ الساعة $\underbrace{13:25}_{time}$ . للمزيد اتصل ب $\underbrace{16789}_{number}$

*English Translation: 620 EGP have been deducted from debit card number 5791 at BERKET-AL-SABA-AL-RAAESY on 25-10 at 13:25. Call 16789 for more information.*

Figure 3: Examples of Arabic samples with NER tagging and their English translations from the semi-structured financial transactions.

| Entity | En Data | Ar Data |
|---|---|---|
| amount | 3511 | 73 |
| supplier | 2968 | 29 |
| currency | 2490 | 34 |
| number | 2465 | 34 |
| full-date | 2234 | - |
| card-number | 1951 | 7 |
| full-time | 1938 | - |
| merchant | 1133 | 7 |
| balance | 494 | 8 |
| time | 135 | 8 |
| month | 99 | 2 |
| date | 10 | 43 |
| **Total Entities** | **19428** | **221** |

Table 1: Unique Named Entities in English and Arabic Datasets. Each row represents a specific named entity, and the corresponding columns indicate the count of occurrences for that entity in each dataset.

the IOB format, each token in the text is tagged with one of three labels: I (inside), O (outside), or B (beginning), indicating whether the token is inside a named entity, outside any named entity, or at the beginning of a named entity, respectively. This format facilitates accurate annotation and training of NER models by clearly delineating entity boundaries. The detailed distribution of unique named entities in these datasets can be found in Table 1.

Additionally, to complement our in-house dataset, we incorporated the WikiANN dataset (Pan et al., 2017) for evaluation purposes. WikiANN offers a diverse range of languages, including Arabic, and serves as a benchmark dataset. Here, English served as the target language, and Arabic as the source language. We selected a random subset of only 100 sentences from the Arabic portion of this dataset for training.

The Arabic language dataset is used unlabeled for the consistency loss and labeled for the supervised loss. The augmented dataset used for consistency loss is generated from this original dataset in the Arabic language.

### 3.4 Experimental Setup

We implement our NER model using the Transformers library and adopt the XLM-RoBERTa (Conneau et al., 2020) architecture for the teacher model and the DistilBERT (Sanh et al., 2019) architecture for the student model. We use AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate $l_r = 2e - 5$. We use a batch size of 28 and train the NER model for 20 epochs.

We run experiments on $\alpha$ over the range of $0, 0.2, 0.5, 0.8, 1$, (as shown in Figure 4). We set $\alpha = 0.8$ based on the best performance. At $\alpha = 0$ (i.e., only unsupervised consistency training loss), the NER model does not learn, and the validation loss increases. At $\alpha = 1$ (i.e., only supervised cross-entropy loss), we observe overfitting on the limited target language data (Arabic), and the validation loss starts to increase after going down. However, at $\alpha = 0.8$ (combination of supervised and unsupervised losses), the NER model gives the best performance on the cross-lingual NER task for low-resource language.

In addition to experiments on our financial dataset, we also conducted experiments on the WikiANN benchmark dataset. For these experiments, we used the English language as the source language and Arabic as the target language. Empirically, we found that setting $\alpha = 0.8$ yielded the best performance on the WikiANN dataset.

Furthermore, we also conducted experiments with a 1000-sample subset of the WikiANN Arabic dataset, and the appropriate $\alpha$ value for this configuration was determined to be 0.2. This observation highlights how the optimal $\alpha$ value can vary depending on the dataset's characteristics, especially with respect to data size and complexity.

