HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

GRADUATION THESIS

Multilingual LLM for Open-Domain Vietnamese Entity Recognition

MAI MINH KHÔI
khoi.mm210492@sis.hust.edu.vn

Program: HEDSPI

Supervisor: Associate Professor Lê Thanh Hương _____

Signature

Department: Computer Science

School: School of Information and Communications Technology

HANOI, 06/2025

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to Assoc. Prof. Lê Thanh Hương for her dedicated supervision, valuable guidance, and constant encouragement throughout the development of this thesis. Her insights and feedback were instrumental in shaping the quality of my work.

I am also deeply thankful to Mr. Nguyễn Văn Tâm, my mentor at Toshiba Software Development VietNam, for his practical advice, technical support, and thoughtful suggestions that helped bridge the gap between academic knowledge and real-world application.

My sincere appreciation goes to my family for their unwavering support and belief in me, and to my friends who were always there with motivation and encouragement.

Last but not least, I thank myself for staying committed through challenges and for the perseverance it took to complete this important milestone in my academic journey.

ABSTRACT

Named Entity Recognition (NER) is a crucial component of Natural Language Processing (NLP), as it involves the identification and classification of entities found within textual data. Traditional NER systems are typically limited to a fixed set of predefined types, while open-domain NER seeks to recognize a broader and more diverse set of entity categories.

Although recent advancements, such as UniversalNER, have shown the promise of using large-scale pre-trained models for open-domain NER, most efforts focus on high-resource languages like English and demand considerable computational resources, making them impractical in low-resource settings.

This thesis addresses these limitations in the context of Vietnamese by applying Low-Rank Adaptation (LoRA) [1] to fine-tune multilingual encoder-decoder models specifically mT0 for open-domain NER. This approach significantly reduces hardware and time requirements while maintaining competitive performance. We also conduct a comprehensive evaluation across various Vietnamese Q&A datasets with both short and long-form answers to assess the models' robustness and generalizability.

Our main contributions are as follows. First, we apply LoRA based fine-tuning to adapt multilingual models for Vietnamese open-domain NER in a resource-efficient manner. Second, we conduct an in-depth empirical analysis of cross-lingual and multitask transfer performance using mT0 on diverse samples labeled types. Finally, we construct a high-quality, open-domain Vietnamese NER dataset with fine-grained annotations to support future research.

These contributions aim to advance open-domain NER for Vietnamese and offer practical insights applicable to other low-resource languages.

Student

(Signature and full name)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Definition
LLM	Large Language Model
LoRA	Low-Rank Adaptation
NER	Named Entity Recognition
NLP	Natural Language Processing
PEFT	Parameter-Efficient Fine-Tuning
Q&A	Question and Answering

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

Named Entity Recognition (NER) is a core task within the broader field of information extraction in Natural Language Processing (NLP). It focuses on detecting and classifying textual spans that refer to real-world entities into predefined categories such as persons, organizations, or locations. NER has long been recognized as a foundational component in many downstream NLP applications, including information retrieval, summarization, question answering, and machine translation. Its effectiveness often reflects a model's ability to incorporate semantic understanding and background knowledge.

While traditional NER systems are typically constrained to a fixed set of entity types, open-domain NER seeks to identify a much broader and more flexible range of entities. Despite its practical significance, open-domain NER remains an under explored area, particularly for low-resource languages like Vietnamese. A major challenge lies in the limited availability of high-quality, annotated datasets that capture the diversity of entities encountered in real-world text.

This thesis addresses the problem of open-domain NER for Vietnamese by investigating the use of multilingual models to bridge the data scarcity gap. Our goal is to build a system that can generalize across a wide array of entity types and adapt effectively to Vietnamese using cross-lingual transfer learning. By reducing reliance on large-scale annotated resources, we aim to make open-domain NER more accessible for low-resource scenarios. The resulting models have the potential to support a variety of NLP applications while requiring minimal task-specific supervision.

1.2 Background and Problems of Research

Named Entity Recognition (NER) is a core task in Natural Language Processing (NLP), supporting applications such as information retrieval, machine translation, and question answering. In its traditional form, NER focuses on extracting entities that fall into predefined types like *Person*, *Organization*, and *Location*. Early methods based on BiLSTM-CRF [2] have shown strong performance across many languages.

The advent of large pretrained language models such as BERT [3], RoBERTa [4], and XLM-R [5] has significantly improved NER systems. These transformer based models benefit from their ability to capture deep contextual representations

and have been widely fine-tuned for entity recognition tasks. For multilingual NER, models such as *xlm-roberta-base-ner-hrl* serve as strong baselines.

However, open-domain NER where entity types are not restricted to a fixed schema presents greater complexity. Earlier approaches leveraged distant supervision or bootstrapping, but were limited by data noise and lack of generalization. The emergence of large language models (LLMs), such as GPT-3 [6] and T0 [7], enabled zero-shot and few-shot NER by leveraging implicit world knowledge. Despite this potential, these models are expensive, opaque, and difficult to deploy in low-resource environments.

More readily available options include instruction-tuned models such as FLAN-T5 [8], along with distilled versions like Alpaca [9] and Vicuna [10]. Nevertheless, their performance still lags behind teacher models, especially for domain-specific tasks [11]. UniversalNER [12] introduced a framework for fine-tuning large language models on broad entity taxonomies and achieved competitive results in English and other high-resource languages. However, its computational demands limit adoption in low-resource contexts.

For Vietnamese, existing NER research remains modest. Benchmarks such as VLSP 2018 [13], VLSP 2021 [14], and PhoNER-COVID19 [15] provide useful datasets, but cover a limited range of entity types. Furthermore, the Vietnamese language presents additional challenges, such as word segmentation ambiguity and a lack of clear word boundaries, which complicate entity recognition.

