PANER: A Paraphrase-Augmented Framework for Low-Resource Named Entity Recognition

Nanda Kumar Rengarajan

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Computer Science at

Concordia University

Montréal, Québec, Canada

August 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Nanda Kumar Rengarajan**

Entitled: **PANER: A Paraphrase-Augmented Framework for Low-Resource Named Entity Recognition**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Jinqiu Yang

_____ Examiner
Dr. Abdelhak Bentaleb

_____ Supervisor
Dr. Jun Yan

_____ Co-supervisor
Dr. Chun Wang

Approved by _____
Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

PANER: A Paraphrase-Augmented Framework for Low-Resource Named Entity Recognition

Nanda Kumar Rengarajan

Named Entity Recognition (NER) is a critical task that requires substantial annotated data, making it challenging in low-resource scenarios where label acquisition is expensive. While zero-shot and instruction-tuned approaches have made progress, they often fail to generalize to domain-specific entities and do not effectively utilize the limited available data. We present a lightweight few-shot NER framework that addresses these challenges through two key innovations: (1) a new instruction tuning template with a simplified output format that combines principles from prior IT approaches to leverage the large context window of recent state-of-the-art LLMs; (2) introducing a strategic data augmentation technique that preserves entity information while paraphrasing the surrounding context, thereby expanding our training data without compromising semantic relationships. Experiments on benchmark datasets demonstrate that our method achieves performance comparable to that of state-of-the-art models on few-shot and zero-shot tasks, with our few-shot approach attaining an average F1 score of 80.1% on the CrossNER datasets. Models trained with our instruction tuning approach exhibit consistent improvements in F1 scores of up to 17% points over comparable baselines, providing a promising solution for groups with limited NER training data and computational resources.

Acknowledgments

This work would not have been possible without the support of many people inside and outside of Concordia University. I want to start by thanking both my supervisors, Dr. Jun Yan and Dr. Chun Wang, for giving me the opportunity to pursue this program and for their continued support and guidance in helping me refine the structure and content of this thesis. I also appreciate their help in securing the MITACS internship at the 7dish organization. To Vincent Trepanier and Simon Oliver Harel, thank you for letting me be part of your project, where I learned and developed the skills I needed for this work.

I'm incredibly grateful to Negin Ashouri and Erfan Fatehi of Femtherapeutics for trusting me to execute their vision and for encouraging me to take courses that expanded my knowledge on the subject—all of which were essential for developing the methods presented here. I also want to thank Charles Frye of Modal for providing me with credits to their platform, where all the experiments in this thesis were conducted, and the Modal team for building such a great platform.

To Dr. Abdelhak Bentaleb and Dr. Jinqiu Yang, thank you for taking the time to review my thesis and for your challenging questions during my defense.

To my parents Kumar and Mohana, thank you for your endless love and support that made this journey possible. Your late-night check-ins from back home to make sure I was eating well and getting enough rest, along with your unwavering support for my dream of studying abroad, mean the world to me. To my friends Sharanyu, San, and Ramya, thank you for your amazing friendship over these past few years, always pushing me to do better while keeping me grounded with the kind of laughter that got me through the tough times. To all my other friends, thank you for sticking with me through this process and for all the good times that made everything worthwhile.

Finally, to my girlfriend/partner/everything, Subhi, thank you for putting up with me, loving me, and always believing in me. You weren't just helpful—you were inspiring. I hope I can be the same support for you. And for all those dates I missed because of this thesis, I'm ready to make up for every single one.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing(NLP) that involves identifying and classifying named entities such as persons, organizations, locations, and domain-specific entities within unstructured text. NER serves as a critical foundation for numerous downstream applications including knowledge graph construction, recommendation systems, and dialogue systems J. Li, Sun, Han, and Li (2020). In knowledge graph construction, NER serves as the basis for extracting structured information from unstructured text, where recognized entities function as nodes that enable the transformation of natural language into structured knowledge that can be processed Al-Moslmi, Ocaña, Opdahl, and Veres (2020). In question answering systems, NER functions as a core component that helps preselect answer candidates and improves overall system performance Mollá, Van Zaanen, and Smith (2006). For information retrieval and extraction, NER has been widely applied to improve system accuracy and user intent understanding J. Li et al. (2020).

Traditional NER systems rely heavily on supervised learning approaches that require extensive manually annotated datasets for specific domains and predefined entity types. This dependency creates significant barriers for organizations operating in specialized domains where comprehensive training datasets are either extremely expensive or unavailable. The emergence of Large Language Models has introduced new paradigms for NER through instruction tuning, demonstrating promising zero-shot and few-shot capabilities that can potentially address these resource constraints.

Recent instruction-tuned approaches have made notable progress but face critical limitations in

practical deployment. GNER Y. Ding et al. (2024) emphasizes the importance of negative instances in improving contextual understanding and entity boundary delineation by including non-entity text in training examples. While this approach demonstrates improved boundary detection through a modified BIO-like generation format, it struggles against the out-of-distribution entities. Additionally, GNER is also a computationally demanding approach (both time and memory-wise) resources which may not be available in resource-constrained environments.

SLIMER Zamai, Zugarini, Rigutini, Ernandes, and Maggini (2024) introduces enriched prompts with entity definitions and annotation guidelines, demonstrating effective performance with minimal training data on unseen entities. However, SLIMER employs a turn-by-turn conversational approach that requires separate queries for each entity type within a given sentence. For instance, when processing a sentence containing multiple entities, the system must pose individual questions such as ”What is the person in this sentence?” and ”What is the organization in this sentence?” This sequential querying mechanism becomes computationally expensive as the number of entity types increases, with inference costs scaling linearly with the number of entities present.

Data augmentation strategies for NER face additional challenges in maintaining semantic consistency while introducing meaningful variation. Existing methods often struggle to preserve entity relationships when dealing with sentences containing multiple entities, where maintaining precise entity relationships is crucial for accurate NER performance. Many approaches either risk corrupting entity information through aggressive modifications or fail to introduce sufficient variation to enhance model generalization.

This thesis addresses these limitations by developing a novel framework that combines strategic paraphrase-based data augmentation with simplified instruction tuning. Our approach aims to reduce the complexity of existing tagging formats while preserving accuracy, implement controlled paraphrasing techniques that maintain entity integrity while expanding diversity, and achieve competitive performance through efficient fine-tuning strategies. By addressing the specific problems identified in existing approaches, this work seeks to provide a practical solution for organizations requiring effective NER capabilities without extensive computational resources or large annotated datasets.

1.1 Objectives

Over the past few years, NER has become increasingly crucial in Natural Language Processing (NLP), enabling advances in information extraction, question answering, and event detection. While traditional supervised approaches have shown strong performance, they remain constrained by their reliance on large annotated datasets and domain-specific training data. The emergence of Large Language Models (LLMs) has introduced new paradigms for addressing NER challenges, particularly in zero-shot and few-shot scenarios, but significant challenges remain in making these approaches practical and accessible for real-world applications. This thesis seeks to address the fundamental challenge of performing NER in low-resource settings by developing novel approaches that combine the strengths of instruction tuning and data augmentation. We aim to bridge the gap between the impressive capabilities of large language models and the practical constraints faced by organizations with limited computational resources and annotated data.

To achieve this goal, we have developed PANER, a comprehensive framework that introduces two key innovations: (1) a refined instruction tuning methodology that combines principles from existing approaches while simplifying the output format, and (2) a strategic paraphrase-based data augmentation technique that preserves entity information while expanding linguistic variety in the training data.

Our research focuses particularly on domain adaptation and few-shot learning scenarios, where we evaluate the effectiveness of our approach across diverse domains including scientific papers, politics, music, and literature. We explore how paraphrase augmentation can enhance model performance while maintaining computational efficiency through techniques like LoRA fine-tuning and optimized prompt design.

This work contributes to the growing body of research in few-shot learning, data augmentation, and instruction tuning, with broader implications for making advanced NLP capabilities more accessible to organizations with limited resources. By demonstrating that competitive performance can be achieved with fewer computational resources and training data, we aim to provide practical solutions for real-world NER applications across various domains and languages. The thesis evaluates our approach through extensive experimentation on multiple benchmark datasets, comparing

against state-of-the-art methods in both zero-shot and few-shot settings, while providing detailed analysis of the factors contributing to improved performance. We also explore the limitations of our approach and suggest future directions for research in low-resource NER.

1.2 Methodology

The methodology employed in this thesis involves two key components. First, we develop a paraphrasing framework using LLAMA 3.3-70B Touvron et al. (2023), a state-of-the-art large language model, to generate high-quality paraphrased training data. This controlled paraphrasing approach preserves entity information while modifying surrounding context, effectively expanding training data without compromising semantic relationships through a strict validation pipeline.

Second, we present fine-tuned versions of instruction-tuned Large Language Models Qwen-2.5-Instruct (7B) Yang et al. (2024), LLAMA-3.1-Instruct (8B) Touvron et al. (2023), and Falcon3-Instruct (10B) Almazrouei et al. (2023) for NER with a simplified word/tag output format enriched with entity definitions and guidelines, and optimized using LoRA E. J. Hu et al. (2022) fine-tuning techniques for computational efficiency in both few-shot and zero-shot scenarios.

The complete framework is evaluated on benchmark datasets, including CrossNER, MIT, and BUSTER, across diverse domains. All implementation code, including instruction tuning templates, paraphrasing framework, and evaluation scripts, is available on GitHub. [1]

1.3 Contributions

This thesis makes several contributions to the field of NLP, particularly in the areas of NER, few-shot learning, and data augmentation:

(1) **Simplified Instruction Tuning for NER:** We propose and evaluate a refined instruction tuning methodology that combines principles from prior approaches while introducing a simplified word/tag output format. This approach leverages enriched prompts with entity definitions and guidelines while avoiding the complexity of the traditional BIO tagging schema. Models using

[1] GitHub repository: https://github.com/parzival11/masters-thesis-NER

our format demonstrated significant improvements in F1 scores compared to traditional approaches, achieving up to 17% improvement over baseline versions.

(2) **Paraphrase-Based Data Augmentation:** We introduce a strategic paraphrasing technique that preserves entity information while expanding linguistic variety in the training data. Our technique consistently improved model performance across multiple domains and datasets, with our few-shot approach attaining an average F1 score of 80.1 on the CrossNER datasets. The detailed implementation and results of this work are provided in Chapter 4.

Overall, this research contributes to making advanced NER more accessible and effective, particularly in low-resource scenarios where traditional approaches requiring extensive training data and computational resources may not be feasible.

1.4 Thesis Structure

This chapter has outlined the motivation, goals, and contributions of this thesis. Chapter 2 provides a comprehensive literature review covering the historical evolution of NER, from traditional supervised approaches to modern Large Language Model-based methods, with particular focus on challenges in low-resource scenarios and recent developments in instruction tuning and few-shot learning. Chapter 3 details our proposed PANER framework, including the paraphrasing-based data augmentation methodology, simplified instruction tuning template design in detail. Chapter 4 describes the experimental setup, covering benchmark datasets, baseline comparisons, and implementation details for evaluating our approach across domains and showcases parameter-efficient fine-tuning strategies that were used. Chapter 5 presents the results and analysis, examining the effectiveness of our simplified tagging format, paraphrase-based augmentation strategies, and comparative performance against state-of-the-art methods in both zero-shot and few-shot scenarios and extensive ablation studies justifying our choices. Chapter 6 discusses the broader implications of our findings, the role of LLM-based paraphrasing in NER tasks, and limitations of the current approach. Finally, Chapter 7 summarizes the key contributions of this work and proposes potential directions for future research in low-resource NER.

Chapter 2

Literature Review

This chapter provides a comprehensive review of the foundational approaches and recent developments in NER, with particular emphasis on techniques designed to address data scarcity challenges. Section 2.1 traces the historical evolution of NER methodologies, examining the progression from traditional supervised approaches through neural network-based methods to contemporary transfer learning paradigms that leverage pre-trained language models. Each evolutionary phase demonstrates how fundamental challenges in entity recognition have been addressed through increasingly sophisticated computational techniques. Following this, Section 2.2 provides a detailed analysis of the multifaceted challenges confronting low-resource NER systems. This section categorizes the main obstacles into distinct categories and what has been done to address those challenges (many of these challenges predate the emergence of large language models). Section 2.3 explores the emergence of instruction tuning and few-shot learning approaches, analyzing how large language models have transformed the landscape of NER by enabling effective performance with minimal training data. Next, Section 2.4 examines data augmentation strategies specifically developed for NER tasks, including traditional methods such as back-translation and entity replacement, as well as recent innovations in synthetic data generation. The focus then shifts to paraphrase-based augmentation techniques, exploring how contextual modifications can preserve entity integrity while enhancing training data diversity.

2.1 Named Entity Recognition

Named Entity Recognition (NER) has evolved significantly since its introduction in the mid-1990s. This section traces the development of NER techniques from traditional supervised approaches through neural network-based methods to the current state-of-the-art transfer learning paradigms that leverage pre-trained language models. The progression of these approaches demonstrates not only the technological evolution in computational linguistics but also how the fundamental challenges of entity recognition have been addressed with increasingly sophisticated methods.

2.1.1 Traditional Supervised Learning for NER

NER emerged as a distinct natural language processing task during the Sixth Message Understanding Conference (MUC-6) in 1995 Grishman and Sundheim (1995). This conference represented a significant milestone in NLP development, as it defined the classic entity categories that would become foundational in the field: *person, location,* and *organization*. The articulation of NER as a focused task, rather than merely a component of broader information extraction systems, enabled researchers to concentrate specifically on the unique challenges of identifying and classifying named entities in text.

In the years following MUC-6, statistical machine learning approaches began to replace the rule-based systems that had dominated earlier information extraction tasks. A significant breakthrough came with the introduction of Hidden Markov Models (HMMs) to NER. Bikel, Miller, Schwartz, and Weischedel (1998) developed a system called "Nymble," which utilized HMMs to model the sequential nature of text and identify named entities without relying on hand-crafted rules. Their approach demonstrated remarkable improvements over previous rule-based systems, particularly in handling previously unseen entities. The probabilistic framework of HMMs allowed the system to learn patterns from large corpora of annotated text, making it more adaptable to different domains and languages.

However, HMMs faced limitations due to their strong independence assumptions, which failed to capture the complex interdependencies between features in natural language. This limitation led to the exploration of Maximum Entropy (MaxEnt) models for NER. Borthwick (1999) proposed a

MaxEnt approach that could incorporate diverse, overlapping features without assuming independence between them. This flexibility allowed for the integration of lexical, syntactic, and semantic features, as well as gazetteer information, resulting in more accurate entity recognition. The MaxEnt framework was particularly effective in handling ambiguous cases where context was crucial for correct entity classification.

An important advancement in statistical NER came with the introduction of Conditional Random Fields (CRFs). Lafferty, McCallum, and Pereira (2001) developed CRFs as a framework for building probabilistic models to segment and label sequence data. Unlike HMMs, CRFs are discriminative models that directly model the conditional probability of the label sequence given the observation sequence. This approach addressed the label bias problem that affected previous probabilistic models. McCallum and Li (2003) applied CRFs to NER tasks and demonstrated their superiority over both HMMs and MaxEnt models. Their work incorporated feature induction techniques and web-enhanced lexicons, showcasing how CRFs could effectively leverage diverse sources of information. The success of CRFs established them as the dominant architecture for NER for nearly a decade, and many subsequent studies built upon this foundation.

As NER techniques matured, researchers began to extend their applications beyond English to other languages. The Conference on Computational Natural Language Learning (CoNLL) played a crucial role in this internationalization by organizing shared tasks focused on language-independent NER. Tjong Kim Sang (2002) introduced the CoNLL-2002 shared task, which focused on NER for Spanish and Dutch, establishing benchmarks for non-English NER systems. This initiative was followed by similar efforts for German and other languages, highlighting the universal challenges and language-specific nuances in NER.

The traditional supervised approaches to NER established fundamental techniques and evaluation metrics that continue to influence the field. They demonstrated the importance of sequential modelling, feature engineering, and the integration of linguistic knowledge. However, these approaches also highlighted the limitations of manual feature engineering and the challenges of adapting models to new domains or languages without substantial annotated data. These limitations would eventually drive the shift toward neural network-based approaches, which offered the potential for more flexible and adaptable NER systems.

2.1.2 Neural Network-Based NER

The transition from traditional supervised methods to neural network approaches marked a paradigm shift in NER research. This evolution was catalyzed by Collobert et al. (2011), who demonstrated that neural networks could learn word representations directly from data, reducing the need for hand-engineered features. Their system, which approached NLP tasks "almost from scratch," represented a significant departure from previous methodologies. By learning continuous vector representations of words, their neural architecture could capture semantic similarities and relationships between words that were difficult to express through discrete features. This approach achieved competitive performance on multiple NLP tasks, including NER, while substantially reducing the effort required for feature engineering.

The revolution in word representation was further advanced by Mikolov, Sutskever, Chen, Corrado, and Dean (2013), who introduced the Word2Vec model for learning distributed representations of words and phrases. Although not specifically designed for NER, Word2Vec's ability to capture semantic relationships between words in a low-dimensional vector space proved invaluable for entity recognition tasks. These pre-trained word embeddings provided a richer starting point for neural NER models, allowing them to leverage semantic information learned from large, unannotated corpora. The incorporation of these word embeddings into NER systems demonstrated how advances in representation learning could benefit specialized NLP tasks.

While word-level representations provided valuable semantic information, they often struggled with morphologically rich languages and out-of-vocabulary words. To address these limitations, Santos and Guimaraes (2015) introduced character-level convolutional neural networks (CNNs) for NER. Their system, CharWNN, combined character-level and word-level representations to better handle morphological variations and rare words. By processing character sequences through convolutional layers, the model could learn to recognize patterns in word formation that were relevant for entity identification. This character-level approach was particularly beneficial for languages with complex morphology and for technical domains with specialized vocabulary.

The most significant architectural innovation in neural NER came with the introduction of

Bidirectional Long Short-Term Memory networks combined with Conditional Random Fields (Bi-LSTM-CRF). Z. Huang, Xu, and Yu (2015) proposed this hybrid architecture, which leveraged the strengths of both neural networks and statistical models. The Bi-LSTM component captured long-range dependencies in both forward and backward directions, providing rich contextual representations for each word. The CRF layer, meanwhile, imposed structural constraints on the output sequence, ensuring that the predicted entity labels followed valid patterns. This combination addressed the limitations of purely neural approaches, which sometimes struggled with consistent sequence labelling.

Ma and Hovy (2016) further refined the Bi-LSTM-CRF architecture by incorporating character-level CNNs, resulting in an end-to-end sequence labelling model. Their system combined the benefits of character-level representations, word embeddings, bidirectional sequence modelling, and structural prediction. This comprehensive architecture achieved state-of-the-art performance on multiple NER benchmarks, demonstrating the effectiveness of integrating different levels of linguistic information within a unified neural framework. The success of this approach established the Bi-LSTM-CRF as the dominant neural architecture for NER for several years.

The neural network-based approaches to NER represented a significant advance over traditional supervised methods. By learning representations directly from data, these models reduced the need for manual feature engineering and demonstrated greater flexibility in adapting to different domains and languages. The integration of character-level and word-level information, along with bidirectional sequence modelling and structural prediction, allowed neural models to capture complex patterns in entity recognition. However, these approaches still required substantial amounts of task-specific training data, a limitation that would be addressed by the subsequent wave of transfer learning and pre-trained language models.

2.1.3 Transfer Learning and Pre-trained Language Models

The most recent paradigm shift in NER has been driven by the application of transfer learning and pre-trained language models. This approach leverages large-scale unsupervised pre-training on diverse textual data, followed by fine-tuning on specific NER tasks. This methodology has dramatically reduced the amount of task-specific annotated data required while simultaneously improving

performance across a wide range of NER applications.

A groundbreaking development in this area was the introduction of Embeddings from Language Models (ELMo) by Peters et al. (2018). Unlike previous static word embeddings like Word2Vec, ELMo provided contextualized word representations that captured how word meaning varies depending on the surrounding context. Generated from a bidirectional language model trained on a large corpus, ELMo embeddings encoded rich syntactic and semantic information that proved highly beneficial for NER. The contextual nature of these representations allowed the model to better handle polysemy and entities based on their usage in specific sentences. When incorporated into existing NER architectures, ELMo embeddings led to significant performance improvements across multiple benchmarks.

The revolution in transfer learning for NER reached new heights with the introduction of Bidirectional Encoder Representations from Transformers (BERT) by Devlin, Chang, Lee, and Toutanova (2019). BERT represented a fundamental shift in NLP architecture, replacing recurrent neural networks with transformer-based models that could process entire sequences in parallel. Pre-trained on massive corpora using masked language modelling and next sentence prediction objectives, BERT captured deep bidirectional contextual information. When fine-tuned for NER, BERT models achieved state-of-the-art performance levels. The pre-training and fine-tuning paradigm introduced by BERT established a new standard for NER development, dramatically reducing the need for task-specific architecture design and feature engineering.