### 3.5 Performance Comparison

#### 3.5.1 Comparison on Financial Dataset

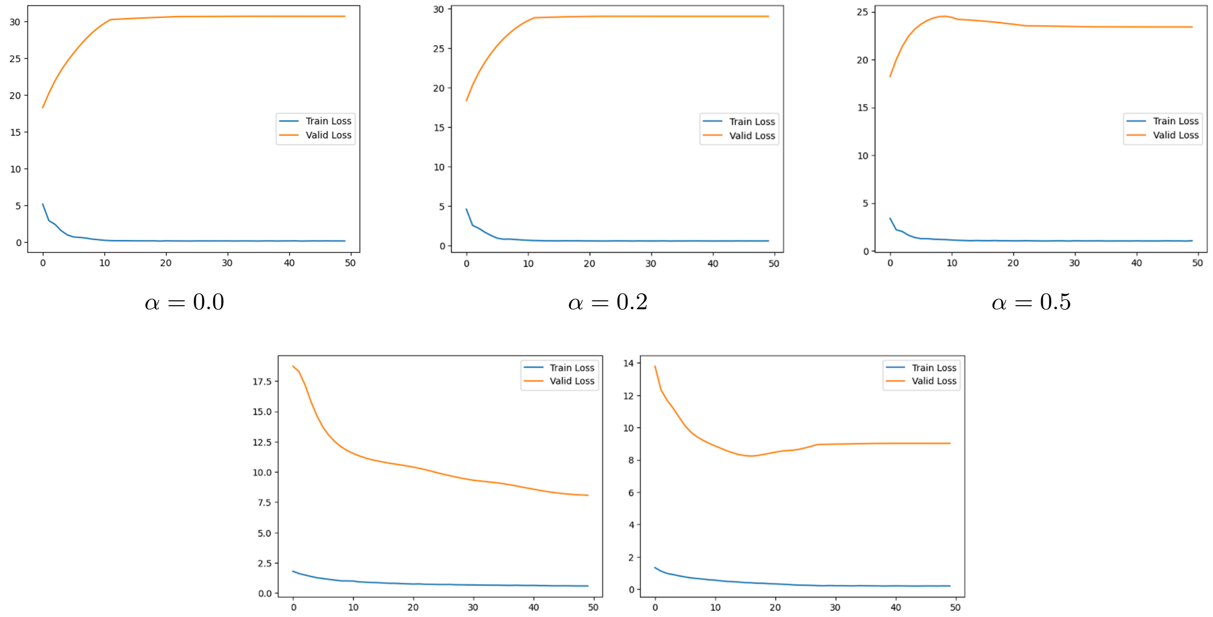
To evaluate the performance of our proposed cross-lingual NER model on our financial transactions dataset, we compare it with that of several baseline models. The baselines include:

1. Teacher Model: A pre-trained large language model (XLM-RoBERTa) fine-tuned on the English language dataset.

2. Student Model: A DistilBERT-based student model trained using knowledge distillation from the teacher model.

3. Naive Benchmark Model: A pre-trained DistilBERT model fine-tuned on the target language (Arabic) dataset.

We report the performance comparison in terms of F1 score and accuracy for NER on both the source (English) and the target (Arabic) datasets.



$\alpha = 0.0$

$\alpha = 0.2$

$\alpha = 0.5$

$\alpha = 0.8$

$\alpha = 1.0$

Figure 4: The training and validation loss of the KD+CT model (our model) over different values of $\alpha$ in $[0, 0.2, 0.5, 0.8, 1]$ when training it on Arabic data. The loss value and the number of epochs are on the $y$-axis and $x$-axis, respectively. The results indicate that when setting the value of $\alpha$ to 0 and 0.2, the model exhibits overfitting behavior on the validation data, as evidenced by an increase in validation loss while the training loss continues to decrease. For $\alpha$ equal to 0.5 and 1, overfitting is still present but not so severe as it was for smaller $\alpha$. Finally, we empirically found that $\alpha = 0.8$ shows the most desirable learning behavior for validation loss, which almost linearly decreases for the duration of the training.

### 3.5.2 Comparison on WikiANN Dataset

In addition to evaluating our model on our financial dataset, we also compared its performance on the WikiANN benchmark dataset with existing state-of-the-art models. We compare our results against MSD (Ma et al., 2022) and ConNer (Zhou et al., 2022). The MSD (Mixture of Short Distillers) model makes use of the rich representations learned from the hidden layers of mBERT (Devlin et al., 2019) instead of distilling knowledge only from the last layer. The ConNer model introduces two variants of consistency training by translating sentences into a different language at the data level and applying dropout at the representation level to induce perturbations, thus forcing the model to learn more general features rather than specific ones. We report results in terms of the F1-score on Arabic data. We do not compare with LLMs as we are targeting low-resource scalable settings. All the models we compare with are trained on consumer GPUs. MSD & ConNER (Ma et al., 2022; Zhou et al., 2022) are both trained on an RTX 3090 GPU. Our model is trained on a T4 GPU available through the free version of Google Colab.