Several Vietnamese NER tools like VnCoreNLP [16], *ner-vietnamese-electra-base* [17], Underthesea [18], and pyvi [19] have been developed using these datasets. Yet, most are restricted to fixed entity types and underperform in open-domain settings. Additionally, open-domain NER for Vietnamese remains underexplored, with few studies addressing the recognition of emerging or domain-specific entities.

This study aims to address these limitations by fine-tuning multilingual encoder-decoder models mT0 [20] for Vietnamese open-domain NER. To reduce computational overhead, we employ parameter-efficient fine-tuning using LoRA [1]. We also explore cross-lingual transfer learning and multitask learning, enabling the models to generalize better in low-resource scenarios where annotated data is scarce.

1.3 Research Objectives and Conceptual Framework

This thesis is centered on enhancing the performance of Named Entity Recognition (NER) in Vietnamese, particularly in open-domain settings where entity types are diverse and not restricted to predefined categories. The research leverages multilingual encoder-decoder models specifically, mT0 to address the dual challenges of limited annotated Vietnamese data and the inadequate generalization of existing models.

The **primary objective** is to develop an effective approach for recognizing a wide variety of named entities in Vietnamese, by utilizing the transfer learning capabilities of multilingual pretrained models.

To achieve this, the study pursues three key research goals:

- **First**, to assess how well multilingual models such as mT0, which have been pretrained on diverse language corpora, can be adapted to perform NER in Vietnamese. The aim is to evaluate their cross-lingual generalization capabilities, especially when applied to a language with limited annotated resources.
- **Second**, to explore the impact of multi-task fine-tuning on the model's ability to recognize entities more accurately. This involves fine-tuning the models on a combination of English NER datasets and other related tasks, with the hypothesis that knowledge from high-resource languages can enhance performance in Vietnamese.
- **Third**, to curate and construct a new high-quality Vietnamese NER dataset with a focus on open-domain coverage. This dataset is intended not only to support model training and evaluation but also to fill the gap left by current Vietnamese NER datasets, which are often narrow in scope and limited in size.

The conceptual framework guiding this research consists of three stages: (1) model adaptation via fine-tuning using LoRA to reduce computational demands, (2) performance evaluation across tasks and languages, and (3) dataset construction and analysis. These components interact to create a comprehensive pipeline aimed at advancing Vietnamese NER capabilities and contributing new insights to the study of multilingual and low-resource NLP.

1.4 Contributions

This thesis contributes to the advancement of Named Entity Recognition (NER) for low-resource languages, with a specific focus on Vietnamese. The contributions

aim to address limitations of traditional BIO tagging schemes while building NER models that are both accurate and adaptable to diverse use cases.

This thesis has three main contributions as follows:

1. **Proposing a training approach for multilingual large language models in open-domain NER tasks:** The proposed method leverages transfer learning from high-resource languages such as English to Vietnamese, enhancing the model's generalization and flexibility in low-resource environments.
2. **Investigating cross-lingual and multitask learning capabilities of the mT0 model:** This study explores how mT0 can transfer knowledge from English to Vietnamese by evaluating its performance across a range of datasets and task types. The research also identifies the most effective data sources for cross-lingual transfer and presents strategies for optimizing training efficiency in terms of time and computational resources.
3. **Developing a cleaned, open-domain Vietnamese NER dataset:** This newly curated dataset addresses current gaps in Vietnamese NER resources and serves as a solid foundation for training, benchmarking, and advancing future NER-related research in the Vietnamese language.

1.5 Organization of Thesis

This thesis continues with its chapters arranged as outlined below. Chapter 2 presents a comprehensive review of the literature, outlining the research context, prior studies, and foundational theories relevant to the topic. This chapter explores existing approaches to Named Entity Recognition, highlights their applications, and examines the unique challenges associated with this task. Furthermore, it discusses recent progress in multilingual models and the benefits of fine-tuning with multiple tasks.

Chapter 3 details the research methodology, including the design of experiments, the process of preparing the data, the training pipeline for the models, and the metrics employed for performance evaluation.

Chapter 4 reports the experimental outcomes, offering an in-depth evaluation of the model's effectiveness on Vietnamese NER, the influence of multi-task learning, and comparative results against baseline methods.

Finally, Chapter 5 summarizes the main contributions of the study, reflects on its significance, and outlines potential avenues for future exploration.

CHAPTER 2. LITERATURE REVIEW

In this chapter, the thesis provides an overview of the research context and related work on the Named Entity Recognition (NER) task, with a particular focus on low-resource languages such as Vietnamese. The chapter first defines the scope of the research problem, highlighting the characteristics of open-domain NER and the associated challenges.

Next, existing studies are analyzed to identify key trends in the field, ranging from traditional approaches to more recent developments involving large pre-trained language models and multilingual transfer learning. This section emphasizes the strengths and limitations of each approach, thereby identifying the research gap that this thesis aims to address.

Additionally, the chapter presents the fundamental concepts required for building and training NER models, including essential machine learning principles, encoder-decoder architectures, transfer learning techniques, and multi-task learning. These foundational topics will serve as the basis for the methodology and experimental chapters that follow.

2.1 Scope of Research

Named Entity Recognition (NER) is a core task in natural language processing that involves identifying and classifying entities such as people, organizations, locations, and other semantic categories within unstructured text. While recent advancements in deep learning have led to highly accurate NER models in resource-rich languages like English, the development of robust NER systems for low-resource languages such as Vietnamese still faces numerous challenges. These include limited annotated data, restricted domain coverage, and difficulty in generalizing to new or unseen entity types.