The multilingual capabilities of pre-trained language models opened new possibilities for cross-lingual transfer in NER. Pires, Schlinger, and Garrette (2019) investigated the cross-lingual effectiveness of multilingual BERT (mBERT) across various tasks, including NER. Their work demonstrated that mBERT, despite not having explicit cross-lingual objectives during pre-training, could effectively transfer knowledge between languages. This was particularly valuable for low-resource languages, where annotated NER data was scarce. The authors compared fine-tuning and supervised approaches to transfer learning, providing insights into the most effective strategies for leveraging pre-trained models in multilingual settings.

Building upon the BERT architecture, Y. Liu et al. (2019) introduced RoBERTa (Robustly Optimized BERT Pre-training Approach), which refined the pre-training methodology through careful

optimization of hyperparameters and training strategies. By removing the next sentence prediction objective, training on longer sequences, dynamically changing masking patterns, and utilizing substantially larger mini-batches and learning rates, RoBERTa achieved significant performance improvements across multiple NLP tasks. In the context of NER, RoBERTa established new state-of-the-art results on the CoNLL-2003 English NER task with an F1 score of 92.4%, surpassing previous benchmarks. Y. Liu et al. (2019) further demonstrated RoBERTa's effectiveness by achieving superior performance on the OntoNotes 5.0 dataset, which features a more diverse set of entity types across multiple genres of text. This established RoBERTa as a powerful foundation for entity recognition systems, with many subsequent studies using it as a starting point for domain-specific adaptations and task-specific enhancements in specialized NER applications.

The potential of pre-trained language models for few-shot learning in NER was explored by J. Huang et al. (2020). Their comprehensive study demonstrated that with appropriate fine-tuning strategies, pre-trained models could achieve remarkable performance on NER tasks with very limited labelled data. This capability was particularly valuable for specialized domains or rare entity types where large annotated datasets were unavailable.

The transfer learning paradigm based on pre-trained language models has transformed the landscape of NER research. The ability to fine-tune pre-trained models on relatively small datasets has democratized NER development, making it possible to build effective systems for specialized domains and low-resource languages. As pre-trained models continue to evolve, with larger architectures and more sophisticated pre-training objectives, the performance and applicability of NER systems are likely to improve further.

The historical evolution of NER from traditional supervised approaches through neural network-based methods to transfer learning with pre-trained language models reflects broader trends in natural language processing. Each new paradigm has built upon the insights and achievements of its predecessors while addressing their limitations. The current state-of-the-art approaches combine the strengths of deep contextual representations, transformer architectures, and transfer learning, resulting in NER systems that are both more powerful and more adaptable than ever before. As research continues, we can expect further innovations that will enhance the performance, efficiency, and applicability of NER across diverse domains and languages.

2.2 Challenges in Low-Resource NER

NER has witnessed significant advances in recent years, particularly with the advent of deep learning techniques. However, these improvements have been primarily observed in resource-rich settings where abundant labelled data is available. In contrast, low-resource NER—scenarios with limited labelled data—continue to present substantial challenges that hinder performance Fritzler, Logacheva, and Kretov (2019). This section examines the primary challenges encountered in low-resource NER settings, focusing specifically on domain adaptation difficulties, entity boundary detection issues, and label inconsistency problems. The fundamental challenge in low-resource NER stems from the scarcity of labelled data, which impedes the model's ability to learn robust representations of entities and their contexts. According to Sun and Yang (2019), even state-of-the-art models like BERT demonstrate significant performance degradation when fine-tuned on limited data. This degradation manifests in various ways, including increased susceptibility to overfitting, reduced generalization capability, and heightened sensitivity to noise in the training data. These fundamental limitations set the stage for more specific challenges that are examined in the following subsections.

2.2.1 Domain Adaptation Challenges

Domain adaptation represents one of the most persistent challenges in low-resource NER. The difficulty primarily stems from the contextual variations that exist across different domains. Z. Liu et al. (2021) introduced CrossNER, a benchmark specifically designed to evaluate domain adaptation capabilities in low-resource settings. Their findings revealed that even state-of-the-art models experience dramatic performance drops when transferred across domains with limited in-domain training data. For instance, pre-trained transformers fine-tuned on the news domain exhibited a performance decline of up to 27% when applied to scientific texts, highlighting the severity of the domain adaptation challenge.

The root of this challenge lies in the domain-specific nature of entities and their contextual patterns. Jia and Zhang (2020) demonstrated that entities in specialized domains often follow unique linguistic patterns and appear in distinct contextual environments that differ significantly

from general domains. Their work on the Multi-Cell Compositional LSTM architecture revealed that conventional transfer learning approaches struggle to capture these domain-specific nuances when in-domain labelled data is scarce. This observation is consistent with their earlier findings Jia, Liang, and Zhang (2019), which established that linguistic patterns differ substantially across domains, creating a significant barrier to effective knowledge transfer. The domain adaptation challenge is further compounded by the presence of domain-specific entities that may not appear in the source domain at all (out-of-distribution entities). Models trained on general domains typically lack the necessary vocabulary and contextual understanding to recognize specialized entities in technical domains. For example, a model trained on news articles might easily recognize person names and locations but would struggle to identify chemical compounds or disease names in biomedical texts. This vocabulary mismatch exacerbates the domain adaptation challenge in low-resource scenarios where limited examples of domain-specific entities are available for training.

2.2.2 Entity Boundary Detection

Accurate identification of entity boundaries presents another significant challenge in low-resource NER. While entity type classification has benefited considerably from transfer learning approaches, boundary detection remains problematic, particularly when training data is limited. Katiyar and Cardie (2018) provided a comprehensive analysis of this challenge, highlighting how models trained on limited data often struggle with entity span identification, even when they correctly identify the entity type. Their work on nested NER revealed that boundary detection errors account for a disproportionately large percentage of overall errors in low-resource settings, suggesting that boundary detection requires more extensive training data than entity type classification.

The boundary detection challenge becomes especially pronounced when dealing with complex entity structures, such as nested or overlapping entities. Tan, Qiu, Chen, Wang, and Huang (2020) found that conventional sequence labelling approaches often fail to accurately identify entity boundaries when multiple entities share tokens or when entities are embedded within other entities. Their boundary-enhanced neural span classification approach demonstrated improved performance but still highlighted the inherent difficulty of the task in low-resource settings.

An innovative perspective on this challenge was offered by X. Li et al. (2019), who considered

NER as a machine reading comprehension (MRC) task to address boundary detection issues. Their unified MRC framework showed promising results in low-resource scenarios by leveraging pre-trained language models' understanding of natural language questions. However, their work also confirmed that boundary detection errors remain a major source of performance degradation in low-resource settings. The authors noted that approximately 60% of all errors in their low-resource experiments were attributable to incorrect boundary detection, underscoring the significance of this challenge X. Li et al. (2019).

2.2.3 Label Inconsistency

The third major challenge in low-resource NER involves label inconsistency issues, which become particularly pronounced when alternative supervision strategies are employed to compensate for the scarcity of manually labelled data. Shang et al. (2018) investigated the use of domain-specific dictionaries as a form of weak supervision and found that dictionary-based approaches often introduce label noise due to inconsistent entity coverage and context-insensitive matching. Their analysis revealed that such inconsistencies can significantly undermine model performance, with false negatives (entities missed by the dictionary) and false positives (incorrect matches) contributing to noisy training signals.

The problem of label inconsistency is further worsened in partially annotated datasets, which are common in low-resource settings due to the cost constraints of comprehensive annotation. Mayhew, Chaturvedi, Tsai, and Roth (2019) directly addressed this issue and demonstrated that partial annotation introduces systematic biases in the training data, leading to inconsistent model behaviour. Their experiments showed that models trained on partially annotated data tend to exhibit highly variable performance across different entity types and contexts, reflecting the inconsistent nature of the underlying annotations.

Weak supervision approaches, while promising for low-resource scenarios, introduce their own set of label inconsistency challenges. Lison, Hubin, Barnes, and Touileb (2020) conducted a comprehensive study of weak supervision methods for NER without labelled data and found that label noise represents a significant obstacle to model performance. Their analysis revealed that different weak supervision sources often produce conflicting annotations for the same entities, creating

confusion during model training. The authors proposed ensemble-based approaches to mitigate these inconsistencies but noted that the fundamental challenge persists in extremely low-resource scenarios.

The label inconsistency problem extends beyond English to other languages with limited resources. Kruengkrai, Nguyen, Aljunied, and Bing (2020) examined low-resource NER in multiple languages and found that label sparsity and inconsistency problems are particularly severe in languages with complex morphological structures and limited annotated resources. Their joint sentence and token labelling approach demonstrated improved performance by leveraging sentence-level signals to compensate for token-level inconsistencies, but the authors acknowledged that label inconsistency remains a fundamental challenge in low-resource multilingual settings.

The challenges discussed in this section—domain adaptation difficulties, entity boundary detection issues, and label inconsistency problems—are deeply interconnected in low-resource NER. Domain adaptation challenges often exacerbate boundary detection issues, as models struggle to identify unfamiliar entity patterns in new domains. Similarly, label inconsistencies become more pronounced when adapting across domains with limited data, as annotation schemes and entity definitions may vary. Understanding these interconnections is crucial for developing effective solutions to low-resource NER problems and advancing the field toward more robust performance across diverse domains and languages.

2.3 Large Language Models for NER

The field of NER has undergone a significant transformation with the advent of large language models (LLMs). This section examines the evolution of LLM applications in NER, tracing developments from foundational work to recent advances in fine-tuning methodologies, output format design, prompt engineering, and zero-shot learning capabilities.

Traditional NER approaches relied heavily on supervised learning with manually annotated datasets, often utilizing sequence labelling architectures such as Conditional Random Fields (CRFs) and later BiLSTM-CRF models Z. Huang et al. (2015). However, the emergence of transformer-based language models, beginning with BERT and culminating in generative models like GPT, has

fundamentally altered this landscape. The introduction of GPT-2 by Radford et al. (2019) marked a significant milestone, demonstrating that generative pre-trained transformers could perform a variety of NLP tasks without task-specific architectures, thereby establishing a foundation for future work in generative approaches to information extraction tasks, including NER.

Following them, Brown et al. (2020) established the potential of large language models for few-shot learning across various NLP tasks. Their GPT-3 model demonstrated a remarkable ability to recognize entities with only a handful of examples, suggesting that the pre-training process encoded substantial knowledge about entities and their contextual patterns. While not specifically focused on NER, this work laid the theoretical foundation for few-shot approaches to entity recognition by showing that LLMs could rapidly adapt to new tasks through in-context learning without parameter updates.

The progression from task-specific models to general-purpose LLMs for NER represents a paradigm shift in how researchers and practitioners approach the challenge of identifying and classifying named entities in text. As Keraghel, Morbieu, and Nadif (2024) comprehensively documents in their survey, this transition has been characterized by increasing model capacity, architectural innovations, and novel training objectives. These developments have collectively enabled models to learn generalizable representations that transfer effectively across domains and languages, addressing longstanding challenges in NER research related to domain adaptation and cross-lingual transfer.

2.3.1 Instruction Tuning Approaches

Instruction tuning has emerged as a powerful technique for adapting LLMs to perform specific tasks like NER with minimal task-specific training data. The groundwork for this approach was established by Wei et al. (2021), who demonstrated that fine-tuning language models on a collection of datasets described via instructions substantially improves zero-shot performance on unseen tasks. Their work showed that instruction-tuned models could follow natural language instructions to perform NER without explicit training on the specific entity types or domains being targeted. This finding was particularly significant for NER, where the diversity of entity types and domains had historically necessitated separate models or extensive domain adaptation.

Building on this foundation, Chung et al. (2024) explored how instruction tuning scales with model size, finding that performance gains increase non-linearly with scale. Their work showed that for information extraction tasks, including NER, larger instruction-tuned models exhibit emergent capabilities not present in smaller variants, including the ability to handle complex entity definitions and nested entity structures. These scaling effects are particularly relevant for NER, where entity definitions can be ambiguous and contextually dependent.

The application of instruction tuning specifically to information extraction tasks was formalized by Wang et al. (2023) with their InstructUIE framework. This approach unifies various information extraction tasks, including NER, relation extraction, and event extraction, under a common instruction-following paradigm. By reformulating NER as a text generation task guided by natural language instructions, InstructUIE achieved state-of-the-art performance across multiple benchmarks while significantly reducing the need for task-specific annotations. A key innovation in this work was the demonstration that instruction tuning enables models to understand and apply complex entity type definitions from natural language descriptions alone, addressing a key limitation of traditional NER approaches that rely on fixed entity type inventories.

The effectiveness of instruction tuning for NER can be attributed to several factors. First, instructions provide a flexible mechanism for specifying entity types and annotation guidelines, allowing models to adapt to new domains and entity types without architectural modifications. Second, instruction tuning leverages the linguistic knowledge encoded in LLMs' parameters, enabling them to recognize entities based on semantic understanding rather than pattern matching. Finally, the instruction-following paradigm aligns naturally with how human annotators approach NER tasks, potentially reducing the gap between model behaviour and human judgment.

2.3.2 Output Format Design: BIO tagging vs. Alternative Formats

The transition from traditional sequence labelling approaches to generative methods has necessitated a reconsideration of output formats for NER. Historically, the BIO (Beginning, Inside, Outside) tagging scheme and its variants have dominated supervised NER systems. However, LLM-based approaches have introduced alternative output formats that may better leverage the generative capabilities of these models.

The UniversalNER framework introduced by W. Zhou et al. (2023) represents a significant advancement in output format design for LLM-based NER. Rather than adhering to the traditional BIO tagging scheme, UniversalNER employs a direct entity span generation approach, where models output entity mentions along with their types in a structured text format. This approach aligns more naturally with the generative capabilities of LLMs and enables more flexible entity type specifications. Through targeted distillation from larger language models, UniversalNER demonstrates that this output format facilitates better knowledge transfer and generalization to unseen entity types.

A critical examination of output formats for generative NER was conducted by Y. Ding et al. (2024), who specifically addressed the challenge of handling negative instances—texts that contain no entities of interest. Their research revealed that conventional span-based output formats often struggle with precision because they lack an explicit mechanism for indicating the absence of entities. To address this limitation, they proposed a novel format that explicitly distinguishes between positive and negative instances, significantly improving precision without sacrificing recall. This work highlights the importance of output format design in balancing precision and recall in generative NER approaches.

Further advancing output format research, Zaratiana, Tomeh, Holat, and Charnois (2023) introduced GLiNER, a generalist approach to NER that employs a bidirectional transformer architecture to address the computational limitations of conversational approaches. Unlike traditional turn-by-turn conversational methods that require separate queries for each entity type, GLiNER treats NER as a matching problem between entity type embeddings and textual span representations in latent space. The model takes entity types as prompts alongside the input text, using special <ENT>tokens to separate different entity types, and computes similarity scores between entity representations and all possible text spans simultaneously.

The evolution of output formats for LLM-based NER reflects a broader trend toward more flexible and expressive annotation schemes that can capture the complexity of real-world entity mentions. As generative approaches continue to mature, we can expect further innovations in output format design that balance the competing objectives of expressiveness, parsability, and alignment with human annotation practices.

2.3.3 Prompt Engineering Strategies for NER

Prompt engineering has emerged as a crucial technique for effectively utilizing LLMs for NER tasks. The design of prompts—the instructions or examples provided to a model—significantly impacts performance, particularly in few-shot and zero-shot scenarios where task-specific fine-tuning may be limited or infeasible.

Reference Sainz et al. (2023) demonstrated that incorporating annotation guidelines directly into prompts substantially improves zero-shot NER performance. Their GoLLIE framework shows that explicitly stating the definition of entity types, providing annotation criteria, and including disambiguation rules enables LLMs to more accurately identify entities in line with human annotator expectations. This approach reduces ambiguity and aligns model outputs with intended annotation standards, addressing a key challenge in deploying LLMs for NER in specialized domains.

Building on similar principles, Zamai et al. (2024) explored the optimal balance between examples and instructions in prompts for zero-shot NER. Their "Show Less, Instruct More" methodology reveals that detailed definitional information and guidelines generally outperform example-heavy prompts when working with capable LLMs. This finding challenges the conventional wisdom that exemplars are the most effective way to guide model behaviour and suggests that LLMs can effectively internalize and apply explicit rules for entity recognition when properly instructed. The authors demonstrate that this approach is particularly effective for specialized entity types that might be underrepresented in the LLM's pre-training data. M. Zhang, Yan, Zhou, and Qiu (2023) introduced PromptNER, which combines retrieval-based and generation-based approaches through k-nearest neighbour search. This method dynamically constructs prompts by retrieving similar examples from a small annotated pool based on input text similarity, then uses these examples to guide the LLM's entity recognition process. By selecting the most relevant examples for each input, PromptNER achieves superior performance compared to static prompting strategies, particularly in few-shot scenarios. This work highlights the potential of hybrid approaches that combine the strengths of retrieval and generation for NER tasks.

The evolution of prompt engineering strategies for NER reflects growing sophistication in how

researchers interact with LLMs. Initial approaches relied heavily on rigid templates and numerous examples, while more recent work emphasizes clear definitions, explicit guidelines, and dynamically selected demonstrations. This progression aligns with a deeper understanding of how LLMs process and apply instructions, moving from treating them as black-box systems to be coaxed through examples toward viewing them as reasoning systems that can follow explicit guidelines.

2.3.4 Parameter-Efficient Fine-tuning and Model Adaptation

As LLMs have grown in size, parameter-efficient fine-tuning methods have become essential for adapting these models to specific NER tasks without prohibitive computational requirements. These approaches enable customization of general-purpose LLMs for specialized NER applications while updating only a fraction of the model parameters.

A pivotal development in this area was the introduction of Low-Rank Adaptation (LoRA) by E. J. Hu et al. (2022). LoRA drastically reduces the number of trainable parameters by representing weight updates as low-rank matrices, making it feasible to fine-tune billion-parameter models on consumer hardware. For NER applications, LoRA enables adaptation to domain-specific entity types and annotation standards without the cost of full model fine-tuning. This breakthrough has democratized access to custom LLM-based NER systems, allowing researchers and organizations with limited computational resources to develop specialized entity recognizers.

The release of open-weight models like the Llama series by Touvron et al. (2023) has further accelerated research on parameter-efficient adaptation for NER. These models provide high-quality pre-trained weights that can serve as starting points for customization while allowing full transparency about model capabilities and limitations. The availability of such models has enabled extensive experimentation with different adaptation techniques for NER, leading to improved understanding of how pre-trained knowledge can be leveraged for entity recognition across diverse domains.

Building on this foundation, Grattafiori et al. (2024) detailed the advancements in the Llama 3 model family, which demonstrates strong performance on NER tasks with minimal fine-tuning. These models incorporate architectural improvements and training methodologies that enhance their ability to understand and follow instructions about entity recognition, resulting in superior zero-shot

and few-shot performance compared to earlier generations. The reduced need for extensive fine-tuning makes these models particularly valuable for rapidly developing NER capabilities for new applications.

Recent model releases like Grattafiori et al. (2024) with Llama-3.3-70B-Instruct, Yang et al. (2024) with Qwen2.5, and the Falcon 3 family of models represent the current state of the art in adaptable LLMs for information extraction tasks. These models incorporate advanced instruction tuning methodologies that enable them to understand complex entity definitions and annotation guidelines with minimal additional training. Their multilingual capabilities also facilitate cross-lingual transfer for NER, addressing the historically challenging problem of developing entity recognizers for low-resource languages.

As LLMs continue to evolve, the trend toward parameter-efficient adaptation and instruction tuning suggests a future where highly customizable NER systems can be rapidly deployed across domains and languages with minimal annotation effort. This evolution promises to address many of the historical limitations of NER technology while enabling new applications that require fine-grained entity understanding.

2.4 Data Augmentation for NER

The challenge of data scarcity in NER has driven considerable research into augmentation techniques that can enhance model performance while preserving the integrity of entity labels. This section examines the evolution of data augmentation approaches for NER, from early token-level manipulations to sophisticated language model-based generation methods, establishing the theoretical foundation that informs our proposed PANER framework.

2.4.1 Traditional Data Augmentation Techniques

The earliest investigations into NER data augmentation emerged from broader efforts to address data scarcity in natural language processing tasks. X. Zhang, Zhao, and LeCun (2015) pioneered the application of systematic lexical replacement for character-level convolutional neural networks,

introducing the foundational principle that controlled perturbations could improve model generalization without compromising semantic integrity. Their approach employed geometric distributions to determine both the number of words to be substituted and the selection criteria for replacement candidates, establishing that probabilistic selection mechanisms could effectively balance augmentation diversity with semantic preservation.