## 3.6 Results and Analysis

### 3.6.1 Performance on Financial Dataset

We compare the performance of our NER model with the Teacher model, the Student model, and the Naive Benchmark model on both the source (English) and the target (Arabic) datasets.

On the English dataset, our model achieves an F1 score of 0.9768. Although the Teacher and Student models exhibit higher F1 scores, our model achieves comparable performance while being smaller than the Teacher model, with an F1 score of 0.9887. On the Arabic dataset, our model significantly outperforms the Teacher and the Student models, reaching an F1 score of 0.6540 and an accuracy of 0.7407. Furthermore, our model performs better than the Naive Benchmark model having an F1 score of 0.6065.

These results show that our model achieves competitive performance on both the English (source) and the Arabic (target) datasets. Despite its smaller size and the limited data available in the target language, our model demonstrates remarkable cross-lingual generalization capabilities. It effectively leverages the knowledge distilled from the

|  | English |  | Arabic |  |
|---|---|---|---|---|
| **Model** | **F1** | **Acc** | **F1** | **Acc** |
| Teacher | 0.9887 | 0.9888 | 0.5929 | 0.6543 |
| Student (only KD) | **0.9957** | **0.9957** | 0.5693 | 0.6852 |
| DistilBERT | 0.6263 | 0.7377 | 0.6065 | 0.7099 |
| KD+CT (Our Model) | 0.9768 | 0.9782 | **0.6540** | **0.7407** |

Table 2: Comparison of the NER performance of the models on English and Arabic datasets. The accuracies and F1 scores are shown for both English and Arabic datasets. Our model's results support our assertion that learning to recognize entities in the high-resource source language (English) can lead to better performance on the low-resource target language (Arabic), even with just a few labeled examples.

Teacher model and further enhances its performance through consistency training on the limited target language data.

### 3.6.2 Performance on WikiANN Benchmark

In addition to evaluating our model on our financial dataset, we also conducted experiments on the WikiANN benchmark dataset. Our results on the WikiANN dataset (as shown in Table 3) are promising and align with our main argument: utilizing a few samples of the target language in semi-supervised learning outperforms unsupervised approaches, even when dealing with smaller datasets. Our model outperforms ConNER with an F1 score of $0.62$ and gives an on-par performance with MSD, while using only a small subset of training data in Arabic. Our model's ability to generalize effectively to Arabic, despite limited labeled data, underscores its potential for cross-lingual NER in low-resource settings. We also tested the benchmark models, ConNER and MSD, by training them on a 100-sample Arabic dataset, similar to our model. However, the results from the benchmark models exhibited bias and poor performance, potentially due to the limited Arabic dataset.

In continuation of our experiments, we also evaluated the performance of both the teacher model and our KD+CT model on the WikiANN Arabic 100-sample dataset. The results (as shown in Table 4) showed that while the teacher model achieved a recall of $0.62$, the KD+CT model demonstrated notably higher precision, reaching $0.87$. This emphasis on precision holds particular significance in domains such as finance and related fields, where accurate identification of entities is crucial.

### 3.6.3 Analysis

The overall superior performance of our model can be attributed to its ability to capture and transfer the underlying patterns learned by the Teacher model, leveraging the knowledge distilled during the training process. By incorporating consistency training, our model achieves more robust predictions by ensuring consistency across augmented versions of the input sequences. This training mechanism enhances the model's ability to adapt to cross-lingual contexts and improve performance. The successful combination of knowledge distillation and consistency training contributes to the model's superior performance in capturing both the general patterns and specific language characteristics required for effective cross-lingual named entity recognition.

Overall, our proposed cross-lingual NER model emerges as a promising approach for low-resource languages. Its ability to achieve competitive performance with a smaller model size makes it a practical and efficient solution for real-world applications.