Traditional NER systems are often constrained by predefined entity categories, making them less adaptable to open-domain scenarios where the diversity of entity types can vary significantly depending on context. Recent research has begun to address this limitation by exploring open-domain NER, where models are expected to recognize a broader and more flexible range of entity types without being restricted to a fixed schema. However, many of these approaches either depend on large-scale English models or assume access to vast computational resources, which limits their feasibility in low-resource settings.

This thesis investigates an alternative approach to open-domain NER by

leveraging mT0 [20], a multilingual encoder-decoder model pre-trained on a wide range of tasks in a variety of languages. Unlike encoder-only models, mT0 allows for task formulation in a natural instruction format, which is particularly useful in low-resource and cross-lingual settings. The study aims to evaluate how well mT0 can be fine-tuned using English data and then transferred to Vietnamese for NER tasks without relying on extensive Vietnamese annotations.

A central objective of this research is to examine the effectiveness of multi-task fine-tuning in improving the generalization ability of mT0 across various types of entities, particularly in open-domain contexts. By training the model simultaneously on multiple NER datasets or task variants, this approach is expected to help the model capture diverse patterns and improve performance on underrepresented entity types.

Through this exploration, the thesis seeks to contribute to the ongoing effort of building practical, scalable, and language-agnostic NER systems for low-resource languages. The research is grounded in the hypothesis that with the right transfer learning strategy and multi-task training setup, even compact multilingual models like mT0 can deliver strong NER performance in Vietnamese without extensive in-language supervision.

2.2 Related Works

Named Entity Recognition (NER) has long been a central task in natural language processing, with early progress driven by rule-based systems and statistical learning methods. With the rise of large-scale pre-trained language models, especially in English, models such as BERT [21] and GPT-3 [6] have become dominant in achieving state-of-the-art performance on NER tasks involving predefined entity types. These models benefit from rich annotated corpora, enabling them to generalize across various domains while maintaining high precision.

In contrast, Vietnamese NER research, although progressing steadily, remains constrained by limited annotated resources. Benchmarks from the VLSP shared tasks and contributions from research labs such as VinAI have pushed the performance boundaries for Vietnamese NER. However, these efforts are largely focused on closed-domain settings, with entity categories restricted to fixed ontologies such as person, organization, and location. As a result, many existing systems lack flexibility when applied to open-domain contexts, where entity boundaries and types are not strictly defined in advance.

A notable recent shift in NER research is the move toward open-domain entity

recognition, which aims to identify a broader and more dynamic range of entity types beyond fixed schemas. Projects like UniversalNER [22] have demonstrated that instruction-tuned models, when combined with targeted distillation techniques, can effectively recognize diverse entities across multiple domains. By compressing knowledge from large-scale models like ChatGPT into smaller architectures such as LLaMA [23], UniversalNER [22] achieved strong results even without explicit supervision. However, these methods remain predominantly focused on English and often require significant computational resources, limiting their practicality in low-resource language settings like Vietnamese.

To address this gap, recent studies have explored the use of multilingual instruction-tuned models, among which mT0 has emerged as a particularly promising candidate. Unlike encoder-only models, mT0 follows a sequence-to-sequence architecture and is trained on a broad mixture of tasks framed as natural language instructions. This setup enables it to generalize across diverse NLP tasks and languages, even in zero-shot or few-shot scenarios. Importantly, mT0's design supports transfer learning from high-resource to low-resource languages by aligning instruction formats rather than relying solely on token-level similarity.

Although mT0 was not originally designed for NER, recent explorations have demonstrated its adaptability to entity recognition tasks when properly fine-tuned. Its capability to handle task prompts as part of the input makes it well-suited for open-domain scenarios, where flexibility and instruction-awareness are critical. Furthermore, multi-task fine-tuning - where NER is trained alongside other classification or extraction tasks - has shown promise in improving generalization, particularly when annotated data is scarce in the target language.

Building upon these findings, this thesis proposes to investigate mT0's capacity for open-domain NER in Vietnamese through cross-lingual transfer from English datasets. By focusing on smaller, efficient variants of mT0, the research aims to balance performance with accessibility. The study contributes to filling the gap in open-domain NER for low-resource languages, offering insights into the practical deployment of multilingual models under realistic constraints.

2.3 Transformer

The Transformer [24] architecture, first introduced by Vaswani et al. in 2017, has revolutionized the field of natural language processing (NLP). Unlike traditional recurrent neural networks (RNNs) [25] or long short-term memory networks (LSTMs) [26], which process sequences sequentially, the Transformer operates entirely through self-attention mechanisms, allowing for

greater parallelization and more effective handling of long-range dependencies.

At its core, the Transformer [24] consists of two main components: an encoder and a decoder, each composed of multiple identical layers. Each encoder layer includes a multi-head self-attention mechanism followed by a position-wise feed-forward neural network. The decoder layers similarly contain self-attention, encoder-decoder attention, and feed-forward components. Positional encoding is added to the input embeddings to preserve the order of the tokens, which is crucial since the Transformer lacks an inherent sense of sequence.

One of the key innovations of the Transformer is the multi-head self-attention mechanism, which enables the model to attend to different parts of a sequence simultaneously. This mechanism computes attention scores between all token pairs in a sequence, capturing contextual relationships without relying on recurrence. This parallelization not only speeds up training but also allows for more flexible representation learning.

Since its introduction, the Transformer has become the foundation of nearly all modern pre-trained language models, such as BERT [21], GPT [6] , T5 [27], and mT0 [20]. These models differ in how they utilize the Transformer: BERT uses only the encoder, GPT uses only the decoder, while T5 [27] and mT0 [20] use the full encoder-decoder architecture. The encoder-decoder design of mT0 [20] is particularly beneficial for tasks framed as instruction-based learning, including Named Entity Recognition in open-domain settings.