Building upon these foundational principles, the machine translation community provided crucial insights that would later influence NER augmentation strategies. Fadaee, Bisazza, and Monz (2017) addressed the specific challenge of rare word handling in low-resource neural machine translation by developing a targeted augmentation approach that generated new sentence pairs containing low-frequency words in synthetically created contexts. Their methodology achieved improvements of up to 2.9 BLEU points over baseline approaches, showing that selective augmentation targeting specific linguistic phenomena could yield substantial performance gains.

The emergence of systematic token-level augmentation techniques was the next advancement in the field. Wei and Zou (2019) introduced Easy Data Augmentation (EDA), a comprehensive framework comprising four fundamental operations: synonym replacement, random insertion, random swap, and random deletion. While originally designed for sentence-level classification tasks, EDA's systematic approach demonstrated particular effectiveness for datasets with fewer than 500 samples, achieving accuracy improvements of up to 3% with 16 augmented sentences per input. The significance of EDA extends beyond its immediate performance gains; it established the principle that simple, interpretable augmentation operations could compete with more complex generative approaches.

However, the application of general-purpose augmentation techniques to NER revealed unique challenges that demanded specialized solutions. Dai and Adel (2020) conducted the first comprehensive analysis of simple data augmentation techniques specifically adapted for NER tasks, systematically comparing label-wise token replacement (LwTR), mention replacement (MR), and synonym replacement (SR) approaches across biomedical and materials science domains. Their work demonstrated that simple augmentation could boost performance for both recurrent and transformer-based models, particularly in small training set scenarios, while highlighting the critical importance of maintaining label consistency during augmentation. The LwTR approach, which replaces tokens

with others sharing the same entity label, inspires our entity masking strategy, where type-specific placeholders preserve semantic relationships while enabling contextual variation.

Contemporary work in this period also explored domain-specific adaptations that addressed the unique characteristics of specialized NER tasks Issifu and Ganiz (2021). The medical domain, in particular, presented challenges related to terminology precision and entity relationship complexity that required careful consideration during augmentation. These domain-specific investigations concluded that successful NER augmentation must balance linguistic diversity with semantic precision.

The evolution of traditional augmentation techniques established several key principles that continue to influence modern approaches: the importance of maintaining semantic coherence, the effectiveness of targeted rather than random modifications, and the need for entity-aware augmentation strategies.

2.4.2 LLM-Based Data Generation and Augmentation

The advent of large language models fundamentally transformed the landscape of data augmentation for NER, introducing unprecedented capabilities for generating coherent, contextually appropriate synthetic text. This paradigm shift from rule-based transformations to generative methods enabled more sophisticated augmentation strategies that could preserve semantic relationships while introducing meaningful linguistic variation.

R. Zhou et al. (2022) introduced MELM (Masked Entity Language Modelling), a pioneering approach that injects entity labels into training contexts to reduce token-label misalignment while improving entity diversity in low-resource and multilingual NER settings. MELM's methodology involved masking entity mentions in sentences and training language models to predict contextually appropriate entities, thereby generating augmented data that maintains semantic coherence while introducing entity-level variation. This approach demonstrated consistent improvements across multiple languages and domains, establishing the viability of entity-focused language modelling for NER augmentation.

The integration of back-translation techniques with NER-specific considerations was introduced by Yaseen and Langer (2021). Their work highlighted both the potential and limitations of translation-based augmentation for token-level tasks, particularly emphasizing the need for careful

entity preservation mechanisms during the translation and back-translation process. While back-translation can introduce beneficial linguistic diversity, the authors noted challenges in maintaining entity consistency across language boundaries, limitations that we plan to address with our PANER framework.

The development of entity-controlled synthetic text generation represented a bridge between traditional rule-based methods and modern neural approaches. Aggarwal, Jin, and Ahmad (2023) developed entity-controlled generation techniques using contextual question-answering with pre-trained language models, focusing on maintaining entity consistency while generating diverse contexts. Their approach demonstrated that large language models could be effectively guided to produce entity-aware synthetic text, establishing the foundation for more sophisticated control mechanisms in neural generation.

A significant breakthrough came with the introduction of unified generative augmentation approaches capable of handling multiple NER task variants. X. Hu et al. (2022) introduced EnTDA (Entity-to-Text based Data Augmentation), proposing the first unified framework capable of addressing flat, nested, and discontinuous NER tasks through a novel entity-to-text generation paradigm. By decoupling entity dependencies through add, delete, replace, and swap operations on entity lists, EnTDA addressed fundamental limitations of previous text-to-entity approaches that struggled with complex entity structures. The method incorporated diversity beam search to increase generation variety while maintaining semantic coherence, demonstrating substantial improvements across thirteen NER datasets.

The most recent advances in LLM-based augmentation have focused on addressing the specific challenges of few-shot NER. Ye et al. (2024) presented LLM-DA, an approach employing 14 contextual rewriting strategies with entity replacement and noise injection specifically designed for few-shot NER scenarios. This work directly confronts the limitations of existing methods that compromise semantic integrity, leveraging large language models' distinctive rewriting capabilities while addressing uncertainty inherent in generated text. LLM-DA consistently outperformed ChatGPT across multiple datasets, demonstrating that targeted utilization of LLMs for NER augmentation could achieve superior results compared to general-purpose language model applications. The contextual rewriting strategies employed in LLM-DA inform our paraphrasing methodology in

PANER, particularly in the systematic approach to preserving entity relationships while introducing linguistic variation. However, we do not hard-code the rewriting strategy since that can heavily restrict the model's ability when switching domains.

The evolution toward LLM-based augmentation reflects a change in understanding how generative methods can be leveraged for NER enhancement. Unlike earlier approaches that relied on simple transformations or external translation services, modern LLM-based methods can generate contextually appropriate text that maintains complex semantic relationships while introducing beneficial diversity. This capability is particularly relevant for specialized domains where entity relationships are nuanced and context-dependent.

However, the application of LLMs to NER augmentation also introduced new challenges related to controllability and consistency. While large language models demonstrate remarkable generation capabilities, ensuring that generated text maintains entity boundaries and type consistency requires sophisticated prompting and validation strategies. Early approaches treated entities as isolated tokens to be preserved during augmentation, while recent work recognizes the complex semantic networks that connect entities to their surrounding contexts.

The synthesis of insights from both traditional and LLM-based approaches informs every aspect of our paraphrasing framework design. The entity masking strategy draws from the label consistency principles established by Dai and Adel (2020), while the controlled generation approach builds upon the contextual rewriting strategies demonstrated by Ye et al. (2024). The type-specific placeholder system addresses the entity preservation challenges identified in back-translation research, while the validation mechanisms ensure the semantic coherence emphasized throughout the literature.

2.5 Summary

The literature examined in this chapter traces the evolution of NER from traditional supervised approaches to modern large language model implementations, establishing the theoretical foundation for our proposed PANER methodology. As demonstrated in Table 2.1, this progression reveals a clear trajectory from high-resource supervised methods requiring extensive manual engineering to

Approach	Training type/Data Requirements	Parameters	Key Innovation
Traditional Supervised Era			
HMM Bikel et al. (1998)	Supervised, high-resource	-	Sequential probabilistic modeling
MaxEnt Borthwick (1999)	Supervised, high-resource	-	Overlapping feature integration
CRF Lafferty et al. (2001)	Supervised, high-resource	-	Global sequence optimization
Neural Network Era			
Bi-LSTM-CRF Z. Huang et al. (2015)	Supervised, high-resource	<100M	Bidirectional context with structured output
Char-CNN X. Zhang et al. (2015)	Supervised, high-resource	<50M	Subword morphological representations
Transfer Learning Era			
BERT Devlin et al. (2019)	Fine-tuning, medium-resource	110M-340M	Bidirectional transformer pretraining
RoBERTa Y. Liu et al. (2019)	Fine-tuning, medium-resource	125M-355M	Optimized pretraining methodology
LLM-based Approaches			
InstructUIE Wang et al. (2023)	Instruction tuning, low-resource	11B	Unified multi-task instruction framework
UniversalNER W. Zhou et al. (2023)	Distillation, zero/few-shot	7B	Large-scale knowledge distillation
GoLLIE Sainz et al. (2023)	Guidelines, zero-shot	7B	Code-structured annotation guidelines
GNER Y. Ding et al. (2024)	Instruction tuning, few-shot	7B-11B	Negative instance incorporation
GLiNER Zaratiana et al. (2023)	Minimal training, zero/few-shot	300M	Bidirectional generalist architecture
SLIMER Zamai et al. (2024)	Guidelines, few-shot	7B	Definition-enriched prompting
PANER	**Augmented few-shot**	**7B-10B**	**Strategic paraphrase-based augmentation**

Table 2.1: Multi-Dimensional Taxonomy of NER Approaches

contemporary LLM-based approaches capable of effective entity recognition with minimal training data. This comprehensive analysis systematically examines three critical areas: the historical development of NER techniques, persistent challenges in low-resource scenarios, and emerging solutions through large language models and data augmentation strategies.

NER was introduced at the Sixth Message Understanding Conference in 1995 and defined the foundational entity categories of person, location, and organization, creating a clear trajectory from rule-based systems to sophisticated statistical models. HMMs initially demonstrated the potential of probabilistic frameworks, followed by Maximum Entropy models that addressed independence assumptions, and ultimately CRFs that dominated the field for nearly a decade by effectively handling sequential dependencies.

The transition to neural network-based approaches marked a paradigm shift, beginning with Collobert et al. (2011)'s demonstration that neural networks could learn representations directly from data, reducing manual feature engineering requirements. The introduction of Word2Vec and character-level CNNs further enhanced representation learning, culminating in the Bi-LSTM-CRF architecture that combined neural sequence modelling with structural prediction constraints. This progression established neural approaches as superior to traditional methods while highlighting the persistent challenge of requiring substantial task-specific training data.

The most recent evolution toward transfer learning and pre-trained language models represents the current state-of-the-art paradigm. ELMo's contextualized representations demonstrated the value of bidirectional language modelling, while BERT's transformer architecture and pre-training objectives achieved unprecedented performance levels. The subsequent development of RoBERTa, with its optimized training methodology, established new benchmarks and demonstrated how methodological refinements could yield substantial performance gains without architectural innovations.

Our review also identified three fundamental challenges that persist in low-resource NER settings. (1) Domain adaptation difficulties represent the most significant barrier, as models trained

on general domains experience dramatic performance degradation when applied to specialized contexts. The CrossNER benchmark reveals performance drops of up to 27% within the encoder-decoder and encoder-only model space when transferring across domains, highlighting the domain-specific nature of entity patterns and contextual environments. (2) Entity boundary detection emerges as a particularly problematic aspect of low-resource NER, with research indicating that boundary errors account for approximately 60% of all mistakes in limited-data scenarios Y. Ding et al. (2024). This challenge becomes especially pronounced with complex entity structures, such as nested or overlapping entities, where conventional sequence labelling approaches fail to maintain accuracy. (3) Label inconsistency problems represent the third major challenge, particularly when alternative supervision strategies are employed to compensate for limited manually labelled data. Dictionary-based weak supervision and partially annotated datasets introduce systematic biases and conflicting annotations that undermine model performance. These interconnected challenges collectively demonstrate why low-resource NER remains a persistent problem despite advances in neural architectures.

The application of Large Language Models to NER tasks represents a fundamental shift from task-specific architectures to general-purpose systems capable of following natural language instructions. This evolution progressed from GPT-2's demonstration of generative capabilities through the development of instruction tuning methodologies that enable models to perform NER without extensive task-specific training.

Instruction tuning approaches, particularly frameworks like InstructUIE and UniversalNER, demonstrate how natural language instructions can guide entity recognition across diverse domains and entity types. The evolution from rigid template-based approaches to sophisticated guideline integration, as exemplified by GoLLIE's incorporation of annotation guidelines, shows how explicit instruction can leverage models' latent entity knowledge. Prompt engineering strategies have evolved from example-heavy approaches to definition-rich instructions that explicitly communicate entity types and annotation criteria.

The evolution of data augmentation techniques specifically designed for NER progresses from traditional approaches through entity-aware strategies to sophisticated LLM-based generation methods. Traditional approaches, beginning with systematic lexical replacement and progressing through

Easy Data Augmentation (EDA), established fundamental principles of controlled perturbation that preserve semantic integrity while introducing beneficial diversity. The recognition that NER requires entity-aware augmentation strategies led to specialized techniques like label-wise token replacement and mention replacement that explicitly consider entity boundaries during transformation. This evolution reflects a growing understanding that successful NER augmentation must balance linguistic diversity with semantic precision. The advent of LLM-based data generation represents the current frontier, with approaches like MELM demonstrating entity-controlled generation and EnTDA introducing unified frameworks for complex entity structures. Recent work like LLM-DA has shown that targeted utilization of large language models for contextual rewriting can achieve superior results compared to general-purpose applications.

This comprehensive analysis establishes that while significant progress has been made in NER through increasingly sophisticated models and techniques, the fundamental challenges of low-resource scenarios—domain adaptation remain unresolved. The convergence of instruction tuning capabilities, parameter-efficient fine-tuning methods, and intelligent data augmentation strategies creates the foundation for our PANER framework, which synthesizes these advances to address persistent low-resource challenges through strategic paraphrase-based augmentation combined with an optimized instruction tuning methodology.

Chapter 3

Methodology

3.1 Introduction

This chapter provides an in-depth description of our methodology for approaching low-resource NER through the PANER (Paraphrase-Augmented Named Entity Recognition) framework. Section 3.2 outlines the comprehensive PANER framework diagrammatically and sets the rest of the section to explore each component in detail. Section 3.3 presents our framework for data augmentation, expanding on the process and the prompt used for generating diverse samples. In Section 3.4 we describe our instruction tuning template design, and explain our choices to diverge from traditional BIO tagging in favour of a simplified word/tag representation format while incorporating negative instances, detailed entity guidelines, and definitions.

3.2 Overview of PANER Framework

Named Entity Recognition (NER) remains a fundamental yet challenging task in Natural Language Processing (NLP), particularly in low-resource scenarios where annotated data is scarce. PANER (Paraphrase-Augmented Named Entity Recognition) is a comprehensive framework designed to address these challenges through two key contributions: strategic data augmentation and optimized instruction tuning.

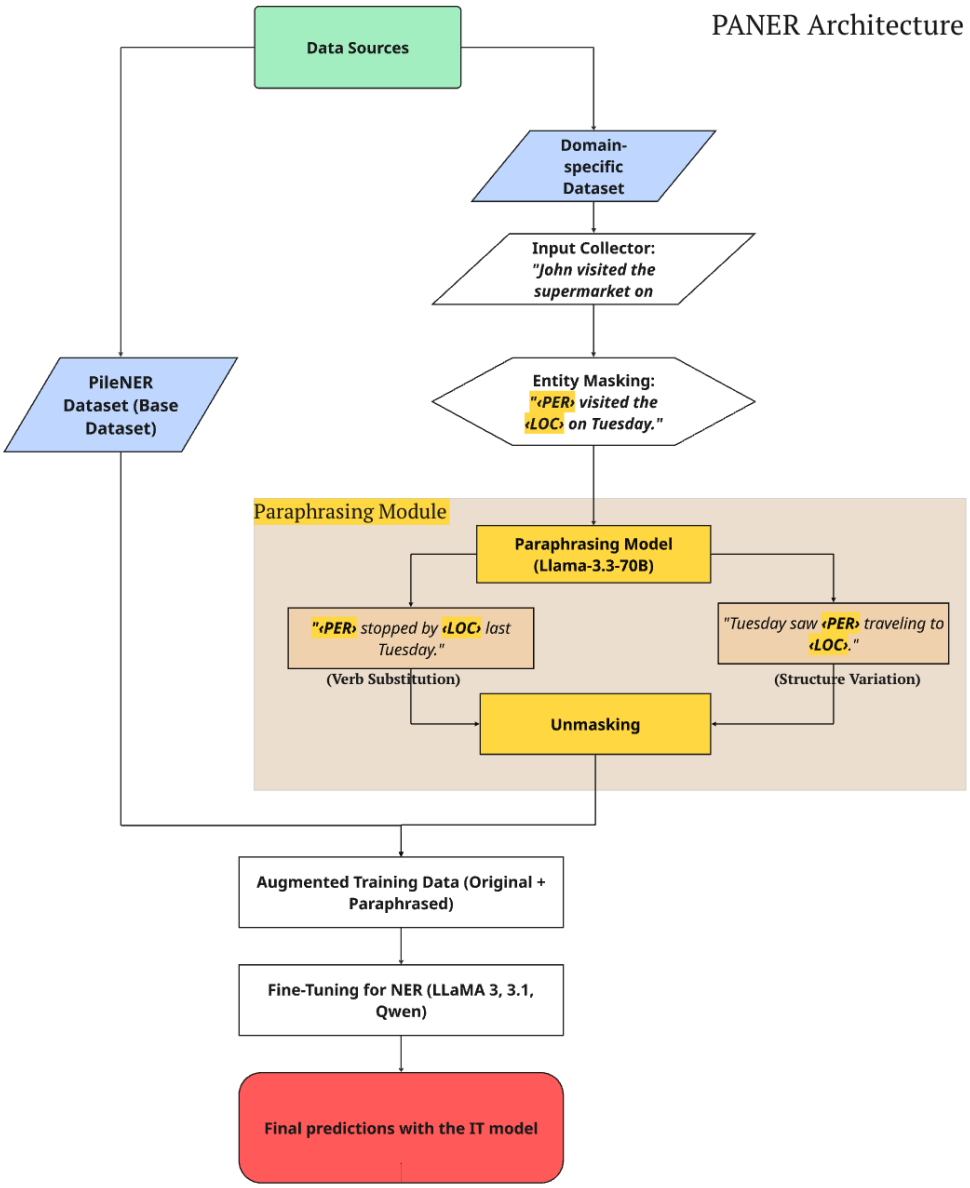

Figure 3.1: Flowchart of the PANER framework.

Figure 3.1 illustrates the overview of the PANER framework from a few-shot learning perspective, where we infuse domain-specific datasets and increase sample size through data augmentation. While this diagram captures the core data processing pipeline, the complete PANER system encompasses additional components, including instruction tuning templates that contain annotations and guidelines, and prompt engineering strategies that operate as complementary elements and will be detailed in subsequent sections.

We present an integration of paraphrase-based data augmentation techniques with refined instruction tuning methodologies, specific to low-resource NER environments. PANER's architecture comprises two primary components that work in synergy:

(1) a controlled paraphrasing mechanism that preserves entity information while diversifying contextual representations, and

(2) an enhanced instruction tuning template that combines principles from prior approaches while implementing a simplified output format.

Unlike conventional NER approaches that require extensive annotated datasets for each target domain, we have designed PANER to operate effectively with minimal labelled data. This efficiency is achieved through our generation of high-quality synthetic samples that maintain semantic integrity while introducing linguistic variation. The framework leverages the extended context windows of modern LLMs, including Qwen-2.5-Instruct (7B), LLAMA-3.1-Instruct (8B), Falcon3-Instruct (10B) and Llama 3.3-70B for both the paraphrasing process and the subsequent entity recognition tasks. These models, with context windows ranging from 32k to 128k tokens, allow the framework to process longer and more complex inputs while maintaining coherent entity recognition. Thus establishing a resource-efficient pipeline that remains accessible to researchers and organizations with limited computational capabilities.

At its core, PANER addresses a critical gap in current NER methodologies by combining the strengths of recent instruction-tuning approaches such as SLIMER's guideline-centric philosophy Zamai et al. (2024) and GNER's negative instance inclusion Y. Ding et al. (2024). This integration enables the framework to better handle domain-specific entities while reducing dependency on extensive labelled datasets. Furthermore, the implementation of a word/tag output format rather than the traditional BIO tagging schema simplifies the annotation process, while also improving

performance metrics on standard benchmark datasets (discussed more in 5).

Through this integration of paraphrase-based data augmentation and optimized instruction tuning, PANER establishes a flexible and efficient framework for addressing the challenges of low-resource NER. The subsequent sections will provide a detailed exploration of each component, including the paraphrasing methodology, instruction tuning approach, and implementation considerations (detailed in Chapter 4) that collectively form the complete PANER framework.

3.3 Paraphrasing Framework for Data Augmentation

The paraphrasing framework represents one of the two main contributions in this paper, addressing the critical challenge of data scarcity in low-resource NER. Unlike traditional data augmentation techniques that often introduce noise or semantic inconsistencies, PANER's approach preserves entity information while generating linguistically diverse contexts, thereby enhancing model generalization without compromising semantic integrity.