## 4 Limitations

While our approach demonstrates promising results in cross-lingual NER, it has several limitations. One key limitation is the inconsistency in results for minority classes. This inconsistency arises from the scarcity of samples for certain classes in the data, which is already limited in the low-resource setting. This could be overcome by choosing well-balanced data and skipping samples with high "$O$" class entities in the target language. It will lead to better generalization

Finally, our approach requires a small amount of labeled data in the target language for consistency training. While this requirement is minimal compared to fully supervised methods, it may still pose challenges in scenarios where even a small amount of labeled data is not available.

|  |  | F1 Score |  |  |
|---|---|---|---|---|
| **Model** | **# Params** | 100% of Samples | 1000 Samples | 100 Samples |
| ConNER | 355M | 0.59 | 0.35 | 0.38 |
| MSD | 111M | **0.62** | 0.52 | 0.16 |
| KD+CT(Our Model) | **66M** | - | - | **0.61** |

Table 3: Comparison of the NER performance (entity-level F1 scores) of the models on the WikiANN dataset. The performances for benchmarks that utilize 100% of Arabic samples are taken directly from their respective papers. Our model provides better or comparative performance to other state-of-the-art models while utilizing only an extremely small fraction of the data used.

| **Model** | **Precision** | **Recall** |
|---|---|---|
| Teacher | 0.64 | **0.62** |
| KD+CT (Our Model) | **0.87** | 0.50 |

Table 4: Comparison of the NER performance of the teacher model and KD+CT model on the WikiANN dataset, in terms of precision and recall. Our model enhances the precision across all entities in the target language (Arabic).

## 5 Conclusion

In this paper, we introduce a novel framework that uses knowledge distillation and consistency training to enhance cross-lingual named entity recognition. Knowledge is transferred from a teacher model pre-trained in English to a smaller student model, which is then fine-tuned for Arabic. Our model, KD+CT, is validated on banking transaction data (semi-structured) in both English and Arabic, showcasing competitive performance compared to state-of-the-art benchmarks on several datasets.

Our modeling approach successfully combines knowledge distillation with consistency training, addressing the challenges of developing accurate cross-lingual NER models for low-resource languages. Importantly, our model significantly outperforms the naive benchmark, the student, and the teacher models in entity recognition on the target language dataset (Arabic) and achieves performance comparable to the larger teacher model while being approximately 3 times smaller in terms of parameters (66 million parameters compared to the teacher model's 270 million parameters) on the source language dataset (English). This demonstrates the remarkable cross-lingual generalization capabilities of our KD+CT model, effectively leveraging the knowledge distilled from the high-resource language and enhancing performance on the low-resource language through consistency training. Additionally, we evaluate our model on the WikiANN dataset, achieving competitive results against state-of-the-art methods, even with minimal labeled data in the target language. Notably, our model showcases an improvement in the precision metric, achieving a precision of 0.87 compared to the teacher model's 0.64. This improvement is particularly significant in the financial sector, where label accuracy is vital.

Our proposed cross-lingual NER model offers valuable contributions to the development of multilingual applications, enabling the extraction of insights, identification of trends, and making well-informed decisions across multiple languages. We hope that our work will inspire further research in this field and facilitate the development of efficient and effective cross-lingual NER models, benefiting low-resource languages and beyond.

## 6 Future Work

Our work establishes an efficient avenue for cross-lingual NER, yet several exciting prospects for further research remain. An immediate extension of our work would involve studying the effect the volume of labeled data has on the performance metrics of our method against other state-of-the-art models.

Another extension of our research includes examining our method's performance across a broader array of low-resource languages. This would give us better insights into the scalability of our method to other low-resource languages.

Furthermore, exploring the potential for resource-efficient NER labeling through the use of commercial Large Language Models (LLMs) as a means for data augmentation offers a compelling new research direction, especially in light of the increasing prevalence of LLMs.

## References

Ebtesam Almansor, Ahmed Al-Ani, and Farookh Hussain. 2020. *Transferring Informal Text in Arabic as Low Resource Languages: State-of-the-Art and Future Research Directions*, pages 176–187.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.*, 54(1).

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Annual Meeting of the Association for Computational Linguistics*.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5308, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *ArXiv*, abs/2304.10428.

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.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2022. ConNER: Consistency training for cross-lingual named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8438–8449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *ArXiv*, abs/2308.03279.