In the context of this thesis, the Transformer architecture underpins the mT0 model, which will be leveraged to perform cross-lingual and open-domain NER. Its attention-based design supports better generalization across languages and tasks, especially when combined with fine-tuning strategies such as multi-task learning. Understanding the Transformer is therefore essential for both the implementation and analysis of the proposed methodology in this work.

2.4 Transfer Learning

Transfer learning has become a cornerstone of modern natural language processing (NLP), dramatically improving performance across a wide range of tasks and languages. At its core, transfer learning refers to the process of leveraging knowledge gained from one domain or task and applying it to another, often distinct, domain or task. This paradigm is particularly valuable in low-resource scenarios, where annotated datasets are scarce or unavailable. For languages like Vietnamese, which lack the large-scale labeled corpora available for English, transfer learning offers a practical and powerful solution to building effective NLP

systems.

Pre-training and Fine-tuning Paradigm

In NLP, transfer learning typically follows the *pre-training and fine-tuning* paradigm. In this setup, a model is first pre-trained on a large, diverse corpus of text data to learn general language representations. The pre-training objective varies depending on the model architecture: for example, masked language modeling (MLM) [3] is used in BERT [21], autoregressive language modeling in GPT [6], and sequence-to-sequence denoising objectives in T5 [27] and mT0 [20]. This stage allows the model to develop a broad understanding of language semantics, syntax, and discourse structure.

After pre-training, the model is *fine-tuned* on a downstream task such as sentiment analysis, question answering, or NER. Fine-tuning involves additional training on a smaller, task-specific dataset, during which the model adapts its general language knowledge to the requirements of the specific task. Importantly, because much of the linguistic structure has already been captured during pre-training, fine-tuning typically requires significantly less data and computational resources.

Multilingual and Cross-lingual Transfer

In recent years, transfer learning has evolved to encompass *multilingual* and *cross-lingual* transfer learning. These approaches involve training models on multiple languages simultaneously, enabling the sharing of knowledge across linguistic boundaries. Models such as mBERT [3], XLM-R [28], mT5, and mT0 [20] are trained on large multilingual corpora and are capable of performing tasks in languages they were not explicitly fine-tuned on - a phenomenon known as *zero-shot* or *few-shot transfer*.

Cross-lingual transfer is especially relevant for low-resource languages like Vietnamese. By pre-training on high-resource languages (e.g., English) and then fine-tuning on limited Vietnamese data, models can transfer structural and semantic knowledge between languages. This is possible due to shared vocabulary, similar grammatical constructs, or even implicit alignment learned during pre-training. In multilingual encoder-decoder models like mT0, instruction-based pre-training further strengthens this capability by encouraging the model to generalize task behavior across languages.

Transfer Learning for NER

NER is a sequence labeling task that greatly benefits from transfer learning, especially in multilingual and open-domain settings. Conventional approaches to NER typically rely on substantial annotated datasets, where entities are labeled according to a predefined set of categories such as PERSON, LOCATION, or ORGANIZATION. However, in open-domain NER, the set of entity types is not predetermined, and the data often come from diverse domains with varying formats. This makes pre-training on generic text and fine-tuning on flexible, open-ended annotations an ideal setup.

Recent research has shown that transfer learning from large English NER datasets can significantly improve performance on NER tasks in other languages. For instance, by fine-tuning multilingual models on English NER tasks and then transferring to Vietnamese with limited additional training, models can generalize entity representations across languages. Instruction-based models like mT0 are particularly suited for this, as they treat NER as a text-to-text task guided by task-specific prompts, making them inherently adaptable.

Instruction Tuning and Multi-task Fine-tuning

An important extension of transfer learning is *instruction tuning*, where the model is trained to follow natural language prompts describing a task. The mT0 model used in this thesis was trained with instruction tuning across various tasks and languages. This framework enables the model to understand what is expected based on the instruction format and to apply its knowledge in flexible ways.

Additionally, this research explores *multi-task fine-tuning*, where the model is trained on multiple NER-related tasks simultaneously. The rationale is that by learning from diverse but related tasks (e.g., nested NER, span classification, entity linking), the model can develop more robust and generalizable representations of entities. This approach is particularly effective in open-domain NER, where the ability to adapt to novel entity types is essential.

Benefits and Challenges

The main advantages of transfer learning in this context include:

- **Data efficiency:** Reduces the amount of annotated Vietnamese data required.
- **Cross-lingual generalization:** Leverages high-resource languages to improve performance in low-resource settings.
- **Model reusability:** Allows pre-trained models to be adapted for various tasks with minimal modification.

- **Open-domain flexibility:** Facilitates entity recognition in diverse and unstructured datasets.

However, transfer learning is not without challenges. One concern is *negative transfer*, where knowledge from the source domain may interfere with performance in the target domain. Another challenge is *domain mismatch*, particularly when pre-training and fine-tuning data differ significantly in style or content. For low-resource languages, tokenization and vocabulary coverage can also limit performance.

Despite these challenges, transfer learning remains a foundational technique in this thesis. Through strategic fine-tuning of the mT0 model - leveraging both cross-lingual and multi-task learning - this research aims to push the boundary of open-domain NER in Vietnamese, demonstrating that scalable and effective entity recognition is achievable even in low-resource settings.

2.5 Large Language Models

Large Language Models (LLMs) have become a foundational component in many natural language processing (NLP) tasks, including Named Entity Recognition (NER). Built primarily on the Transformer architecture, these models are pre-trained on massive text corpora using self-supervised learning objectives, enabling them to capture rich contextual and semantic information.