3.3.1 Entity Masking and Semantic Preservation

The entity masking process transforms input sentences into masked templates where named entities are replaced with semantic placeholders corresponding to their entity types. Consider a sample input sentence: "Microsoft announced new products in Seattle yesterday." We generate a masked version: "<ORG> announced new products in <LOC> yesterday." Unlike generic masking approaches that use uniform placeholders, PANER employs type-specific tags (e.g., <PER>, <LOC>, <ORG>), providing the paraphrasing model with essential context about the entity type being masked. This design choice proved critical for generating high-quality paraphrases, as our experimental iterations revealed that using generic <ENTITY> tags provided insufficient contextual guidance to the language model. Type-specific tags enable the paraphrasing model to understand the semantic role of each entity within the sentence structure, allowing it to generate contextually appropriate variations while maintaining entity integrity. For instance, knowing that <ORG> represents an organization entity and <LOC> represents a location enables the model to produce semantically coherent paraphrases such as "<ORG >revealed new products in <LOC >yesterday"

or "Yesterday, <ORG> launched products in <LOC>," where the organization-location relationship and temporal context remain logically consistent across all generated variants. A very basic example of the process is illustrated in Figure 3.2.

An important optimization involves the handling of consecutive entity tags. In early experiments, we assigned individual masks to each word within multi-worded entities (e.g., a four-word organization name). However, this method significantly increased the complexity of the paraphrasing task and often resulted in inconsistent output during generation. The paraphrasing model struggled to preserve entity boundaries and frequently introduced an inconsistent number of masks in the output sentence. To mitigate these issues, we pivoted to a consolidated tagging scheme that merges adjacent words of the same entity type into a single entity mask. This modification simplifies the input structure and improves the paraphrasing model's ability to maintain entity consistency throughout the process.

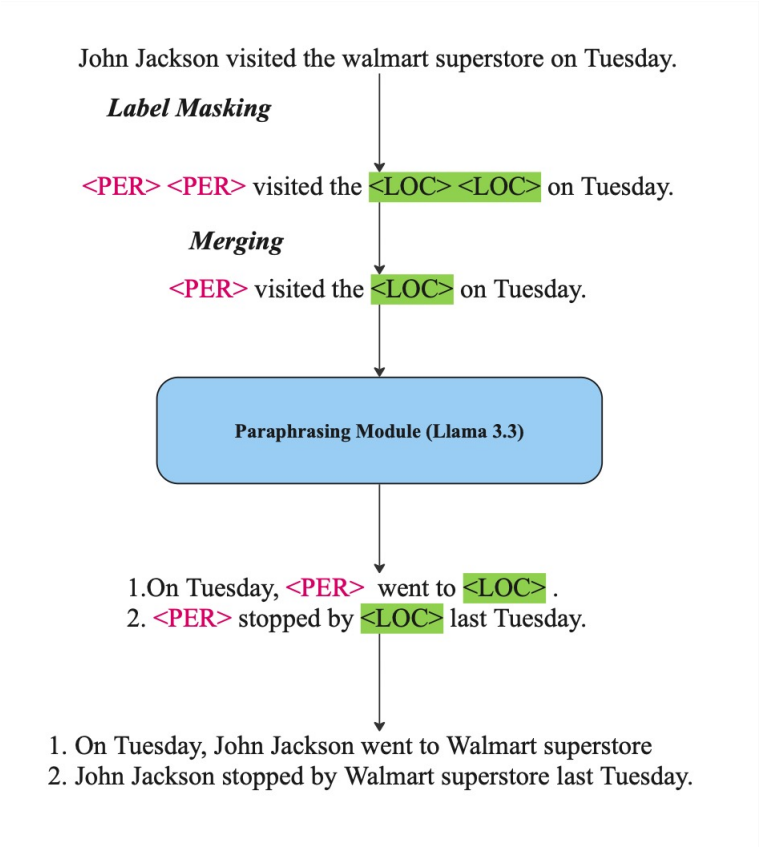

Figure 3.2: Illustration of a paraphrasing-based data augmentation process.

Paraphrasing Prompt
Task Description:
You are a helpful assistant. I have a sentence with certain entities that I want to preserve in spirit, but you may modify the sentence slightly to add variety. Your task is:

(1) Read the Original Sentence provided.

(2) Create 2 new sentences (variants) that:

- DO NOT MODIFY any word enclosed in <<>> tags or move them around (do not introduce any new <<>> tags that weren't in the original).
- May adjust phrasing, structure, or add contextual details while maintaining logical coherence and meaning.
- Minor modifications are allowed, but retain the core entity references and do not transform them into something else.

(3) Return the output in a valid JSON format with the generated variants.

Original Sentence: *Input*

Figure 3.3: Prompt used for generating paraphrases.

3.3.2 Paraphrase Generation Techniques

Building upon the entity-masked templates, PANER employs a simple paraphrase generation technique to balance diversity with sentence coherence (an example of the type of generation done can be seen in Fig. 3.1). The generation process leverages the LLAMA 3.3-70B model, selected for its optimal balance of performance and efficiency relative to larger models while operating with substantially fewer parameters (70B vs 405B), making it more practical for real-world applications.

The paraphrasing prompt, shown in Figure 3.3, provides explicit instructions for the language model to maintain entity integrity while introducing contextual variations.

Through experimental validation (Section 5.7), we determined that generating two paraphrased versions per input sentence achieves the optimal balance between diversity and quality. Attempts to produce three or more variations frequently resulted in redundancy or phrases not adhering to requested output format. Additionally, generating more variants often led to inconsistencies in the number of entity tags compared to the input sentence.

Using our approach with the masked example "<ORG>hired <PER>as <MISC>" yields variations such as:

- "<ORG>appointed <PER>as <MISC>"
- "<PER>was recruited by <ORG> for the <MISC> position"

Each variant maintains the entity relationships of the original sentence while introducing natural diversity in structure and word choice.

A critical component of our paraphrasing framework is a robust quality control and validation pipeline developed using the instructor package J. Liu (2024). Unlike structured models that inherently produce constrained outputs, generative large language models have the flexibility to produce unlimited textual variations, which, while advantageous, can also lead to inconsistent formatting or outputs that deviate from the required specifications. To constrain this ability while maintaining control over output quality and format, we implement structured constraints that guide the LLM's responses into predetermined formats suitable for automated processing. We process the model's output in JSON format through the instructor package (implementation shown in the Appendix A), which enforces schema validation and ensures consistent structure across all generated paraphrases. This constraint mechanism is essential for preventing the model from producing nonsensical outputs or deviating from the specified paraphrase generation task, while simultaneously enabling efficient parsing and validation of the generated samples without requiring manual inspection of individual examples. For each paraphrase, we verify that:

(1) The number of entity tags matches the input sentences.

(2) These sentences meet a minimum cosine threshold(0.6) to preserve semantic relevance.

When a generated paraphrase fails these validation checks, we rerun the generation with adjusted model parameters like temperature, top-p, etc. This validation process helps maintain the integrity of the augmented dataset while allowing for natural variations in sentence structure and word choice.

The generation process remains configurable, allowing users to generate additional variations by adjusting the temperature parameter during generation. For applications requiring greater sample

diversity, larger and more capable models can be employed to generate three or more variations per input sentence, potentially incorporating directional prompts that guide specific types of linguistic modifications (e.g., formal-to-informal tone shifts, syntactic restructuring, or domain-specific vocabulary substitutions). This flexibility enhances PANER's adaptability across different NER tasks and domains, making it a versatile solution for low-resource scenarios that can scale with available computational resources and specific augmentation needs.

The generation process remains configurable, allowing users to generate additional variations by adjusting the temperature parameter during generation. For applications requiring greater sample diversity, larger and more capable models can be employed to generate three or more variations per input sentence, potentially incorporating directional prompts that guide specific types of linguistic modifications (e.g., formal-to-informal tone shifts, syntactic restructuring, or domain-specific vocabulary substitutions). This scalable approach accommodates varying computational budgets and performance requirements, where organizations with access to more powerful models can achieve enhanced augmentation diversity, while those with limited resources can still benefit from the two-variant configuration demonstrated in our experiments. This flexibility enhances PANER's adaptability across different NER tasks and domains, making it a versatile solution for low-resource scenarios that can scale with available computational resources and specific augmentation needs.

3.4 Instruction Tuning Template Design

The instruction tuning template design in PANER is the second contribution we are proposing, combining insights from recent research while addressing their limitations. Instruction tuning has emerged as a leading paradigm for adapting large language models to downstream natural language tasks, and the capabilities of LLMs particularly benefit low-resource scenarios. However, one of the problems with using LLMs for NER is that they often struggle with complex entity boundaries, domain adaptation, and computational efficiency.

3.4.1 Simplified Tagging Formats

Traditional supervised NER approaches typically employ the BIO (Beginning, Inside, Outside) tagging scheme, which distinguishes between the beginning and continuation of entity spans. While effective for many supervised learning scenarios, this scheme introduces unnecessary complexity in the context of instruction-tuned language models, particularly for few-shot and zero-shot settings.

Our approach revisits and refines the instruction-tuning methodologies by diverging from the traditional BIO tagging schema in favour of a simplified word/tag representation format. This format annotates each word with its corresponding entity tag using a forward slash (/) separator, as shown in the prompt (Figure 3.5). By removing the distinction between "B-" and "I-" labels, we want the model to focus on entity type classification rather than complex boundary detection (which will benefit from the inclusion of negative instances).

We acknowledge the limitations inherent in this design choice. Specifically, eliminating the "B-" and "I-" tags results in the loss of our ability to distinguish between start and end spans of consecutive entities of the same type. A common example would be citations of multiple author names in academic papers, where traditional BIO tagging would differentiate between separate consecutive person entities. We address this limitation by relying on punctuation marks, conjunctions, and other sentence structures to differentiate between distinct entities. While we understand that this approach cannot differentiate entities in all cases, our decision is based on analysis of the datasets used in this paper, where sufficient lexical and syntactic cues are present to support this direction. We provide a more detailed discussion of these limitations and their implications in Section 6.3.2.

Our experimental results demonstrate that the simplified word/tag format consistently outperforms the BIO-based approaches across diverse domains. For example, on the CrossNER datasets, our approach achieved an average F1 score of 67.13 without guidelines, compared to 51.92 for the BIO-based format (explained in detail in Section 5.1). This significant improvement underscores the effectiveness of simplifying the tagging schema for instruction-tuned models.

Refined Instruction Tuning Approach Our refined instruction tuning approach incorporates two key design elements beyond the simplified word/tag output format: (1) the integration of detailed definitions and guidelines for each entity type, and (2) the inclusion of negative instances in the

input/output format. These design choices are grounded in recent empirical findings and are specifically engineered to complement each other in addressing the challenges of low-resource NER.

Domain-Specific Entity Guidelines and Definitions

AI Domain - ALGORITHM:
Definition: 'Algorithm' entities refer to specific computational procedures or methods designed to solve a problem or perform a task within the field of computer science or related disciplines.
Guidelines: Avoid labelling generic technology or software names without specific algorithmic context. Exercise caution with terms that may denote both a specific algorithm and a generic concept, such as 'neural network'.

Literature Domain - LITERARY_GENRE:
Definition: 'Literary genre' refers to a category or type of literary composition characterized by a particular style, form, or content, such as novel, poetry, science fiction, or picaresque.
Guidelines: Avoid labelling general literary terms like 'book' or 'writing'. Consider the specific stylistic, structural, or thematic elements that define a genre, and be mindful of overlapping genres, such as historical fiction or science fiction thriller.

Music Domain - MUSICAL_ARTIST:
Definition: 'Musical artist' refers to individuals primarily known for their contributions to the field of music, including singers and musicians.
Guidelines: Exercise caution with terms that may have multiple meanings, such as 'John Denver' (could refer to a musical artist or a geographic location). Be mindful of context, as not all mentions of names in a musical context necessarily indicate a musical artist (e.g., 'Novoselic' might refer to a person but not necessarily a musical artist).

Figure 3.4: Examples of domain-specific entity definitions and guidelines across AI, Literature, and Music domains.

3.4.2 Definitions and Guidelines

We adopt the definition and guideline framework pioneered by SLIMER Zamai et al. (2024), which demonstrated significant improvements in zero-shot NER performance, particularly for unseen entity types. The SLIMER approach addresses a fundamental limitation in existing instruction-tuned NER models: while models like UniNER W. Zhou et al. (2023) and GNER Y. Ding et al. (2024) achieve high performance on entity types seen during training (95-100% overlap with test sets), they struggle with truly unseen entities.

SLIMER's approach of enriching prompts with entity-specific definitions and annotation guidelines yielded substantial improvements, with their model achieving competitive performance while being trained on only a fraction of the data used by other approaches. The effectiveness stems from providing explicit semantic context that enables models to generalize beyond memorized entity patterns. For instance, when evaluating on never-seen-before entities in the BUSTER dataset, SLIMER achieved an F1 score of 45.3, significantly outperforming models like GNER-T5 (27.9) and UniNER (37.8) (We ran this experiment with our approach; the results can be found in Section 5.3.2).

The definitions and guidelines are generated using GPT-3.5-turbo by selecting three examples from the dataset for each entity type and fed to the model with a carefully designed prompt template (detailed in Appendix A.2) that produces consistent, structured definitions and guidelines. Figure 3.4 illustrates examples of generated definitions and guidelines across different domains.

3.4.3 Incorporation of Negative Instances

Our second design choice builds upon GNER's empirical finding that explicit annotation of non-entity tokens significantly improves NER performance Y. Ding et al. (2024). GNER's comprehensive analysis revealed that "negative instances contribute to remarkable improvements by introducing contextual information, and delineating label boundaries."

The GNER framework demonstrated through controlled experiments that traditional entity-centric approaches, which focus only on entity portions during training, overlook crucial contextual information provided by non-entity text. Their systematic evaluation using three error metrics—Unlabeled Error (UE), Noisy Error (NE), and Boundary Error (BE)—showed clear improvements when incorporating negative instances:

Context	**UE (%)**	**NE (%)**	**BE (%)**	**F1 (%)**
Without context (entity-centric)	7.8	16.7	3.8	59.0
With full context	7.7	14.9	3.7	61.0

Table 3.1: Results taken from *the GNER paper* Y. Ding et al. (2024). This table compares performance metrics (UE, NE, BE, F1) with and without the non-entity tokens.

This proved that incorporating contextual information around entities led to progressive performance improvements. The effectiveness of negative instances is particularly pronounced in their dual role for precision and recall improvement. As Y. Ding et al. (2024) explains: "The context surrounding an entity often leads to a more accurate determination of its type, and the model is guided to make judgments on every token in a sentence (including those in non-entity texts), which helps recall more entities.".

3.4.4 Integrated Framework Design

Our approach combines these two evidence-based strategies within the simplified word/tag format. The complete instruction tuning template, shown in Figure 3.5, integrates task description, entity-specific definitions and guidelines, and explicit negative instance mentions in our training data and simplified output format.

This integration addresses complementary aspects of the NER challenge: while definitions and guidelines provide semantic clarity for entity type determination, negative instances enhance boundary detection and contextual understanding. When applying this new instruction prompt to the CrossNER Z. Liu et al. (2021) Science dataset with 16 entity types in the training prompt, including task description, annotations, and guidelines for all named entities, added up to 1700 tokens, well below the context length of the models used in our experiments. Our experimental validation confirms the effectiveness of this integration.

3.5 Summary

This chapter presented PANER, a comprehensive framework for low-resource NER that addresses limited annotated data through strategic data augmentation and optimized instruction tuning. The methodology centers on two key innovations: (1) a paraphrase-based data augmentation technique that preserves entity information while modifying surrounding non-entity tokens, and (2) a refined instruction tuning template combining recent approaches with a simplified output format. Section 3.3 introduces a controlled approach to generating synthetic training examples through entity masking and context modification. By preserving entity boundaries while introducing linguistic

Instruction Tuning Prompt

Task Description:
Please analyze the sentence provided, identifying the type of entity for each word on a token-by-token basis. Each word in the sentence should be annotated with its corresponding named entity tag, using a forward slash / between the word and the tag. Output format is: word_1/label_1, word_2/label_2, ...
Guideline:
1) Use O for words that are not part of any named entity.
2) For multi-word entities, label each word with the same entity tag.

Use the specific entity tags: $l_1, l_2, \ldots, l_m$, and O.

To help you, here are dedicated DEFINITION and GUIDELINES for each entity tag.
{ l_1: {
DEFINITION : ,
GUIDELINES : }
}

Input: $x_1 x_2 \ldots x_n$
Output: $x_1/\hat{y}_1 x_2 \hat{y}_2 \ldots x_n/\hat{y}_n$

Figure 3.5: Prompt used for Instruction-tuning LLMs

variation, this technique expands training datasets without compromising recognition accuracy. The quality control pipeline ensures generated paraphrases are without errors and maintain semantic relevance.

Section **??** introduced an instruction tuning template that diverges from traditional BIO tagging in favour of a simplified word/tag format that reduces complexity while improving boundary recognition. Combined with negative instances and detailed entity definitions, this approach enhances the model's ability to distinguish between entity and non-entity tokens while adapting to domain-specific entity types.

These methodological components form a cohesive framework addressing fundamental low-resource NER challenges, offering a promising solution for domain-specific tasks where labelled data is scarce. The following chapter details our experimental setup and evaluation framework for assessing PANER's performance.

Chapter 4

Experimental Setup

4.1 Introduction

This chapter presents the experimental setup for evaluating the PANER framework. Section 4.2 describes the datasets used for both training and evaluation, including our primary training corpus and benchmark datasets. Section 4.3 outlines the baseline approaches selected for comparison in zero-shot, few-shot, and data augmentation experiments. Section 4.4 details the backbone models employed for fine-tuning within our framework. Section 4.5 presents the data configuration setups used for zero-shot and few-shot evaluation scenarios. Section 4.6 provides comprehensive implementation details, including the training platform, optimization strategies for fine-tuning, computational resources, and reproducibility settings. Finally, Section 4.7 describes our evaluation framework and scoring methodology, including the specific metrics used and measures taken to ensure fair comparison across baselines.

4.2 Datasets

4.2.1 Training Corpus

PileNER serves as the primary training corpus for the instruction tuning phase of our framework. Introduced by W. Zhou et al. (2023), this dataset was constructed by sampling 50,000 passages from the Pile corpus—an 825 GiB English text dataset—annotated using ChatGPT (gpt-3.5-turbo-0301)

to identify entities and their associated types W. Zhou et al. (2023). PileNER has become a standard training corpus for instruction tuning in recent NER literature, particularly in few-shot and zero-shot setups. Notable studies such as those by Y. Ding et al. (2024); Zamai et al. (2024); W. Zhou et al. (2023) have used it as a primary corpus for their instruction-based NER model training, citing its breadth of entity types and depth in number of samples per entity present in the corpus. Each passage is limited to 256 tokens, resulting in a corpus of 45,889 input-output pairs, encompassing 240,725 entities and 13,020 unique entity types after filtering for parsing ability and excluding noisy labels (e.g., "NA", "MISC", or "ELSE").

A key characteristic of PileNER is the long-tailed distribution of entity types. The most frequent 1% of entity types account for 74% of total frequency counts, while the remaining 99% are split across increasingly rare categories. This distribution supports broad generalization capabilities, especially for few-shot and long-tail NER scenarios. Table 4.1, adapted from W. Zhou et al. (2023), illustrates the range and diversity of entity types by frequency bracket. Notably, entity types span general domains (e.g., person, organization) to fine-grained and domain-specific categories (e.g., immune response, input device), offering both breadth and granularity that are particularly valuable for instruction tuning in low-resource settings.

Frequency	**Entity types**
Top 1% (74%)	person, organization, location, date, concept, product, event, technology, group, medical condition, ...
1%–10% (19%)	characteristic, research, county, module, unit, feature, cell, package, anatomical structure, equipment, ...
10%–100% (7%)	attribute value, pokemon, immune response, physiology, animals, cell feature, FAC, input device, ward, broadcast, ...

Table 4.1: Examples of entities across different frequency ranges, along with the percentage of total frequencies for each range W. Zhou et al. (2023).

On top of the inherent diversity and long-tailed structure of PileNER, we applied a series of preprocessing steps to further refine the dataset for instruction tuning and few-shot experimentation. These filtering operations were designed not only to enhance the consistency and reliability of the training samples, but also to align with annotation standards used in prior NER research. Several of these steps were also applied to other datasets used in this work to maintain uniformity in data preparation.