LLMs can be categorized into encoder-only models (e.g., BERT), decoder-only models (e.g., GPT), and encoder-decoder models (e.g., T5, mT0). For NER tasks, encoder-decoder models like mT0 [20] offer greater flexibility, especially in scenarios involving instruction tuning or multitask learning.

Multilingual LLMs such as mT0 [20] are particularly useful for low-resource languages. By leveraging large-scale multilingual training and natural language prompts, mT0 [20] can generalize across languages and tasks, making it a suitable choice for open-domain NER. Its instruction-tuned setup allows it to handle diverse entity types without needing rigid schema design.

In this research, mT0 [20] is selected for its multilingual capabilities and efficiency, enabling the development of a practical NER model for Vietnamese that balances performance with accessibility. While LLMs bring challenges in terms of computational cost and prompt sensitivity, their flexibility and adaptability make them powerful tools in addressing the limitations of traditional NER systems.

2.6 Sequence-to-Sequence Model: T0

T0 is a sequence-to-sequence (seq2seq) model derived from the T5 architecture, which reformulates all NLP tasks into a unified text-to-text format. Built as an encoder-decoder Transformer, T0 processes natural language instructions and generates corresponding textual outputs, making it highly flexible and task-agnostic.

Unlike traditional classification-based models used in Named Entity Recognition (NER), which rely on assigning fixed labels to tokens, T0 predicts textual spans directly. This makes it particularly suitable for open-domain NER, where entity categories may not be predefined, and output spans are more naturally expressed in free-form text.

T0 is instruction-tuned, meaning it is fine-tuned using a diverse collection of NLP datasets presented in the form of natural language prompts. This allows the model to learn to follow various instructions across tasks, increasing its adaptability and generalization capability - especially in multilingual and low-resource settings.

In this thesis, T0 is employed to tackle Vietnamese NER in an open-domain context. The sequence-to-sequence design of the model enables it to generate entity spans directly based on context and prompt, while its multilingual pretraining and instruction tuning allow it to transfer knowledge from English and other high-resource languages to Vietnamese. Furthermore, the use of smaller variants of T0 ensures a balance between model performance and computational efficiency.

Overall, T0 provides a powerful and flexible foundation for building robust NER systems in scenarios where annotated data is limited and entity types are not strictly constrained.

2.7 Multilingual Large Language Models: mT0

Multilingual language models have become a cornerstone in modern Natural Language Processing (NLP), especially in applications involving low-resource languages. Among these, mT0 stands out as a multilingual variant of the T0 model, designed to handle a wide range of NLP tasks across different languages. mT0 retains the sequence-to-sequence architecture of T5 and T0, but it is pretrained and instruction-tuned on multilingual datasets, making it particularly suited for cross-lingual tasks such as Named Entity Recognition (NER).

The core innovation of mT0 lies in its combination of multilingual pretraining and task instruction fine-tuning. During pretraining, the model is exposed to

massive amounts of data from over 50 languages, enabling it to learn shared representations across linguistic boundaries. This is followed by instruction tuning, where the model is trained to follow prompts written in natural language for a variety of tasks. This dual approach not only strengthens the model's multilingual understanding but also improves its generalizability to unseen tasks and languages.

In the context of NER, mT0 offers several advantages. First, its multilingual capacity allows the model to leverage annotated data from high-resource languages (e.g., English) to benefit low-resource ones (e.g., Vietnamese). This is particularly valuable for Vietnamese, where high-quality annotated datasets are scarce and limited in scope. Second, by treating NER as a text generation task rather than token classification, mT0 can handle open-domain scenarios where entity types may be varied and not predefined. This generation-based formulation aligns well with the flexible nature of open-domain entity recognition.

Moreover, the model's instruction-following capability plays a critical role in aligning task requirements with model behavior. For example, mT0 can be prompted with instructions such as "Identify all people mentioned in the following sentence" or "List the organization names found in the text," which allows for adaptable and interpretable interactions. This is in contrast to traditional models that require rigid label definitions and cannot easily accommodate variations in entity types or task formats.

Another practical strength of mT0 is the availability of multiple model sizes (e.g., mT0-small, mT0-base, mT0-large), making it suitable for research environments with limited computational resources. In this thesis, smaller variants of mT0 are employed to balance between performance and efficiency. Through fine-tuning on both English and Vietnamese data, the model can be adapted to the specific characteristics of Vietnamese texts while benefiting from the transfer of linguistic knowledge learned from other languages.

In summary, mT0 represents a powerful approach for addressing open-domain NER in low-resource settings. Its multilingual, instruction-tuned, sequence-to-sequence design enables flexible entity extraction and robust cross-lingual transfer, which are essential for building practical and scalable NER systems in Vietnamese and other under-resourced languages.

2.8 Low-Rank Adaptation (LoRA)

Fine-tuning large language models (LLMs) is often computationally expensive, requiring substantial hardware resources and long training times. To address this challenge, Low-Rank Adaptation (LoRA) [1] was introduced as an efficient

parameter-efficient fine-tuning (PEFT) [29] technique. Instead of updating all parameters of a pre-trained model, LoRA [1] freezes the original model weights and injects trainable low-rank matrices into each layer of the Transformer architecture, typically within the attention and feed-forward components.

The core idea of LoRA [1] is to approximate the full-rank weight updates using low-rank decomposition, significantly reducing the number of trainable parameters while maintaining competitive performance. This approach is particularly effective when dealing with large models in low-resource environments, as it allows for faster training with limited memory and compute.

In the context of this thesis, LoRA [1] is applied to the mT0-base model to fine-tune it for the open-domain NER task in Vietnamese. By adopting LoRA, the training process becomes more efficient, enabling experimentation with multilingual models under constrained hardware settings. Additionally, LoRA facilitates easier model deployment due to the reduced size of the fine-tuned components, which can be merged with the base model or stored separately.

CHAPTER 3. METHODOLOGY

3.1 Overview

Traditional NER models typically follow sequence tagging formats like IOB or IOB2, where each token in a sentence is labeled to indicate whether it is at the beginning, inside, or outside of a named entity. While this method performs well for predefined sets of entity types, it becomes less effective in open-domain scenarios, where texts may contain a wide variety of entity types, many of which are not explicitly defined in advance.

To address this limitation, this thesis proposes a more flexible formulation: constructing a model that accepts a Vietnamese sentence and a specified entity type as input, and generates the entities of that type present in the sentence. This design enables the model to better interpret both the semantics of the entity type and the contextual cues within the sentence, improving its ability to accurately extract relevant entities. Inspired by UniversalNER [12], this approach simplifies the NER task by focusing on one entity type per inference, which enhances precision, interpretability, and ease of generalization.

In this study, the mT0-base model was selected due to its compatibility with prompt-based generation, its multilingual adaptability, and its prior training on tasks such as Named Entity Recognition (NER). Designed with an instruction-tuning paradigm and supporting zero-shot capabilities, mT0-base aligns well with the objective of approaching the problem without requiring task-specific training data.

Choosing a moderately sized model like mT0-base is also part of the research's resource-optimization strategy. Instead of investing in retraining large-scale models such as LLaMA, which demand substantial computational resources, mT0 offers comparable performance across many tasks - especially in low-resource language settings like Vietnamese.

To further reduce computational costs during fine-tuning, this study employs LoRA (Low-Rank Adaptation) [1], an efficient fine-tuning method that allows updating only a small subset of parameters while keeping the majority of pretrained weights frozen.

Specifically, only 884,736 parameters (approximately 0.1517% of the total 583 million parameters) are updated during the fine-tuning process. This enables the model to adapt quickly and effectively to new tasks while remaining

computationally efficient and suitable for environments with limited memory and processing power.

One of the major challenges in Vietnamese natural language processing is the scarcity of high-quality training data compared to resource-rich languages like English. To address this, the approach in this study leverages the abundance of English data by incorporating Vietnamese datasets that serve as a semantic bridge - specifically, Question Answering (QA) and Machine Translation (MT) datasets.

These datasets were selected because they not only capture rich semantic structures in Vietnamese but also enable the model to learn cross-lingual semantic alignments between English and Vietnamese. QA datasets are particularly valuable for training models to understand context, perform information retrieval, and reason across texts - core competencies for a wide range of NLP tasks. Meanwhile, MT data reinforces the model's bilingual representational capacity, allowing it to better comprehend Vietnamese by grounding it in high-quality English-language representations.

This strategy acts as a form of knowledge transfer, helping the model extend its understanding to Vietnamese without requiring full retraining on large-scale monolingual Vietnamese data.

The methodology consists of two primary phases. In the first phase, the mT0-base model is fine-tuned using LoRA [1] on English open-domain NER datasets to take advantage of cross-lingual transfer learning. In the second phase, the model is further fine-tuned and evaluated on Vietnamese data, with an optional exploration of multi-task fine-tuning strategies to enhance generalization and robustness. The performance is measured using the F1 score on both English and Vietnamese datasets, aiming to evaluate the model's effectiveness in low-resource, open-domain NER scenarios.

3.2 Data Preparation

This section describes the datasets used during the fine-tuning and training process. A total of eight datasets are employed for fine-tuning, including five English and Vietnamese open-domain NER datasets, two Vietnamese question-answering datasets, and one English-Vietnamese machine translation dataset. The evaluation process in this study involves two standard Vietnamese NER datasets, namely PhoNER_COVID19 [15] and VLSP NER 2018 [30].

3.2.1 English open-domain NER Dataset

In this thesis, the English open-domain NER dataset is constructed using data derived from the Pile corpus, inspired by the methodology proposed in the UniversalNER [12] framework. Specifically, 50,000 English-language text segments were selected across a wide range of domains and were subsequently annotated with named entities using a large language model (ChatGPT-3.5), without predefined constraints on entity categories. This open-ended setting enables the emergence of diverse entity types reflective of natural language use.

To align the dataset with instruction-based learning paradigms, each entity type was associated with a natural language prompt. These prompts followed a consistent structure such as *"What describes [entity type] in the text?"*, (see figure 3.1) , thereby framing the task as a form of question-answering suitable for few-shot learning scenarios. This design encouraged the model to extract specific entities given a query and a contextual passage, making it ideal for use with instruction-following models.