Preprocessing Pipeline Our preprocessing pipeline for PileNER involved the following key steps:

- **Length Filtering**: We established a minimum sentence length threshold of 10 words to ensure sufficient context for meaningful entity recognition. This filtering removes short fragments or incomplete sentences that might introduce noise into the training process.
- **Language Filtering**: To maintain consistency and maximize our performance, we retained only English text, removing sentences in other languages.
- **Entity Type Filtering**: We restricted our training data to a predefined subset of 423 named entity types, as introduced by Zamai et al. (2024). This subset was curated by the original authors through a multi-step process: retaining only entity types with at least 100 instances, merging semantically identical or orthographically variant labels (e.g., *organization* and *organization*), and discarding vague or hallucinated categories such as *unknown*, *miscellaneous*, and *general entity type*. By relying on this filtered entity set and its associated annotation guidelines, we ensured that all entity types used during instruction tuning were well-defined, frequently occurring, and supported by consistent annotation protocols.

After applying these filtering criteria, the resulting dataset contained *23,402* high-quality samples.

Domain	# paragraph	# sentence	# tokens	# Train	# Dev	# Test	Entity Categories (examples)
Reuters	-	-	-	14,987	3,466	3,684	person, organization, location
Politics	2.76M	9.07M	176.56M	200	541	651	politician, person
Natural Science	1.72M	5.32M	98.50M	200	450	543	scientist, person, university
Music	3.49M	9.82M	194.62M	100	380	456	music genre, song
Literature	2.69M	9.17M	177.33M	100	400	416	book, writer, award
Artificial Intelligence	97.04K	287.62K	5.20M	100	350	431	algorithm, task, product

Table 4.2: Data statistics of unlabeled domain corpora, labelled NER samples, and representative entity categories for each domain in the CrossNER dataset (from Z. Liu et al. (2021).)

4.2.2 Benchmark Datasets

For our various experiments, we have leveraged four established NER datasets on top of PileNER as our primary evaluation and domain-specific training sources.

(1) **CrossNER** Z. Liu et al. (2021): A comprehensive cross-domain dataset spanning diverse subject areas including scientific papers, politics, music, and literature. The diversity across

domains, combined with relatively small training and development datasets, makes it particularly suitable for few-shot learning scenarios. We leverage portions of CrossNER's training data for few-shot domain adaptation experiments, allowing us to assess how effectively our paraphrase-augmented approach adapts to specialized vocabularies and domain-specific entity characteristics with minimal supervision. More details regarding the data split can be found in 4.2

(2) **MIT Restaurant and Movie Review Datasets** J. Liu, Pasupat, Cyphers, and Glass (2013): These datasets feature user-generated content with informal language and domain-specific terminology. Similar to CrossNER, we incorporate subsets of the MIT training data for few-shot learning scenarios, particularly to evaluate performance on conversational and review-style text that differs significantly from formal training corpora.

- **MIT Restaurant** consists of 9,448 sentences (7,660 train / 1,943 test) annotated with 8 entity types, including restaurant type, location, price range, amenities, cuisine, dish, time, and party size.
- **MIT Movie (Eng)** 10,841 sentences (7,660 train / 3,181 test), annotated with 12 entity types, including movie name, actor name, director name, genre, year, rating, and others.

(3) **BUSTER** Zugarini, Zamai, Ernandes, and Rigutini (2024): A document-level financial domain NER benchmark that we employ primarily for zero-shot evaluation on out-of-distribution entities. Following the evaluation protocol established by Zamai et al. (2024), we use BUSTER to assess the robustness of instruction-tuned models on completely unseen domain-specific terminology and entity relationships. Due to the document-level nature of this dataset, we perform sentence-level segmentation based on our context length constraints to ensure compatibility with our training framework.

(4) **CoNLL-2003** Sang and De Meulder (2003): The widely-used multilingual NER benchmark featuring standard entity categories (PERSON, LOCATION, ORGANIZATION). We utilize

CoNLL-2003 specifically to evaluate the effectiveness of our paraphrase-based data augmentation strategy, as it provides a well-established baseline for comparing augmentation techniques against existing methods in the literature.

The diversity of these datasets ensures that our evaluation framework can comprehensively assess the strengths and limitations of the proposed PANER approach across a wide range of real-world NER challenges, with particular emphasis on low-resource scenarios.

4.3 Baseline Approaches for Comparison

To comprehensively evaluate the effectiveness of our proposed PANER framework, we selected diverse state-of-the-art models for comparison. Each baseline represents a distinct methodological approach or architectural design, providing a robust comparative framework for assessing both zero-shot and few-shot performance, as well as the efficacy of our data augmentation technique.

4.3.1 Zero-shot and Few-shot NER Baselines

(1) **GoLLIE** Sainz et al. (2023): A generative model based on CodeLLAMA, specifically designed to leverage annotation guidelines formatted in a code-like representation. This approach is notable for its innovative encoding of labelling criteria directly within the prompt structure, representing the first explicit attempt to incorporate annotation guidelines in instruction tuning. For our evaluation, we utilized the 7B parameter variant of GoLLIE to maintain comparability with other models of similar scale in our study.

(2) **GLiNER-L** Zaratiana et al. (2023): An encoder-only model based on DeBERTa with 304 million parameters. Despite being the smallest model among the selected baselines, GLiNER-L has demonstrated competitive performance in out-of-distribution (OOD) zero-shot NER tasks. Its inclusion provides valuable insights into the trade-offs between model size and performance, particularly important for resource-constrained applications.

(3) **GNER** Y. Ding et al. (2024): GNER is a generative NER framework that improves zero-shot and supervised performance by incorporating negative instances non-entity tokens labelled

as "O"—into the instruction tuning process. GNER emphasizes the role of contextual and boundary information provided by surrounding non-entity text. The model uses BIO-tagged token-by-token outputs during generation. To handle the alignment between generated outputs and original tokens, GNER introduces a tailored Longest Common Subsequence (LCS) Matching algorithm to ensure efficient mapping despite possible generation errors. These design choices significantly enhance the model's recall and precision, especially in zero-shot settings.

(4) **SLIMER** Zamai et al. (2024): A model based on the LLAMA-2-7B chat architecture, fine-tuned with LoRA for 10 epochs. SLIMER integrates structured annotation guidelines, making it a particularly relevant benchmark for evaluating our guideline-based NER approach. The model emphasizes the use of enriched prompts that incorporate definitions and annotation guidelines to improve performance on unseen entities. They employed a turn-by-turn conversational style information extraction for each entity type present in the sentence.

(5) **InstructUIE** Wang et al. (2023): A unified information extraction framework utilizing a Flan-T5-xxl and Flan-T5 architecture fine-tuned on a diverse set of information extraction datasets.

(6) **UniNER** W. Zhou et al. (2023): Employs a conversational template with LLAMA for universal NER. The specific variant we compare against, UniNER-type+sup, incorporates both type information and supervised learning, representing a strong baseline for zero-shot performance.

This diverse set of baselines was selected based on their reliance on large language models for NER, ensuring fair comparison within similar model parameter spaces and computational requirements. The chosen methods share comparable training paradigms and architectural foundations, allowing us to evaluate PANER against approaches that operate within the same resource constraints while employing different strategies for prompt engineering, instruction tuning, and model optimization.

4.3.2 Data Augmentation Baselines

To specifically evaluate the effectiveness of our paraphrase-based data augmentation approach, we selected the following data augmentation methods for comparison:

(1) **DAGA** B. Ding et al. (2020): Utilizes a one-layer LSTM-based language model trained on linearized labelled sentences from CoNLL and other sequence-tagging datasets to generate synthetic training data. DAGA represents a traditional approach to data augmentation that focuses on statistical patterns learned from existing datasets rather than leveraging large language models.

(2) **MELM** R. Zhou et al. (2022): A data augmentation framework that ensures label-consistent entity replacements by fine-tuning XLM-RoBERTa with masked entity prediction. MELM introduced a strategy that injects entity labels into the training context, reducing token-label misalignment and improving entity diversity, particularly in low-resource and multilingual NER settings. It represents a strong baseline for entity-preserving augmentation techniques.

(3) **Traditional Augmentation Approaches**: We also consider conventional methods including:

- **Label-wise Substitution**: A straightforward entity replacement strategy proposed by Dai and Adel (2020) that generates augmented samples by randomly substituting named entities with existing entities of the same entity type from the original training set.
- **MLM-Entity**: A context-aware augmentation method that randomly masks entity tokens and directly utilizes a pretrained Masked Language Model (MLM) for data augmentation without additional fine-tuning or labelled sequence linearization.
- **Gold-Only**: Training exclusively with the original gold-standard annotated samples without any data augmentation, serving as a control baseline to quantify the impact of various augmentation strategies.

These data augmentation baselines provide a comprehensive framework for evaluating the effectiveness of our paraphrase-based approach, particularly for NER.

4.4 Backbone Models

The selection of appropriate model architectures was crucial for the success of PANER, particularly in low-resource scenarios where computational efficiency must be balanced with performance. Our approach leverages state-of-the-art Large Language Models (LLMs) as backbone architectures while employing efficient adaptation techniques to minimize computational overhead.

We selected three cutting-edge instruction-tuned models— Qwen-2.5-Instruct (7B) Yang et al. (2024), LLAMA-3.1-Instruct (8B) Touvron et al. (2023), and Falcon3-Instruct (10B) Almazrouei et al. (2023) —as the core backbone architectures for PANER. Our selection focused on models in the 7–10 billion parameter range, guided by practical considerations around computational feasibility: these are the largest models that can be reliably deployed on a single GPU with 40 GB of memory, making them ideal for cost-conscious research and deployment without compromising performance.

Initial experiments with smaller models in the 3B parameter range revealed substantial limitations in output quality, including lower consistency in output format and weaker generalization across domains. The performance gap relative to the 7B models was nontrivial, and the marginal computational savings did not justify the trade-off.

Moreover, many of the 7B-class models we selected are distilled or optimized versions of significantly larger architectures—often replicating the performance characteristics of models with over 30B or even 70B parameters. This distillation offers a favourable balance between performance and cost, enabling PANER to maintain state-of-the-art capabilities in a lightweight and accessible framework.

All three selected models are optimized for instruction-following tasks and feature extended context windows—128K tokens for Qwen-2.5 and LLAMA-3.1, and 32K tokens for Falcon3—making them well-suited for processing long, complex sequences. These models were chosen not only for their long-context handling capabilities, but also for their state-of-the-art performance on benchmarks such as MT-Bench Bai et al. (2024) and Alpaca WC (AlpacaEval) Dubois, Galambosi, Liang, and Hashimoto (2024), which respectively test multi-turn conversational coherence and single-turn instruction-following quality.

MT-Bench is a challenging, 80-plus prompt benchmark that evaluates a model's ability to maintain context, reasoning, and adaptability over multiple turns. Alpaca WC (AlpacaEval) is a fast, single-turn instruction benchmark using GPT-4 as a judge and validated against human preferences, providing reliable insights into prompt compliance and response quality. Together, these benchmarks affirm our backbones' capabilities in precise instruction following competencies for improving entity recognition. The corresponding benchmark scores are shown below in Table 4.3.

Model	**MTBench (multi-turn)**	**Alpaca WC (single-turn)**
Qwen-2.5-Instruct (7B)	8.08	—
LLAMA-3.1-Instruct (8B)	8.98	57.6
Falcon3-Instruct (10B)	8.37	31.97

Table 4.3: MT-Bench and Alpaca WC (AlpacaEval) evaluation scores for backbone models

We acknowledge the temporal differences between model generations in our evaluation framework. While our backbone models (Qwen-2.5, LLAMA-3.1, Falcon-3) represent more recent architectures compared to some baseline approaches, we maintain fairness through parameter-scale equivalence, comparing models within similar sizes (7B-11B parameters). This approach mirrors established practices in the field, where GNER and SLIMER similarly compared against earlier methods like UniNER and InstructUIE despite architectural advances. Critically, the extended context windows available in newer models (32K-128K tokens versus traditional 2K-4K limits) were essential for implementing our instruction tuning template with comprehensive entity definitions and guidelines an approach that would be infeasible with shorter context constraints. Our evaluation thus represents a fair assessment within equivalent parameter budgets while leveraging architectural capabilities necessary for our proposed approach.

4.5 Training Data Configurations by Evaluation Type

Zero-shot Evaluation: We fine-tuned the backbone models (4.4) on 23,402 PileNER samples, as described in Section 4.2.1. Although this represents a larger sample size than that used by Y. Ding et al. (2024), our training setup optimizes for it by doing single epoch LORA finetuning as opposed to the training setup of our baseline models.

Few-shot Evaluation: For few-shot scenarios, we utilized 10,000 examples randomly sampled from PileNER as our base dataset (aligned with Y. Ding et al. (2024) and systematically added domain-specific examples from the benchmark datasets (CrossNER and MIT) as necessary. We experimented with varying numbers of augmented samples (0, 100, and 300) to quantify the impact of our paraphrase-based augmentation technique.

Data Augmentation Experiments: For the CoNLL dataset specifically, we simulated low-resource scenarios by using 100, 200, and 400 gold samples, following the setup of R. Zhou et al. (2022), to facilitate direct comparison with existing data augmentation techniques, and then generating 200, 400, and 800 augmented samples, respectively, using our paraphrasing technique.

For experiments involving datasets with longer sentence structures and complex entity annotations, such as BUSTER Zugarini et al. (2024), we implement a chunking strategy to segment the input into manageable units. Unlike sentence-level NER datasets, BUSTER contains extended sequences that include multiple clauses, specially designed for long context NER. While chunking is often framed as a computational optimization, our primary motivation is to align with the constraints of our output format and to ensure that each broken-down segment remains interpretable and structurally coherent for both annotation and generation.

This strategy is particularly critical when using the Falcon3 model, which has a 32K context window—smaller than that of Qwen-2.5 and LLAMA-3.1 but remains well-suited for moderate sentences with dense entity definitions and guidelines. To ensure uniformity across models and reduce variability in output length, we restrict context length to *2048* tokens and automatically segment longer sequences accordingly. This limit helps standardize input handling and reduces the likelihood of truncation errors, especially in entity-dense regions.

4.6 Experimental Parameters and Training Configurations

4.6.1 Training Platform and Framework Selection

Our experimental infrastructure leverages Modal, a serverless cloud platform specifically designed for AI and machine learning workloads *Modal* (2023). Modal enables seamless scaling from

zero to thousands of GPUs through a Python-native interface, eliminating the traditional complexities of cloud infrastructure management. As illustrated in Figure B.2, the platform provides instant container deployment with optimized file systems and automatic resource scaling, making it particularly well-suited for iterative machine learning experiments (Figure B.3).

We employ Axolotl *axolotl-ai* (2023), an open-source framework that streamlines large language model fine-tuning through configuration-driven approaches. Axolotl enables rapid experimentation by allowing users to specify training parameters, datasets, and model configurations through simple YAML files, while automatically handling the underlying complexities of distributed training, memory optimization, and state management. This configuration-based approach facilitates reproducible experiments and enables quick iteration across different hyperparameter settings without requiring extensive code modifications.

4.6.2 Parameter Configuration and Optimization Strategy

All models were fine-tuned using Low-Rank Adaptation (LoRA) E. J. Hu et al. (2022) with carefully selected hyperparameters based on established best practices. Following the widely-adopted convention in the literature *axolotl-examples* (2023); Raschka (2025), we set the rank $r = 8$ and scaling factor $\alpha = 16$, maintaining the recommended $\alpha = 2 \times r$ ratio. This configuration has been demonstrated to provide an optimal balance between model adaptability and computational efficiency, with the 2:1 alpha-to-rank ratio serving as a standard guideline that ensures appropriate scaling of learned weight updates during fine-tuning.

We also employed the AdamW optimizer Loshchilov and Hutter (2017) with a cosine learning rate schedule, implementing a warm-up phase covering 4% of training steps and peaking at 2×10^{-5}. All models were fine-tuned for exactly one epoch, demonstrating significant computational efficiency compared to traditional approaches requiring multiple epochs. The complete YAML configuration file for Llama 3.1-8B-Instruct is provided in Appendix A.1, enabling full reproducibility of our experimental setup.

4.6.3 Paraphrasing Infrastructure and Structured Output Generation

For paraphrase generation, we employed a cloud-hosted version of LLAMA 3.3-70B due to its substantial computational requirements. Our implementation utilizes the Instructor package J. Liu (2024) in conjunction with Pydantic to ensure reliable structured output generation. Instructor provides a powerful framework for constraining open-ended LLM outputs into validated, structured formats using Pydantic models, enabling automatic retries when validation fails and ensuring consistent JSON output formatting. The code for which is shown in Appendix A. The complete prompt structure and generation parameters are detailed in Fig 3.3, as referenced earlier in our methodology.

4.6.4 Computational Resources and Scalability

We dynamically scaled our GPU usage based on model size and experimental requirements, utilizing between 1-8 NVIDIA A100 GPUs (40GB) for fine-tuning operations and a single NVIDIA A10 GPUs for inference tasks (Figure B.1) Modal's serverless architecture enables automatic scaling between these configurations (for inference), optimizing resource utilization and cost-effectiveness. For paraphrase generation, we hosted the LLAMA 3.3-70B model in an 8 * NVIDIA A100 GPU cluster in modal.

4.6.5 Reproducibility and Experimental Controls

All experiments employ fixed random seeds and standardized configurations to ensure reproducible outcomes. Our experimental framework implements comprehensive logging and metrics collection, with all hyperparameters and training configurations explicitly documented. The modular architecture of our implementation allows for easy replication and extension, with complete YAML configuration files and detailed setup instructions provided in the appendix to facilitate reproduction by other researchers.

4.7 Evaluation Framework and Scoring Methodology

We employ micro-averaged F1 score as our primary evaluation metric, which is the universal metric for NER performance evaluation in the literature, ensuring direct comparability with existing approaches. For our novel instruction tuning template that deviates from traditional BIO tagging standards, we employ entity-level F1 scores for both the BIO and word/tag formats to ensure fair comparison. While the formats differ in presentation, our evaluation criteria remain consistent across both approaches: an entity prediction is considered correct only when both the entity type and its complete boundaries match the gold standard annotation. Table 4.4 illustrates this scoring methodology with concrete examples, demonstrating how partial entity identification or incorrect boundary detection is treated as an error, regardless of whether the entity type was correctly identified.

Gold Standard	**BIO Prediction**	**Word/Tag Prediction**	**Score**
"New York University" (ORG entity)	B-ORG I-ORG I-ORG	New/ORG York/ORG University/ORG	Correct
"New York University" (ORG entity)	B-ORG I-ORG O	New/ORG York/ORG University/O	Wrong boundary

Table 4.4: Examples of entity-level scoring methodology for BIO and word/tag formats

To ensure robustness and statistical validity of our findings, all reported results represent averages over three experimental runs. We conducted systematic ablation studies to isolate the contributions of various components of our approach, including comparisons between our word/tag format against traditional BIO tagging schema and evaluating the impact of including entity definitions and guidelines in the instruction tuning process.

4.8 Summary

This chapter establishes the comprehensive experimental setup that will be used to evaluate the PANER framework across diverse NER scenarios. Our experimental design addresses both the paraphrase-based data augmentation strategy and the novel word/tag instruction tuning template through rigorous zero-shot, few-shot, and data augmentation evaluation paradigms.

Section 4.2 describes our dataset selection, utilizing PileNER as the primary training corpus with systematic preprocessing, alongside benchmark datasets spanning multiple domains: CrossNER for cross-domain assessment, MIT Restaurant and Movie datasets for informal user-generated content, BUSTER for out-of-distribution evaluation, and CoNLL for direct comparison with existing augmentation techniques. Section 4.3 outlines the comprehensive baseline selection for comparison across zero-shot, few-shot, and data augmentation scenarios.

Section 4.4 details our selection of three state-of-the-art instruction-tuned language models based on their extended context windows, strong instruction-following performance, and optimal balance between computational efficiency and capability. Section 4.5 presents the specific data configurations employed across different evaluation scenarios, while Section 4.6 describes the experimental infrastructure leveraging Modal's serverless platform with Axolotl framework for streamlined operations.

Our evaluation methodology, outlined in Section 4.7, employs entity-level F1 scoring with consistent criteria across both traditional BIO and our novel word/tag formats, ensuring fair comparison while validating the effectiveness of our simplified tagging approach. Comprehensive documentation of training configurations and evaluation protocols ensures reproducibility and facilitates future comparative analyses.

In the following chapter, we present the experimental results demonstrating PANER's performance across different datasets, resource levels, and baseline comparisons, providing detailed analyses of individual component contributions to validate our framework's overall performance improvements in low-resource NER scenarios.