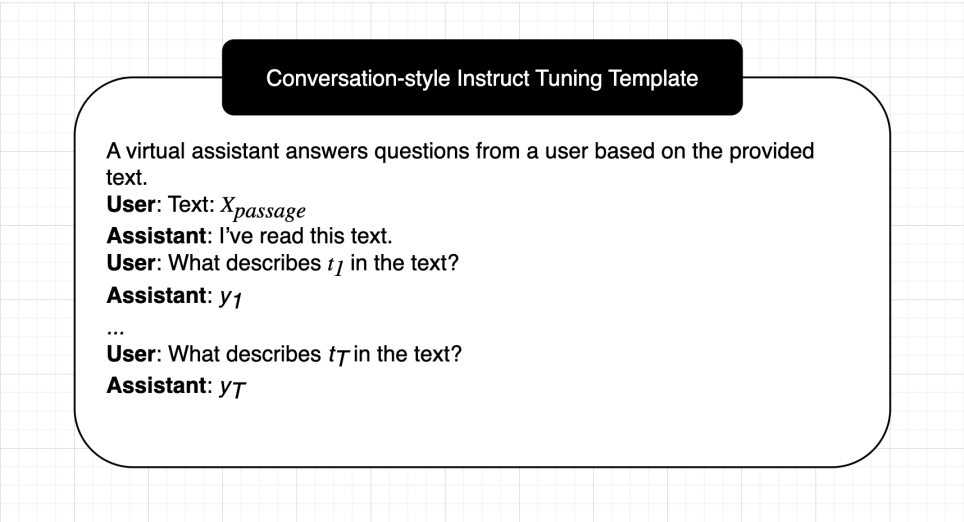


Figure 3.1: Conversation-style Prompt Template

Since this thesis employs sequence-to-sequence architectures such as mT0, the original conversational-style data was reformatted into a structure more compatible with encoder-decoder training. Drawing inspiration from Alpaca-style [31] prompt engineering (see figure 3.2), each original passage was split into multiple samples, each consisting of a unique entity-type query, the original passage as input, and a list of corresponding entity mentions as the output.

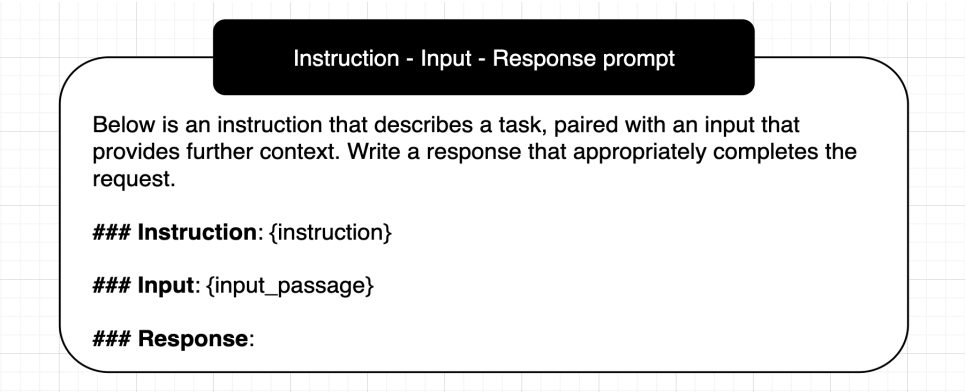


Figure 3.2: Instruction-style (Alpaca-like) Prompt Template

This reformatted dataset resulted in a total of 358,181 samples. To maintain consistency, all entity type labels were converted to lowercase. Additionally, to ensure compatibility with the token limits of the model (maximum 1024 tokens for input and output combined), samples exceeding this threshold were excluded. The final dataset used for training contained 358,002 samples. Each instance consisted of a prompt (including the query and the passage) and an output string listing the extracted entity mentions. This structured format is well-suited for training sequence-to-sequence models on open-domain NER tasks. Below is a sample of the dataset as shown in figure 3.3.

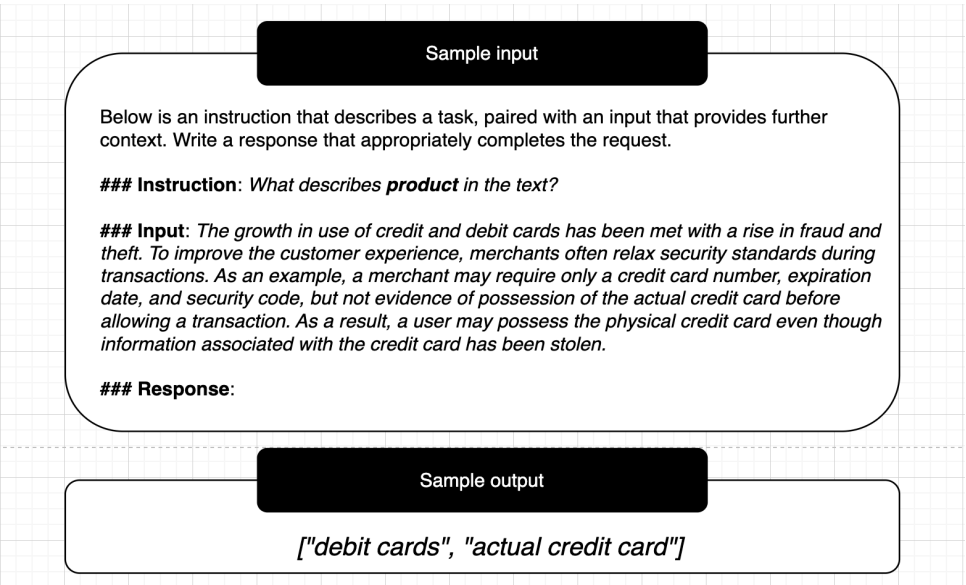


Figure 3.3: An Example of English NER Data

3.2.2 Vietnamese open-domain NER Dataset

To optimize the data generation cost, I opted for LLaMA 3 70B 8192 - an open large language model [32] developed by Meta - instead of relying on commercial APIs such as ChatGPT-3.5. The source data was sampled from the BKAI News

Corpus [33], a large-scale Vietnamese news collection published by the Foundation Models Lab at BKAI-HUST. This corpus, which comprises around 32 million news articles, was compiled by merging newly collected news with the existing Binhvq News Corpus [34], followed by fuzzy deduplication to eliminate redundant entries.

From this corpus, I extracted 6,600 document samples to construct a Vietnamese open-domain NER dataset. These samples were concatenated and segmented into approximately 16,000 passages, each consisting of 150 to 256 tokens. Care was taken to ensure that each passage included at least one complete sentence to preserve the natural structure of the text.

To annotate the entities, I employed a prompting strategy inspired by the UniversalNER [12] framework, modified specifically to handle Vietnamese input (refer to figure 3.4). The generation temperature was fixed at 0 to enforce output consistency. This step resulted in 16,000 annotated examples, where each sample comprises an input passage and the corresponding entity list predicted by LLaMA 3.