Chapter 5

Results and Analysis

5.1 Introduction

This chapter presents a comprehensive analysis of the experimental results obtained from evaluating the PANER framework. We begin by examining the performance of our instruction tuning template, validating our core methodological design choices and assessing the effectiveness of our output format combined with entity guidelines and definitions. The evaluation then progresses to analyze our paraphrasing-based data augmentation technique in few-shot NER scenarios, establishing the quality and effectiveness of our augmentation strategy by systematically comparing against existing augmentation methods. We conduct additional evaluations to determine optimal configurations, including sample size selection and the number of paraphrases per input sentence. The scope of our tests involves testing involves both few-shot and zero-shot learning capabilities, with particular emphasis on demonstrating our approach's effectiveness in low-resource scenarios.

5.2 Comparison of Tagging Formats

We first validate the effectiveness of our word/tag format by comparing it against the BIO-style output format presented in Y. Ding et al. (2024). For this experiment, LLAMA 3.1-8B-Instruct was used as the backbone architecture, and we leveraged the same PileNER dataset W. Zhou et al. (2023) filtered for sentences with more than 10 words only in English, resulting in approximately 23,402

samples for training. The model was fine-tuned with LoRA using the hyperparameters specified in Section 4.6.2. To maintain consistency with GNER Y. Ding et al. (2024), we evaluated the model on the same five datasets (CrossNER), with 200 random samples per dataset, for zero-shot performance analysis. Results reported in Table 5.1 represent averages over three test runs to ensure robustness.

The evaluation methodology for comparing different tagging formats, as established in Section 4.7, employs entity-level F1 scores for both the BIO and word/tag formats to ensure fair comparison. While the formats differ in presentation, our evaluation criteria remain consistent across both approaches: an entity prediction is considered correct only when both the entity type and its complete boundaries match the gold standard annotation. This means that in both formats, partial entity identification or incorrect boundary detection is treated as an error, regardless of whether the entity type was correctly identified.

Reference Y. Ding et al. (2024) performed a similar boundary analysis with different tagging formats to determine the optimal approach. While their results differ from ours, this discrepancy could be attributed to our use of a larger model, the LLAMA 3.1-8B, compared to their Flan-T5 large model with 780M parameters. Additionally, we employ a LoRA fine-tuning approach for a single epoch, in contrast to their full fine-tuning over three epochs, which may also contribute to the observed differences in performance.

Table 5.1 demonstrates a significant improvement in F1 scores when adopting the word/tag format compared to the traditional BIO tagging schema. The GNER-BIO approach achieved an average F1 score of 51.92% across the five domains (AI, Literature, Music, Politics, and Science), while our word/tag format without guidelines (Ours-slash w/o) reached 67.13%, representing a substantial improvement of 15.21 percentage points. When definitions and guidelines are included alongside the new format (Ours-slash), we observe a further increase to 68.58%, yielding an additional 1.45 percentage point improvement. While this increase may appear marginal, it offers substantial benefits when integrated with our paraphrase-augmented synthetic data during few-shot testing. This improvement underscores the complementary nature of clear guidelines and the word/tag format in enhancing model accuracy and adaptability.

While the improvement from adding guidelines and definitions appears modest in the aggregate results (from 67.13% to 68.58%, a 1.45 percentage point increase), domain-specific analysis reveals

	AI	Lit	Music	Pol	Sci	Avg
BIO	52.1	51.1	58.5	54.1	43.8	51.92
Ours-slash w/o*	59.1	67.4	72.25	70.8	66.1	67.13
Ours-slash	63.9	67.2	75.3	67.8	68.7	**68.58**

Table 5.1: Comparison between instruction formats (F1 scores in %)
*w/o: prompt without guidelines

more nuanced benefits that align with research done on definition and guideline effectiveness by Zamai et al. (2024). As shown in Table 5.1, the inclusion of entity guidelines and definitions led to particularly significant improvements in the AI domain (from 59.1% to 63.9%, a 4.8 percentage point increase) and the Music domain (from 72.25% to 75.3%, a 3.05 percentage point increase).

These domain-specific improvements can be understood through the complexity and specificity of entity definitions required across different domains, as illustrated in Figure 3.4. The AI domain benefits substantially from explicit definitions due to its rapidly evolving terminology and the need for precise algorithmic context disambiguation. This observation is consistent with prior research demonstrating that definitions and guidelines are particularly effective for disambiguating polysemous tags and detecting novel named entities in specialized domains Zamai et al. (2024). Similarly, the Music domain shows marked improvement with additional contextual guidance, requiring careful distinction between overlapping categories and context-dependent identification, as exemplified by the need to differentiate musical artists from geographic locations sharing the same name.

The Literature domain showed a marginal decrease in performance when guidelines were added (from 67.4% to 67.2%), despite having detailed genre-specific definitions that distinguish between broad literary terms and specific stylistic categories. This finding suggests that domains with well-established entity types may experience diminishing returns from additional guideline complexity, particularly when the entity categories are already well-understood in natural language contexts. This observation aligns with Zamai et al. (2024), who also pointed out in their work that not all definitions and guidelines were beneficial to entity recognition.

5.3 Performance of Instruction Tuning Template in Zero-shot NER

5.3.1 Out-of-distribution Named Entities

While our primary goal is to improve few-shot NER, the proposed instruction-tuning template also performs competitively with state-of-the-art methods in zero-shot NER. Table 5.2 presents the zero-shot performance compared to existing state-of-the-art benchmarks. Our framework demonstrates the effectiveness of combining principles from both GNER and SLIMER approaches. Compared to SLIMER, which achieved an average F1 score of 54.0%, our approach shows substantial improvements across all backbone models, with gains of 6.5 percentage points (Qwen-2.5: 60.5%), 7.7 percentage points (LLAMA-3.1: 61.7%), and 10.8 percentage points (Falcon-3: 64.8%). This improvement validates our integration of GNER's negative instance handling with SLIMER's guideline-centric philosophy through our simplified word/tag output format.

When compared to the GNER variants, our framework achieves competitive results despite requiring significantly fewer resources. While GNER-T5 leads with 69.1% F1 using an 11B parameter model and full fine-tuning over three epochs, our Falcon-3 implementation achieves 64.8% F1 with a 10B parameter model using LoRA fine-tuning for only one epoch. The 4.3 percentage point difference represents a reasonable trade-off considering the substantial computational efficiency gains. Similarly, when comparing similar-sized models, GNER-LLAMA (7B, 66.1% F1) slightly outperforms our LLAMA-3.1 variant (8B, 61.7% F1) by 4.4 percentage points, but again at the cost of three-fold increased training time.

Domain-specific analysis reveals interesting performance patterns. Our Falcon-3-10B-Instruct model demonstrates exceptional performance in the Movie domain (69.4% F1), notably outperforming GNER-T5 (62.5%) by 6.9 percentage points and GNER-LLAMA (68.6%) by 0.8 percentage points. In the Music domain, we achieve competitive results (75.8% F1) compared to GNER-T5 (81.2%) and GNER-LLAMA (75.7%), with only a 5.4 percentage point gap from the leading method. However, the Restaurant domain presents challenges for all approaches, with our best performance at 43.3% F1 compared to GNER-T5's 51.0%, indicating that annotations and guidelines are not helpful for all categories of entities.

Model	Backbone	#Params	Movie	Restaurant	AI	Literature	Music	Politics	Science	Average
ChatGPT	gpt-3.5-turbo	-	5.3	32.8	52.4	39.8	66.6	68.5	67	47.5
InstructUIE	Flan-T5-xxl	11B	63	21	49	47.2	53.2	48.2	49.3	47.3
UniNER-type+sup.	LLAMA-1	7B	61.2	35.2	62.9	64.9	70.6	66.9	70.8	61.8
GoLLIE	Code-LLAMA	7B	63	43.4	59.1	62.7	67.8	57.2	55.5	58.4
GLiNER-L	DeBERTa-v3	0.3B	57.2	42.9	57.2	64.4	69.6	72.6	62.6	60.9
GNER-T5	Flan-T5-xxl	11B	62.5	51	68.2	68.7	81.2	75.1	76.7	69.1
GNER-LLAMA	LLAMA-1	7B	68.6	47.5	63.1	68.2	75.7	69.4	69.9	66.1
SLIMER	LLAMA-2-chat	7B	50.9	38.2	50.1	58.7	60	63.9	56.3	54
PANER	Qwen-2.5-Instruct	7B	51.5	37.3	62	61.7	75.9	69.72	65.63	60.5
PANER	LLAMA-3.1-Instruct	8B	52	37	63.9	67.2	75.3	67.8	68.7	61.7
PANER	Falcon3-Instruct	10B	69.4	43.3	65.5	61.3	75.8	70.3	68.3	64.8

Table 5.2: Comparison of Zero-shot Learning Performance F1 (%) scores

Model	Backbone	#Params	Pr.	R	F1
GNER-LLAMA	LLAMA-1	7B	14.68	59.97	23.58
GLINER-L	DeBERTa-v3	0.3B	42.55	19.31	26.57
GoLLIE	Code-LLAMA	7B	28.82	26.63	27.68
GNER-T5	Flan-T5-xxl	11B	19.31	50.15	27.88
UniNER-type+sup.	LLAMA-1	7B	31.4	47.53	37.82
SLIMER	LLAMA-2-chat	7B	47.69	43.09	**45.27**
PANER	Llama-3.1-8B-Instruct	8B	27.28	37.50	**31.58**
PANER	Falcon3-Instruct	10B	29.92	38.38	**33.63**

Table 5.3: Zero-shot result comparison on BUSTER dataset F1 (%) scores

5.3.2 Never-seen-before Named Entities

Further, the out-of-domain performance of our instruction-tuning template is showcased in Table 5.3.2, where we compare it against the above-mentioned models on the BUSTER dataset. The results demonstrate that our PANER Falcon-3-10B-Instruct model performs better than both GNER models with an F1 of 33.63% compared to GNER-LLAMA (23.58%) and GNER-T5 (27.88%), representing improvements of 10.05 and 5.75 percentage points, respectively. Our approach achieves performance comparable to SLIMER, which holds the state-of-the-art F1 score of 45.27%, with an 11.64 percentage point difference. This out-of-domain performance is particularly noteworthy since we truncated the input in-order to fit into the context of our models with our output format, whereas SLIMER is a turn by turn conversational entity recognition system and achieved superior performance.

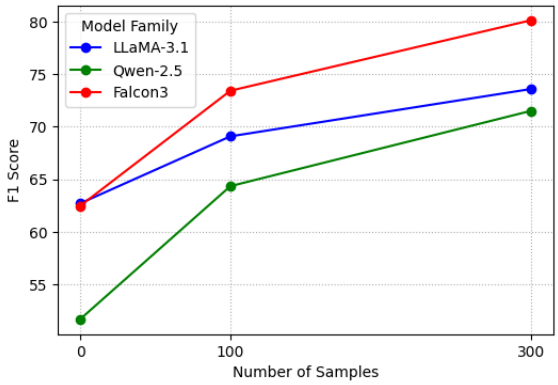

Figure 5.1: Impact of augmented sample size on model performance (F1 score, in %) for CrossNER dataset.

5.4 Performance of Paraphrasing in Few-shot NER

We evaluate the effectiveness of our paraphrasing-based data augmentation approach using the experimental setup described in Section 4.5 and the backbone models detailed in Section 4.4, we test three configurations: baseline with no augmentation or original samples, few-shot with 100 original domain-specific samples, and augmented configurations combining 100 original samples with 200 paraphrased variants from the 10,000 PileNER samples as mentioned in the data preparation step.

The experimental results reveal substantial improvements across all tested model architectures when compared against the baseline (0 original, 0 augmented samples). LLAMA-3.1-8B-Instruct showed progressive improvements from the baseline average F1 of 55.3% to 60.8% with 100 original samples, and further to 67.4% with augmented data (12.1 percentage point total improvement). The augmentation strategy alone contributed an additional 6.6 percentage points beyond the few-shot baseline, with notable gains in the Movie domain (from baseline 43.3% to 64.2% with augmentation).

Qwen-2.5-7B-Instruct demonstrated even more substantial gains from its baseline average F1 of 48.9% to 57.4% with original samples, and 65.8% with augmentation (16.9 percentage point total improvement). The paraphrasing augmentation contributed 8.4 percentage points beyond the few-shot performance, with particularly strong responses in the Literature domain (from baseline 54.9% to 73.3% with augmentation).

Falcon-3-10B-Instruct achieved the most compelling results, improving from a baseline average F1 of 59.3% to 66.8% with original samples, and 74.4% with augmentation (15.1 percentage point total improvement). The augmentation strategy contributed 7.6 percentage points beyond the few-shot baseline, establishing new benchmarks with 85.3% F1 in Music and 82.3% F1 in Science.

Notably, we also observe a consistent pattern where the Restaurant domain from the MIT dataset shows lower performance gains across all three model architectures compared to other domains. This phenomenon can be attributed to the challenging nature of applying guidelines and annotations in this specific domain, which is further substantiated in the previous Table 5.2 which examinies the zero-shot performance results. In the zero-shot setting, GNER-T5 and GNER-LLAMA, which do not incorporate entity guideline annotations, achieve Restaurant domain F1 scores of 51.0% and 47.5% respectively, significantly outperforming SLIMER (38.2%), which utilizes the same annotation and guideline approach as our method. This pattern suggests that the Restaurant domain may be inherently resistant to guideline-based improvements, as the domain-specific entity relationships and contextual dependencies may not benefit from explicit guidance. As demonstrated in Section 5.2 and annotations effectiveness across domains, certain entity types and domains show diminishing returns from additional guideline complexity. The modest improvements seen in Restaurant domain performance (ranging from 30.3% to 42.8% across models) indicate that while our paraphrasing approach maintains entity relationships effectively, the combination with guideline-based instruction tuning may not provide the optimal framework for this particular domain's unique characteristics.

As illustrated in Figure 5.1, there is a clear correlation between increased augmented samples and F1 score improvements across all models. These results demonstrate that our data augmentation technique can enhance model performance significantly in few-shot scenarios. As shown in Table 5.5, our PANER framework using Falcon-3 with augmented samples achieves superior performance compared to supervised techniques from the original CrossNER paper, which utilized all available training data within the CrossNER datasets. Our approach achieves an average F1 score of 80.1%, substantially outperforming the best supervised baseline methods, including NER-BERT (72.7%), and DAPT (69.6%). While NER-BERT achieves the highest individual performance in the AI domain (76.1% vs. our 72.7%), our method demonstrates clear increases across the remaining four domains. This is particularly noteworthy given that these supervised methods had access to the

complete training and development sets, whereas our approach achieves superior results using only 100 original samples plus 200 augmented variants per domain.

The consistent improvements demonstrated across different model architectures, domain-specific few-shot comparisons, and supervised baseline methods indicate that combining lightweight few-shot learning with intelligent data augmentation offers a comprehensive solution for domain-specific NER tasks. Our approach not only presents a viable path for bridging the performance gap between resource-constrained few-shot scenarios and supervised methods but surpasses traditional supervised approaches while requiring significantly less labelled data. This positions our framework as a promising direction for organizations operating in scenarios where labelled data is scarce and computational resources are limited.

Model Family	**# of Original Samples**	**# of Augmented Samples**	**Movie**	**Restaurant**	**AI**	**Literature**	**Music**	**Politics**	**Science**	**Average**
LLAMA-3.1-8B-Instruct	0	0	43.3	30.3	59.4	61.5	68.2	62.1	62.3	55.3
	100	0	45.1	35.4	61.0	67.2	75.9	70.4	70.9	60.8
	100	200	64.2	39.8	64.0	77.0	80.8	73.2	72.9	**67.4**
Qwen-2.5-7B-Instruct	0	0	50.3	33.6	46.5	54.9	53.9	54.0	49.1	48.9
	100	0	54.2	25.6	57.8	63.4	70.4	64.2	65.9	57.4
	100	200	65.1	38.2	59.5	73.3	79.2	70.2	75.3	**65.8**
Falcon-3-10B-Instruct	0	0	64.5	38.1	63.7	58.8	68.6	61.8	59.4	59.3
	100	0	63.7	37.0	67.8	67.6	82.2	72.0	77.5	66.8
	100	200	77.5	42.8	72.7	79.0	85.3	81.3	82.3	**74.4**

Table 5.4: Few-shot F1 (%) scores using augmented samples Across Different Domains

	AI	**Lit**	**Music**	**Pol**	**Sci**	**Avg**
BERT	68.7	64.9	68.3	63.6	58.8	64.9
CDLM	68.4	64.3	63.5	59.5	53.7	61.9
DAPT	72.0	68.8	75.7	69.0	62.6	69.6
NER-BERT	**76.1**	72.1	80.2	71.9	63.3	72.7
PANER (*Ours*)	72.7	**79**	**85.3**	**81.3**	**82.3**	**80.1**

Table 5.5: Comparison of F1 (%) scores on CrossNER for supervised techniques

5.5 Effectiveness of Paraphrase-Based Augmentation Compared to Data Duplication and In-Domain Expansion

To isolate the specific impact of our paraphrasing methodology, we conducted a controlled experiment comparing our data augmentation approach against conventional alternatives, including simple data duplication and in-domain sample expansion using the Falcon-3-10B-Instruct model. This experiment systematically evaluated three training configurations, each maintaining a total of 300 samples while differing in their compositional approach:

(1) 100 original in-domain samples augmented with 200 paraphrased variants

(2) 300 distinct original in-domain samples

(3) 100 original in-domain samples duplicated twice

The experimental results, presented in Table 5.6, show that the configuration utilizing 300 distinct original samples achieved the highest average F1 score of 75.3%, which represents the expected upper bound given the inherent value of diverse and authentic training samples. However, our hybrid approach combining 100 original samples with 200 paraphrased variants performed remarkably well, achieving an F1 score of 73.2%, representing only a 2.1 percentage point difference from the all-original configuration. This minimal performance gap suggests that our paraphrasing strategy successfully preserves essential entity relationships while introducing beneficial linguistic variation.

In contrast, the simple data duplication approach yielded substantially lower performance at 66.8%, confirming that mere repetition of training examples provides no meaningful diversity to enhance model generalization. These findings validate our augmentation approach as an effective strategy when additional authentic in-domain samples are unavailable or prohibitively expensive to obtain, demonstrating nearly comparable performance to training with three times the amount of original data.

The results provide compelling evidence for the effectiveness of LLM-generated paraphrases in enhancing model generalization for NER tasks.

	AI	Lit	Music	Pol	Sci	Avg
100 OG + 200 dup	60.8	67.5	72.7	68.4	64.9	66.8
100 OG + 200 aug	63.9	77.1	80.1	71.9	72.7	73.2
300 OG	67.4	79.0	80.2	73.3	76.5	75.3

Table 5.6: Comparison of F1 (%) scores on CrossNER for augmentation composition with Falcon -3-10B-instruct

5.6 Comparative Analysis with Alternative Augmentation Strategies

To establish the relative effectiveness of our paraphrasing-based approach, we conducted comprehensive comparisons with established data augmentation techniques in the NER domain. This comparative analysis provides crucial insights into the unique advantages of LLM-based paraphrasing compared to traditional augmentation methods, following the experimental framework established in the literature. Our comparison included several representative approaches ranging from LSTM-based approaches to masked entity language models mentioned in Section 4.3.2.

To simulate low-resource scenarios, we followed the methodology established by MELM R. Zhou et al. (2022) using 100, 200, and 400 gold samples from the CoNLL dataset, generating 200, 400, and 800 augmented samples, respectively. Our LLAMA 3.1-8B model was trained using these configurations, with our approach utilizing a 2× augmentation ratio compared to MELM's 3× ratio, making our method more data-efficient.

The results demonstrate clear superiority of our LLM-based paraphrasing approach across all tested sample size configurations. With 100 gold samples, our method achieved an F1 score of 80.52%, substantially outperforming MELM (75.21%) and DAGA (68.06%). This 5.31 percentage point improvement over MELM is particularly significant given that MELM utilizes 1.5x the number of augmented samples compared to our approach, highlighting the efficiency advantages of our method. The performance advantage remains consistent across different data availability scenarios, with our approach achieving 85.74% F1 compared to 82.91% for MELM and 79.11% for DAGA at 200 gold samples. At 400 gold samples, our method reached 88.11% F1, outperforming MELM (85.73%) and DAGA (84.36%).

The superior performance of our approach can be attributed to several factors. First, our use of LLMs for prediction leverages models that are already familiar with the entity types in the CoNLL

dataset (PERSON, LOCATION, ORGANIZATION), providing a substantial advantage over approaches that must learn these patterns from limited training data. Second, our entity-preserving paraphrasing strategy maintains critical semantic relationships while introducing meaningful linguistic diversity, contrasting with MELM's masked entity prediction approach, which may not capture the same level of contextual sophistication.

#Gold	Method	F1 Score (in %)
100	Gold-Only	50.57
	Label-wise	61.34
	MLM-Entity	61.22
	DAGA	68.06
	MELM	**75.21**
	PANER (Ours)	**80.52**
200	Gold-Only	74.64
	Label-wise	76.82
	MLM-Entity	79.16
	DAGA	79.11
	MELM	**82.91**
	PANER (Ours)	**85.74**
400	Gold-Only	81.85
	Label-wise	84.62
	MLM-Entity	83.82
	DAGA	84.36
	MELM	**85.73**
	PANER (Ours)	**88.11**

Table 5.7: Performance comparison of different augmentation methods on English (En)

The controlled nature of our paraphrasing process, guided by specific instructions and validation mechanisms, ensures higher quality augmented samples compared to the autoregressive generation used in DAGA. While DAGA generates entire sentences from scratch using an LSTM-based language model trained on gold NER data, our approach focuses on modifying only the context surrounding entities, preserving the original entity relationships while introducing beneficial variation.

Furthermore, the efficiency gains observed in our approach extend beyond mere performance metrics. The reduced augmentation ratio required to achieve superior results translates to lower computational costs during training. This practical advantage makes our approach particularly suitable for resource-constrained environments where both data and computational resources are limited.

5.7 Quality Analysis of Generated Paraphrases

To establish the optimal number of paraphrased variants per original sentence, we conducted a systematic analysis evaluating both the quality and diversity of generated augmentations. Our investigation examined augmentation counts ranging from 1 to 3 paraphrases per original sentence, measuring quality dimensions to identify the point of diminishing returns. We assessed augmentation quality using cosine similarity with Sentence-BERT Reimers and Gurevych (2019) and lexical diversity analysis (Type-Token Ratio). The evaluation excluded entity placeholder tokens (e.g., <PER>, <LOC>) from similarity calculations to focus on contextual variation.

Augmentation	**Cosine Similarity**†	**Lexical Diversity**‡
1st Augmentation	0.72	0.82
2nd Augmentation	0.68	0.79
3rd Augmentation	0.85	0.63

Table 5.8: Quality Analysis of Paraphrase Generation by Augmentation Count
†Average cosine similarity between original and augmented sentences using Sentence-BERT embeddings (lower = more diverse)
‡Type-Token Ratio excluding entity placeholders (higher = more diverse)

Our analysis reveals that augmentation quality exhibits a clear degradation pattern beyond the second paraphrase, as shown in Table 5.8. The cosine similarity scores demonstrate that while the first two augmentations maintain reasonable diversity (0.72 and 0.68, respectively), the third augmentation shows a significant increase to 0.85, indicating reduced linguistic variation.

More critically, manual inspection revealed a substantial degradation in JSON format compliance beyond two augmentations. Despite employing the Instructor package J. Liu (2024) to enforce structured response generation and class-based parsing for extracting paraphrases, we encountered significant parsing errors when generating three augmentations. The primary issue stemmed from inconsistent entity tag counts, where the generated paraphrases contained a different number of <ENT> placeholders compared to the original sentences. These malformed samples were excluded from our analysis, which explains the relatively modest numerical differences in Table 5.8 despite the underlying degradation of quality. Consequently, we established a hard constraint of two augmented sentences per original sentence to maintain integrity and minimize errors.

This limitation could potentially be addressed through the use of more advanced language models; however, at the time of conducting these experiments, LLAMA-3.3-70B demonstrated superior performance across diverse benchmarks among open-source alternatives, even surpassing the substantially larger LLAMA-3.1-405B model Touvron et al. (2023). We deliberately avoided proprietary models such as OpenAI's GPT series, O-series, or Anthropic's Claude models due to their closed-source nature and associated transparency limitations. Our approach prioritizes reproducibility and industry-wide adoption, particularly in privacy-sensitive domains where open-source solutions are essential for maintaining data sovereignty and transparency.

The lexical diversity analysis further supports these findings, with Type-Token Ratio scores decreasing from 0.82 and 0.79 for the first two augmentations to 0.63 for the third, indicating that subsequent paraphrases increasingly rely on repetitive vocabulary and sentence structures. Based on these empirical findings, we adopted a 2:1 augmentation ratio (two paraphrases per original sentence) as the optimal balance between sample diversity and generation quality. This configuration maximizes linguistic variation while maintaining high format compliance and entity consistency, ensuring reliable integration with our instruction tuning pipeline.

5.8 Base Dataset Sample Size Ablation Study

We also conducted an ablation study of the number of PileNER base samples required for effective zero-shot domain transfer. We did so by varying the training corpus size from 1,000 to 23,402 samples. Although GNER Y. Ding et al. (2024) established 10,000 samples as their baseline for full fine-tuning experiments, our approach was designed to assess whether our LoRA-based approach could achieve comparable performance with fewer training instances (since we incorporate guidelines and annotations), thus reducing computational requirements and training time.

The results presented in Table 5.9 demonstrate a clear positive correlation between base dataset size and zero-shot performance across all CrossNER domains. The substantial performance gap between smaller sample sizes (1,000-2,000 samples) can be attributed to insufficient training data leading to inadequate instruction adherence, resulting in reduced F1. The marked improvement from 41.22% to 53.92% average F1 score between 1,000 and 5,000 samples indicates that our instruction

Samples	AI	Lit	Music	Pol	Sci	Avg
1000	35.2	42.1	48.6	41.3	38.9	41.22
2000	48.7	44.3	51.2	45.6	42.1	46.38
5000	56.4	51.8	58.9	52.2	50.3	53.92
7500	60.1	54.5	62.1	55.0	54.5	57.24
10,000	62.7	58.8	68.6	61.8	59.4	62.26
15,000	63.1	59.6	71.5	64.2	63.8	64.44
23,402	65.5	61.3	75.8	70.3	68.3	68.8

Table 5.9: Ablation study showing F1 scores (%) across varying numbers of PileNER base samples using Falcon-3-10B-Instruct on CrossNER datasets.

tuning approach requires a critical mass of examples to effectively learn the word/tag formatting conventions and entity boundary detection principles.

While performance continues to improve steadily up to 10,000 samples, the rate of improvement begins to moderate beyond this point. The progression from 10,000 to 23,402 samples yields a 6.54 percentage point increase (62.26% to 68.8%), demonstrating continued but gradual enhancement. This slower improvement rate beyond 10,000 samples underscores the significant potential of few-shot learning approaches, where the addition of domain-specific samples can achieve substantial performance gains that would otherwise require exponentially more general training data. Our final configuration using 23,402 samples, as mentioned in Section 4.5, was added to establish the baseline that we use for our zero-shot experiments.

5.9 Summary

This chapter presented comprehensive experimental validation of the PANER framework across multiple evaluation dimensions. Table 5.10 gives a summarized outlook on the particular tests run on PANER. The initial analysis demonstrated the superiority of our word/tag instruction format over traditional BIO tagging, achieving a 15.21 percentage point improvement (67.13% vs 51.92% F1) with additional gains from entity guidelines and definitions. Domain-specific analysis revealed that guidelines provide particularly significant benefits in complex domains like AI (4.8 percentage point increase) and Music (3.05 percentage point increase), while showing diminishing returns in well-established domains like Literature.

Our paraphrasing-based data augmentation strategy consistently improved performance across

Experiment Type	Training Data Configuration	Evaluation Datasets	Sample Sizes	Specific Objectives
Instruction Format Validation				
BIO vs Word/Tag Format Comparison	PileNER (23,402 samples)	CrossNER (5 domains)	200 samples per domain	Validate word/tag format effectiveness against traditional BIO tagging
Zero-shot Performance Evaluation				
Zero-shot NER	PileNER (23,402 samples)	CrossNER, MIT, BUSTER	Full test sets	Assess generalization to unseen domains without task-specific examples
Few-shot Learning Assessment				
Base Dataset Size Ablation	PileNER (0 to 10,000 samples)	CrossNER, MIT	Varying configurations	Determine optimal base dataset size for few-shot learning
Paraphrase Augmentation Impact	10,000 PileNER + 0/100/300 domain samples	CrossNER, MIT	0, 100, 300 augmented samples	Quantify paraphrase augmentation effectiveness
Data Augmentation Comparison				
Augmentation vs Duplication	100 original + 200 augmented vs 300 original vs 100 duplicated	CrossNER	300 total samples	Isolate paraphrasing benefits vs simple data expansion
CoNLL Augmentation Benchmark	100/200/400 gold samples	CoNLL-2003	200/400/800 augmented samples	Compare against established augmentation methods (DAGA, MELM)

Table 5.10: Summary of Experimental Setup Configurations for PANER Framework Evaluation

all backbone models in few-shot scenarios. LLAMA-3.1-8B showed 12.1 percentage point improvement with augmentation, Qwen-2.5-7B achieved 16.9 percentage point gains, and Falcon-3-10B demonstrated 15.1 percentage point improvement. Notably, our Falcon-3 implementation with augmented samples achieved an average F1 score of 80.1% on CrossNER, surpassing supervised baseline methods including NER-BERT (72.7%) while using only 100 original samples plus 200 augmented variants per domain.

Comparative analysis against established augmentation methods on CoNLL datasets revealed substantial advantages of our LLM-based approach. Our method achieved 80.52% F1 with 100 gold samples, outperforming MELM (75.21%) and DAGA (68.06%) while requiring fewer augmented samples (2× vs 3× ratio). The controlled experiment comparing paraphrasing against data duplication and in-domain expansion validated our approach's effectiveness, achieving 73.2% F1 compared to 75.3% for 300 original samples and significantly outperforming simple duplication (66.8%).

Quality analysis of generated paraphrases established the optimal 2:1 augmentation ratio, with cosine similarity scores (0.72 and 0.68 for first two augmentations) indicating appropriate diversity while maintaining semantic consistency. The third augmentation showed quality degradation (0.85 similarity) and increased format compliance issues, confirming our constraint design.

Zero-shot evaluation demonstrated competitive performance with state-of-the-art methods while requiring significantly fewer computational resources. Our Falcon-3 model achieved 64.8% average F1, approaching GNER-T5 (69.1%) and GNER-LLAMA (66.1%) performance using LoRA fine-tuning for one epoch versus full fine-tuning for three epochs. Out-of-domain evaluation on BUSTER dataset showed superior performance (33.63% F1) compared to both GNER variants, validating the robustness of our instruction tuning template.

The base dataset ablation study revealed the importance of sufficient training data, with performance plateauing around 10,000 samples but continued gradual improvement up to 23,402 samples. These findings collectively establish PANER as an effective framework for bridging the performance gap between resource-rich and resource-constrained NER environments, offering competitive performance with substantially reduced computational requirements and training data needs.

Chapter 6

Discussions

This chapter is an extension of the previous chapter, where we presented the results and tried to interpret them individually. In this chapter, we try to examine the theoretical and practical implications of our PANER framework, discussing how our contributions advance the field of low-resource NER. We also talk about the broader limitations of our framework.

6.1 Results Interpretation and Analysis

6.1.1 Paraphrasing-Based Data Augmentation Efficacy in Few-Shot NER

The experimental results presented in Section 5.4 reveal fundamental insights that advance our understanding of effective strategies for addressing data scarcity in NER tasks. The most important finding concerns the substantial performance improvements achieved through our paraphrasing-based data augmentation approach, which consistently delivers gains ranging from 6.6 to 8.4 percentage points across different model architectures in few-shot learning scenarios establishes that even modest data augmentation can help resource-constrained environments achieve good results.

More specifically, our Falcon-3-10B-Instruct model also approaches state-of-the-art performance levels, and with minimal augmentation (100 original samples plus 200 paraphrased variants), achieved 80.1% F1 on CrossNER—substantially outperforming supervised baseline methods including NER-BERT (72.7%) while using fewer labelled samples. These improvements present a new, effective

way to utilize limited training data to achieve competitive performance levels, especially in domain-specific areas.

6.1.2 Output Format Optimization: Theoretical Implications for Instruction-Tuned Sequence Labelling

Our simplified word/tag output format performed better than traditional BIO tagging schemes, proving that text-generation models have an ease of adaptation to simpler output formats that align more naturally with their language understanding capabilities, especially for sequence labelling tasks. The 15.21 percentage point average improvement observed when transitioning from BIO to our simplified format suggests that the constraints imposed by complex tagging schemes may unnecessarily limit model performance, particularly in low-resource scenarios. Although there is a limitation associated with letting go of start and end information within sequencing tasks, we have to consider the huge performance upside from adopting a simpler output format while handling that limitation through other means.

6.1.3 LLM-based Augmentation Comparative Analysis Against Established Augmentation Methodologies

Our comparative analysis against established augmentation methods reveals the fundamental advantages of leveraging large language models for generating synthetic training data. Our methods consistently outperform methods like MELM and DAGA—achieving 80.52% F1 versus 75.21% and 68.06% respectively with 33% less data, showcasing that qualitative generation methods introduce linguistic diversity that yields superior returns compared to approaches that simply increase data volume.

6.1.4 Zero-Shot Instruction Tuning Template Performance

Beyond augmentation effectiveness, our instruction tuning template demonstrates robust standalone performance within the zero-shot framework. The domain-specific effectiveness of entity guidelines and definitions reveals that technical domains such as AI and Music show particularly strong improvements (4.8 and 3.05 percentage points, respectively), while well-established domains

like Literature show marginal decreases, indicating that guideline utility is not always beneficial for better entity identification.

Perhaps most significantly, our results demonstrate that competitive NER performance can be achieved with substantially reduced computational requirements compared to existing state-of-the-art approaches, even in a zero-shot setting. The fact that our instruction tuning template, infused with guidelines and annotations trained on single-epoch LoRA fine-tuning, performs comparably to methods requiring multiple epochs of full fine-tuning—as evidenced by our Falcon-3 model achieving 64.8% F1 approaching GNER-T5's 69.1% with three-fold computational reduction—represents the standalone capabilities of our proposed instruction tuning template. This makes our contribution much more effective for resource-constrained NER environments.

6.2 Cost-Effective NER Development: Foundations and Practical Deployment Implications

The implications of our PANER framework extend far beyond the specific performance improvements demonstrated in our experiments. It also changes the cost-benefit analysis for developing effective NER systems in data-scarce environments. The most immediate implication concerns the dramatic reduction in annotation requirements needed to achieve target performance levels. Traditional approaches to domain-specific NER often require thousands of annotated examples to achieve acceptable performance, creating significant barriers for organizations working in specialized domains or emerging fields. Our framework demonstrates that competitive performance—surpassing supervised baselines like NER-BERT by 7.4 percentage points—can be achieved with as few as 100 domain-specific examples.

6.3 Limitations

6.3.1 Model Dependencies and Augmentation Quality

The quality of the generated paraphrases depends on the capabilities and training distribution of the underlying paraphrasing model. Our approach relies on LLAMA 3.3-70B for generating

paraphrases, and the effectiveness is inherently limited by this model's understanding of domain-specific terminology and contextual relationships. Highly specialized domains with terminology that falls outside the training distribution of the paraphrasing model may experience reduced augmentation quality. This limitation is particularly evident in cutting-edge scientific domains or highly specialized professional fields where the surrounding terminology evolves rapidly.

Our strict entity preservation constraints, while effective for maintaining semantic relationships, restrict the diversity of generated samples to a degree. During our analysis, we observed that augmented sentences often exhibit limited structural variation when multiple entities appear nearby, as our approach compresses them into a single entity placeholder, to reduce paraphrasing complexity and drastically reduce the sentence length. This trade-off between entity integrity and linguistic diversity represents an inherent tension within our current implementation.

6.3.2 Entity Boundary Detection Challenges

A significant limitation emerges in handling nested NER scenarios, where entities are embedded within other entities. Our current framework does not adequately address these complex entity structures, which are common in specialized domains such as biomedical texts or legal documents.

More critically, the boundary detection of consecutive entities presents substantial challenges that stem from our design choice to abandon traditional BIO tagging in favour of our simplified word/tag format. While this decision enables faster inference and improved overall entity identification performance, it creates inherent difficulties when multiple entities of the same type appear sequentially (though combated by our use of external lexicons-based separation, it is an added step of complexity). Our approach struggles to differentiate the end boundary of the first entity from the start boundary of the second entity, as we deliberately chose to forgo the explicit start and end index tracking that BIO tagging provides. This limitation becomes particularly problematic in citation-heavy texts, where sequences of author names appear consecutively (e.g., "Smith, Johnson, Williams, and Davis"), making it challenging to determine where one PERSON entity ends and another begins.

This represents a fundamental trade-off in our architectural decisions. By prioritizing computational efficiency and simplifying the annotation schema, we accepted reduced precision in scenarios

involving consecutive same-type entities. While traditional BIO-based approaches theoretically address this issue through their explicit Beginning-Inside-Outside labelling scheme, they face their own challenges with annotation complexity and increased computational overhead. Even state-of-the-art BIO-based models struggle with boundary detection accuracy, particularly in domains with frequent entity adjacency, achieving entity-level F1 scores that drop significantly when strict boundary matching is required. However, our simplified approach exacerbates this challenge by removing the structural scaffolding that BIO tagging provides for distinguishing entity boundaries, making it a notable limitation that affects the practical applicability of our framework in entity-dense texts.

The approach to include guidelines and annotations for all entity types does not benefit cases where specific entities are negatively affected by the guidelines. Prior work by SLIMER Zamai et al. (2024) conducted entity-by-entity analysis demonstrating that some entities do not require or benefit from the presence of guidelines and annotations. Since we process entire sentences and extract all entities in a single request, it is difficult to selectively include or exclude guidelines based on specific entity types, which could impact performance for certain categories.

6.3.3 Generalization Constraints

Domain transfer limitations emerge when applying our approach to domains that differ significantly from those represented in our evaluation datasets (shown in the BUSTER dataset analysis). While we demonstrate cross-domain generalization capabilities, the effectiveness may diminish for domains with substantially different linguistic patterns or entity types.

Our comprehensive assessment focuses exclusively on English-language datasets, leaving important questions about multilingual effectiveness unresolved. While the underlying principles of paraphrasing-based augmentation appear theoretically sound across languages, the practical effectiveness may vary significantly for languages with different linguistic structures, morphological complexity, or limited representation in large language model training corpora.

Chapter 7

Conclusions

This chapter provides a comprehensive summary of the research contributions presented in this thesis, while outlining promising directions for future investigation.

7.1 Summary of Contributions

This thesis presents PANER, a paraphrase-augmented framework for NER that addresses data scarcity in low-resource scenarios through two core contributions.

Our first contribution introduces an optimized instruction tuning methodology that combines principles from existing approaches to leverage the extended context windows of modern state-of-the-art language models. We integrate the negative instance inclusion strategy from GNER Y. Ding et al. (2024) with the guideline-centric philosophy of SLIMER Zamai et al. (2024), while proposing a simplified word/tag output format that replaces traditional BIO tagging schemes. Our comparative analysis demonstrates an average of 16.66% improvement when transitioning from BIO to our simplified format, providing empirical evidence that incorporating entity definitions and guidelines significantly benefits NER performance. Through extensive validation, we show this approach achieves competitive results in both zero-shot and few-shot settings.

Our second contribution was the paraphrasing-based data augmentation strategy that leverages

large language models to generate high-quality synthetic training examples. Unlike traditional augmentation approaches, our method preserves entity information while diversifying contextual patterns through controlled paraphrasing. The systematic evaluation demonstrates consistent performance improvements ranging from 6.6% to 8.4% across multiple model architectures in few-shot learning scenarios, with our best configuration achieving an average F1 score of 80.1% on CrossNER datasets.

These contributions collectively demonstrate that competitive NER performance can be achieved through strategic data augmentation combined with optimized instruction tuning, reducing computational barriers while maintaining effectiveness across both zero-shot and few-shot scenarios in low-resource environments.

7.2 Future Work

The findings and limitations identified in this thesis point toward several promising directions for future research that could significantly extend the impact and applicability of paraphrasing-based data augmentation for NER.

Future work can explore more flexible entity augmentation strategies that preserve semantic relationships while allowing controlled entity variations (such as pairing our approach with replacing entities with semantically equivalent alternatives within the same type R. Zhou et al. (2022)) and adaptive paraphrasing approaches that adjust constraint strictness based on sentence complexity and domain characteristics.

The development of domain-adaptive paraphrasing models that can be quickly specialized for specific fields or applications would address current limitations in handling highly specialized terminology. Additionally, exploring ensemble paraphrasing approaches or a multi-stage augmentation pipeline that combines multiple generation strategies could potentially improve both quality and diversity of augmented samples.

Our current correction approach to failed augmentation is very reactive right now and presents a bottleneck. Dynamic methods to enhance the responsiveness to errors can also be explored. Reinforcement learning approaches for optimizing paraphrasing strategies based on downstream task

performance could provide more sophisticated optimization of augmentation quality and error handling.

Another key direction for future work is refining selective guideline inclusion, where entity-specific constraints could be dynamically applied during instruction tuning. Currently, our approach processes entire sentences and extracts all entities in a single request, making it difficult to selectively include or exclude guidelines based on specific entity types that may not benefit from comprehensive annotation guidelines. Investigation into boundary detection mechanisms for nested NER scenarios, where our current word/tag format could be extended to handle overlapping entities and complex entity hierarchies.

Further, while our approach has demonstrated strong performance in English-language datasets, its multilingual effectiveness remains unexplored. A critical next step is to study how paraphrase-based augmentation can be effectively applied to other languages, particularly for morphologically rich languages, where entity boundaries and paraphrasing strategies may require language-specific adaptations.

Finally, broader application research could extend our paraphrasing-based augmentation approach to other natural language processing tasks beyond NER. Investigation of applications to relation extraction, sentiment analysis, text classification, and other sequence labelling tasks could demonstrate the general utility of our methodology. Developing task-agnostic augmentation frameworks based on our paraphrasing principles could benefit the broader NLP community, especially in the wake of LLMs.

These research directions collectively offer substantial opportunities for advancing the field of data augmentation for natural language processing while addressing some of the broader limitations of PANER.

References

Aggarwal, K., Jin, H., & Ahmad, A. (2023, July). Ecg-qalm: Entity-controlled synthetic text generation using contextual q&a for ner. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023*. Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.349/ doi: 10.18653/v1/2023.findings-acl.349

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., ... Malartic, Q. (2023). The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Al-Moslmi, T., Ocaña, M. G., Opdahl, A. L., & Veres, C. (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, *8*, 32862–32881.

axolotl-ai. (2023). https://github.com/axolotl-ai-cloud/axolotl. ([Accessed 15-06-2025])

axolotl-examples. (2023). https://github.com/axolotl-ai-cloud/axolotl/tree/main/examples. ([Accessed 15-06-2025])

Bai, G., Liu, J., Bu, X., He, Y., Liu, J., Zhou, Z., ... Ouyang, W. (2024, August). MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*. Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.401/ doi: 10.18653/v1/2024.acl-long.401

Bikel, D. M., Miller, S., Schwartz, R. M., & Weischedel, R. M. (1998). Nymble: a high-performance learning name-finder. *CoRR*, *cmp-lg/9803003*. Retrieved from http://arxiv.org/

abs/cmp-lg/9803003

Borthwick, A. E. (1999). *A maximum entropy approach to named entity recognition* (Unpublished doctoral dissertation). New York University, Computer Science Department, USA. (AAI9945252)

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... Brahma, S. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, *25*(70), 1–53.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. P. (2011). Natural language processing (almost) from scratch. *CoRR*, *abs/1103.0398*. Retrieved from http://arxiv.org/abs/1103.0398

Dai, X., & Adel, H. (2020, December). An analysis of simple data augmentation for named entity recognition. In D. Scott, N. Bel, & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.343/ doi: 10.18653/v1/2020.coling-main.343

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 4171–4186. Retrieved from https://aclanthology.org/N19-1423/ doi: 10.18653/v1/N19-1423

Ding, B., Liu, L., Bing, L., Kruengkrai, C., Nguyen, T. H., Joty, S., ... Miao, C. (2020, November). DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp).* Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.488/ doi: 10.18653/v1/2020.emnlp-main.488

Ding, Y., Li, J., Wang, P., Tang, Z., Yan, B., & Zhang, M. (2024). Rethinking negative instances for generative named entity recognition. *arXiv preprint arXiv:2402.16602*.

Dubois, Y., Galambosi, B., Liang, P., & Hashimoto, T. B. (2024). Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Fadaee, M., Bisazza, A., & Monz, C. (2017, July). Data augmentation for low-resource neural machine translation. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 567–573). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P17-2090/ doi: 10.18653/v1/P17-2090

Fritzler, A., Logacheva, V., & Kretov, M. (2019). Few-shot classification in named entity recognition task. In *Proceedings of the 34th acm/sigapp symposium on applied computing* (pp. 993–1000).

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... Vaughan, A. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grishman, R., & Sundheim, B. (1995). Message Understanding Conference- 6: A brief history. In *COLING 1995 volume 1: The 16th international conference on computational linguistics.* Retrieved from https://aclanthology.org/C96-1079/

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, *1*(2), 3.

Hu, X., Jiang, Y., Liu, A., Huang, Z., Xie, P., Huang, F., ... Yu, P. S. (2022). Entity-to-text based data augmentation for various named entity recognition tasks. *arXiv preprint arXiv:2210.10343*.

Huang, J., Li, C., Subudhi, K., Jose, D., Balakrishnan, S., Chen, W., ... Han, J. (2020). Few-shot named entity recognition: A comprehensive study. *arXiv preprint arXiv:2012.14978*.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Issifu, A. M., & Ganiz, M. C. (2021). A simple data augmentation method to improve the performance of named entity recognition models in medical domain. In *2021 6th international conference on computer science and engineering (ubmk)* (p. 763-768). doi: 10.1109/UBMK52708.2021.9558986

Jia, C., Liang, X., & Zhang, Y. (2019). Cross-domain ner using cross-domain language modeling.

In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2464–2474).

Jia, C., & Zhang, Y. (2020). Multi-cell compositional lstm for ner domain adaptation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5906–5917).

Katiyar, A., & Cardie, C. (2018). Nested named entity recognition revisited. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies* (Vol. 1).

Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*.

Kruengkrai, C., Nguyen, T. H., Aljunied, S. M., & Bing, L. (2020). Improving low-resource named entity recognition using joint sentence and token labeling. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5898–5905).

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml* (Vol. 1, p. 3).

Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE transactions on knowledge and data engineering*, *34*(1), 50–70.

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2019). A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

Lison, P., Hubin, A., Barnes, J., & Touileb, S. (2020). Named entity recognition without labelled data: A weak supervision approach. *arXiv preprint arXiv:2004.14723*.

Liu, J. (2024). *Structured ouputs with ollama and instructor.* https://python.useinstructor.com/examples/ollama/#further-reading. ([Accessed 15-06-2025])

Liu, J., Pasupat, P., Cyphers, S., & Glass, J. (2013). Asgard: A portable architecture for multilingual dialogue systems. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 8386–8390).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., ... Fung, P. (2021). Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 13452–13460).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ma, X., & Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Mayhew, S., Chaturvedi, S., Tsai, C.-T., & Roth, D. (2019). Named entity recognition with partially annotated training data. *arXiv preprint arXiv:1909.09270*.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on natural language learning at hlt-naacl 2003* (pp. 188–191).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Modal. (2023). https://modal.com. ([Accessed 15-06-2025])

Mollá, D., Van Zaanen, M., & Smith, D. (2006). Named entity recognition for question answering. In *Australasian language technology association workshop* (pp. 51–58).

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, June). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-1202/ doi: 10.18653/v1/N18-1202

Pires, T., Schlinger, E., & Garrette, D. (2019, July). How multilingual is multilingual BERT? In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4996–5001). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1493/ doi: 10.18653/v1/P19-1493

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raschka, S. (2025). *Build a large language model (from scratch)*. Manning Publications.

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*. Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1410/ doi: 10.18653/v1/D19-1410

Sainz, O., García-Ferrero, I., Agerri, R., de Lacalle, O. L., Rigau, G., & Agirre, E. (2023). Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*.

Sang, E. F., & De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Santos, C. N. d., & Guimaraes, V. (2015). Boosting named entity recognition with neural character embeddings. *arXiv preprint arXiv:1505.05008*.

Shang, J., Liu, L., Ren, X., Gu, X., Ren, T., & Han, J. (2018). Learning named entity tagger using domain-specific dictionary. *arXiv preprint arXiv:1809.03599*.

Sun, C., & Yang, Z. (2019). Transfer learning in biomedical named entity recognition: an evaluation of bert in the pharmaconer task. In *Proceedings of the 5th workshop on bionlp open shared tasks* (pp. 100–104).

Tan, C., Qiu, W., Chen, M., Wang, R., & Huang, F. (2020). Boundary enhanced neural span classification for nested named entity recognition. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 9016–9023).

Tjong Kim Sang, E. F. (2002). 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)*. Retrieved from https://aclanthology.org/W02-2024/

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... Azhar, F. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, X., Zhou, W., Zu, C., Xia, H., Chen, T., Zhang, Y., ... Gui, T. (2023). Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J., & Zou, K. (2019, November). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*. Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1670/ doi: 10.18653/v1/D19-1670

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., ... Wei, H. (2024). Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yaseen, U., & Langer, S. (2021, December). Data augmentation for low-resource named entity recognition using backtranslation. In S. Bandyopadhyay, S. L. Devi, & P. Bhattacharyya (Eds.), *Proceedings of the 18th international conference on natural language processing (icon)*. National Institute of Technology Silchar, Silchar, India: NLP Association of India (NLPAI). Retrieved from https://aclanthology.org/2021.icon-main.43/

Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., & Huang, X. (2024). *Llm-da: Data augmentation via large language models for few-shot named entity recognition*. Retrieved from https://arxiv.org/abs/2402.14568

Zamai, A., Zugarini, A., Rigutini, L., Ernandes, M., & Maggini, M. (2024). Show less, instruct more: Enriching prompts with definitions and guidelines for zero-shot ner. *arXiv preprint arXiv:2407.01272*.

Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2023). Gliner: Generalist model for named entity recognition using bidirectional transformer. *arXiv preprint arXiv:2311.08526*.

Zhang, M., Yan, H., Zhou, Y., & Qiu, X. (2023). Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *arXiv preprint arXiv:2305.12217*.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, *28*.

Zhou, R., Li, X., He, R., Bing, L., Cambria, E., Si, L., & Miao, C. (2022, May). MELM: Data augmentation with masked entity language modeling for low-resource NER. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*. Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.160/ doi: 10.18653/v1/2022.acl-long.160

Zhou, W., Zhang, S., Gu, Y., Chen, M., & Poon, H. (2023). Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

Zugarini, A., Zamai, A., Ernandes, M., & Rigutini, L. (2024). Buster: a" business transaction entity recognition" dataset. *arXiv preprint arXiv:2402.09916*.

Appendix A

Paraphrase generation code using Instructor

```
from tqdm import tqdm
class AugmentationOutput(BaseModel):
    variants: List[str] = Field(..., description="List of two scenario-based
    augmented sentences")
client = instructor.from_openai(
    OpenAI(
        base_url="https://ba79772637721.notebooks.jarvislabs.net/v1",
        api_key=jarvis_labs_api_key,
    ),
    mode=instructor.Mode.JSON,
)

output_list = []
for dataset_name in [
    'crossnermusic', 'crossnerpolitics', 'crossnerliterature',
    'mit-movieslash', 'mit-restaurantslash', 'crossnerai'
]:
    print(dataset_name)
    for idx in tqdm(range(len(combined_processed_data[dataset_name]))):
        resp = client.chat.completions.create(
```

```
            model="llama3.3:70b",
            messages=[
                {
                    "role": "user",
                    "content": create_prompt(combined_processed_data[dataset_name
    ][idx]['processed_sent']),
                }
            ],
            response_model=AugmentationOutput,
        )
        combined_processed_data[dataset_name][idx]['variants'] = resp.variants
```

A.1 Axolotl Config for Finetuning Llama 3.1-8B

```
base_model: meta-llama/Llama-3.1-8B-Instruct
model_type: LlamaForCausalLM
tokenizer_type: AutoTokenizer

load_in_8bit: true       # Load model in 8-bit precision to reduce memory usage
load_in_4bit: false      # 4-bit loading disabled; would save even more memory if
     enabled
strict: false

datasets:
  - path: data.jsonl
    ds_type: json
    type: alpaca

dataset_prepared_path:
val_set_size: 0.05
output_dir: ./outputs/lora-out
deepspeed: /workspace/axolotl/deepspeed_configs/zero3_bf16.json  # Zero3 offloads
     optimizer states to reduce GPU memory
```

```
sequence_len: 4096
sample_packing: false
pad_to_sequence_len: true

adapter: lora
lora_model_dir:
lora_r: 32
lora_alpha: 16
lora_dropout: 0.05
lora_target_linear: true
lora_fan_in_fan_out:

gradient_accumulation_steps: 4   # Accumulates gradients over steps to simulate
    larger batch size with less memory
micro_batch_size: 2              # Per-device batch size; kept small to stay
    within memory limits
num_epochs: 1
optimizer: adamw_bnb_8bit        # 8-bit optimizer from bitsandbytes to reduce
    memory usage
lr_scheduler: cosine
learning_rate: 0.0002

train_on_inputs: false
group_by_length: false
bf16: auto
fp16:                          # Mixed precision left empty (disabled); could
    reduce memory if enabled
tf32: false

gradient_checkpointing: true  # Saves memory by recomputing activations during
    backprop
early_stopping_patience:
resume_from_checkpoint:
local_rank:
logging_steps: 1
```

```
xformers_attention:           # Left empty; could enable memory-efficient
    attention
flash_attention: true         # Enables efficient attention with reduced memory
    usage
s2_attention:

warmup_ratio: 0.04
evals_per_epoch: 1
eval_table_size:
eval_max_new_tokens: 128
saves_per_epoch: 1
debug:
deepspeed:
weight_decay: 0.0
fsdp:
fsdp_config:
special_tokens:
   pad_token: <|end_of_text|>
```

A.2 Prompt for generating Annotations and Guidelines from SLIMER

```
    {
    named_entity: ACTOR,
    real_name: actor,
    sentences_as_example: [
      {
        sentence: did jane fonda do an period picture,
        entities: [
          jane fonda
        ]
      },
      {
        sentence: i want an r rated four stars western with jean arthur,
        entities: [
```

```
        jean arthur
      ]
    },
    {
      sentence: is there a rated pg 13 movie in the year 1940 s with jordanna
  brewster,
      entities: [
        jordanna brewster
      ]
    }
  ],
  prompt_length: 445,
  output_length: 62,
  gpt_answer: {Definition: 'actor' refers to individuals who perform in films,
   television shows, or theater productions., Guidelines: Avoid labeling
  characters or fictional individuals. Exercise caution with ambiguous names
  that may refer to both real persons and fictional characters, such as 'Robin
   Hood' or 'James Bond'.},
  finish_reason: stop
},
```

Appendix B

Modal Platform Images

This appendix contains platform screenshots and deployment-related figures from the Modal environment used in our experiments.

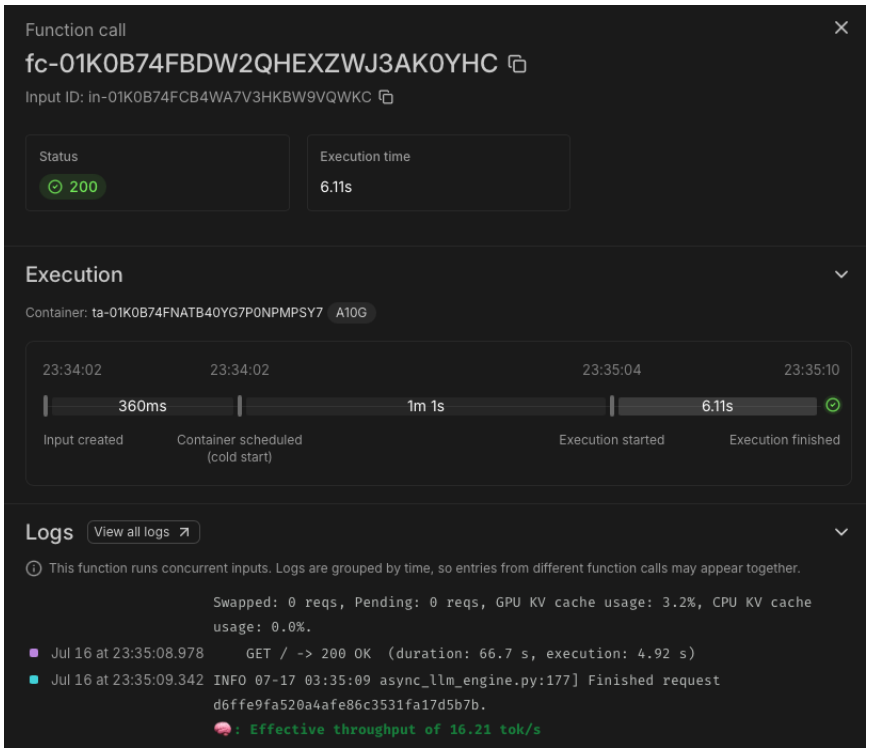

Figure B.1: Elaboration of each call - shows the time taken for execution and throughout

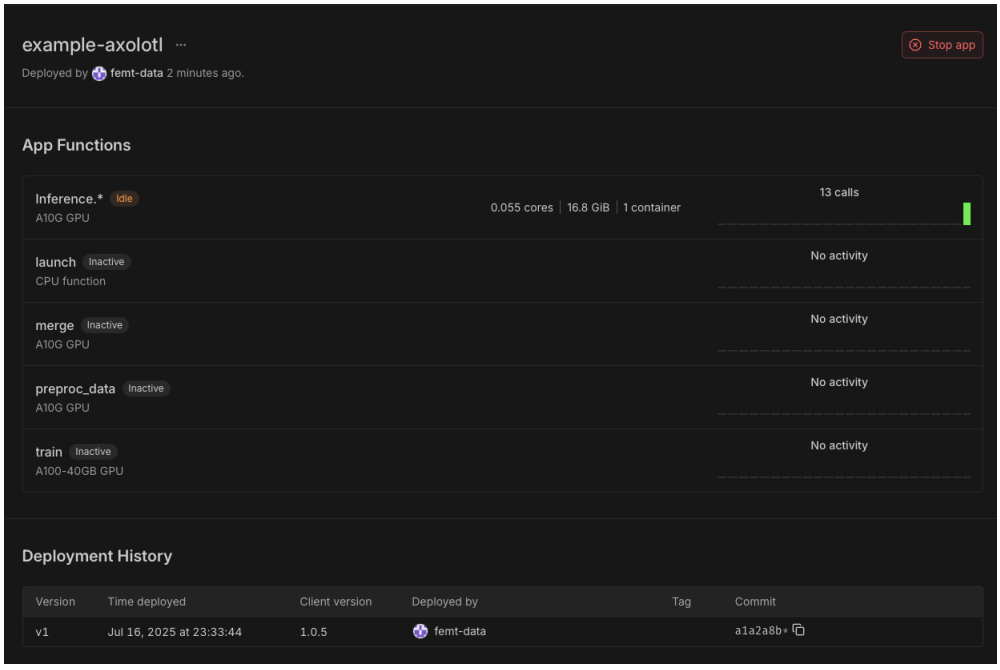

Figure B.2: Modal Training and deployment container - showcases the different function within it and what GPU is being used.

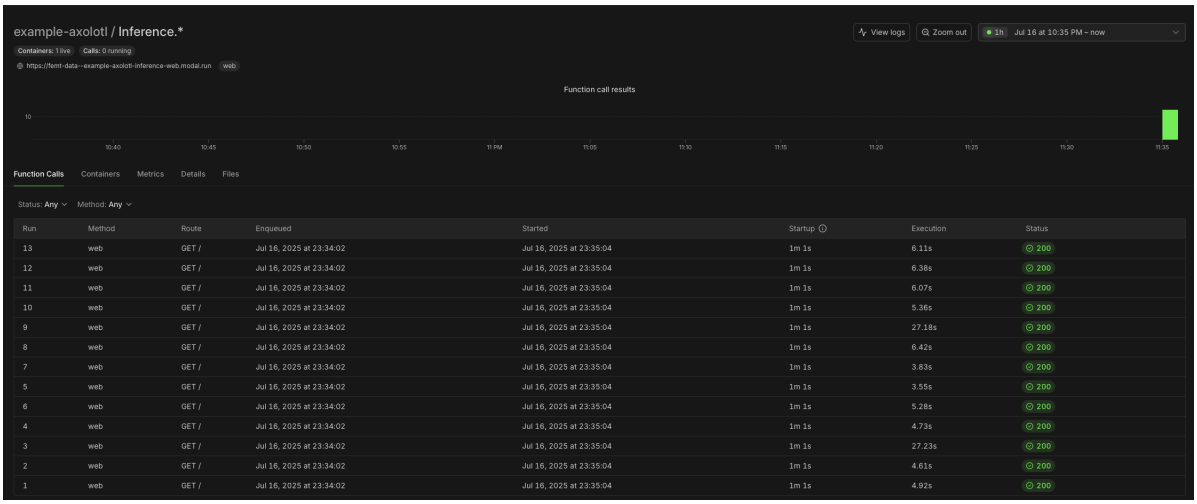

Figure B.3: Inference calls made parallel to deployed Modal - Status 200 means okay