Following the same postprocessing procedure used for the English dataset, I transformed the annotated Vietnamese data into instruction-style samples, each focusing on a single entity type. This conversion yielded 46,169 distinct samples, suitable for training or evaluating instruction-tuned NER models.

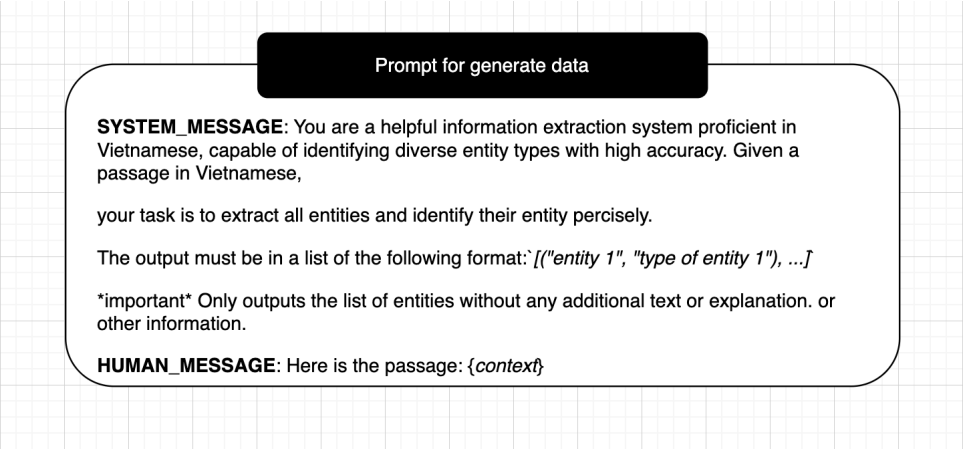


Figure 3.4: Data Construction With LLaMA 3 Prompt Template

The entity types produced by LLaMA 3 were notably varied, with several label variants representing the same conceptual category but differing in format and structure. Some were labeled using code-like syntax, such as *"entity_type:organization"* or *"entity_type:field_of_study"*, while others were malformed. For example, *"_organization"*, *".organization"*, or abbreviated as *"org"* instead of the full form *"organization"*. Additionally, entity labels originally in

Vietnamese were translated into English to ensure consistency across the dataset. These irregular formats were cleaned and normalized through pre_processing to make the entity types more intuitive and user-friendly. Furthermore, entity labels written in other languages, such as Chinese, as well as those with ambiguous or unclear meanings, were manually inspected and removed. Finally, samples containing hallucinated entities - those that did not actually appear in the input text - were also filtered out from the dataset.

To maintain consistency with the English open-domain NER dataset, a final filtering step was applied to discard any samples in which either the input text or the associated labels exceeded 1024 tokens. This step was crucial to ensure the dataset remained suitable for training sequence-to-sequence models. After filtering, the resulting dataset contained a total of 45,834 examples, encompassing 2190 distinct entity types drawn from a broad array of domains.

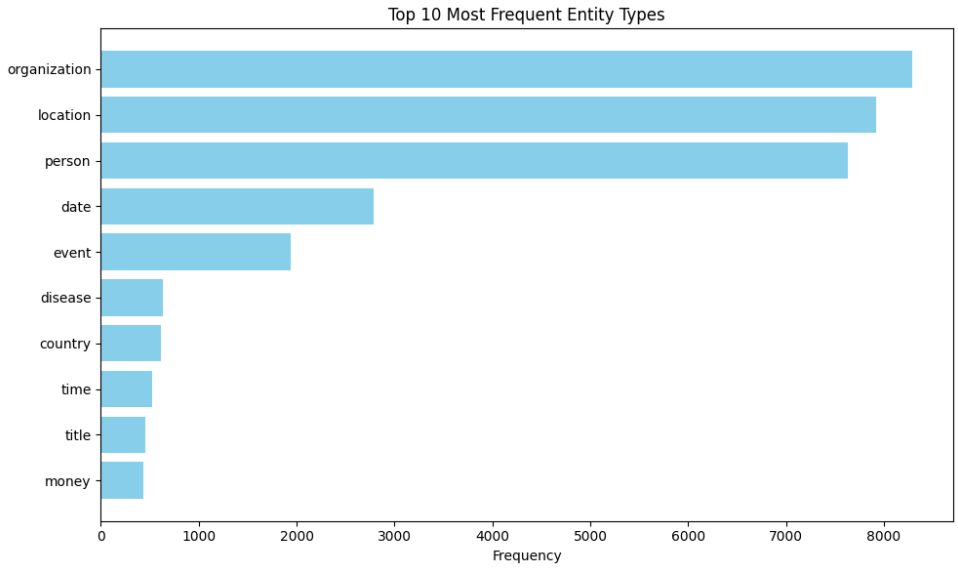


Figure 3.5: Top 10 Most Frequent Entity Types

Based on the *figure 3.5* chart, it is evident that the dataset predominantly focuses on real-world entities, with *organization*, *location*, and *person* appearing most frequently. This suggests that the data likely centers around news events, institutional activities, or personal profiles. The consistent presence of *date* and *time* entities also highlights a strong temporal component, indicating the dataset's emphasis on event-based or time-sensitive information. Interestingly, the relatively high frequency of *disease* points to a potential relevance to health or medical domains. However, there is a noticeable imbalance in entity distribution, as the top three types occur far more frequently than others like *title* or *money*, which may pose challenges for balanced NER training. The presence of *event* and

country entities further reinforces the dataset's context-rich nature, aligning with themes of geopolitical or socio-cultural reporting. Figure 3.6 shows an example of Vietnamese NER data.

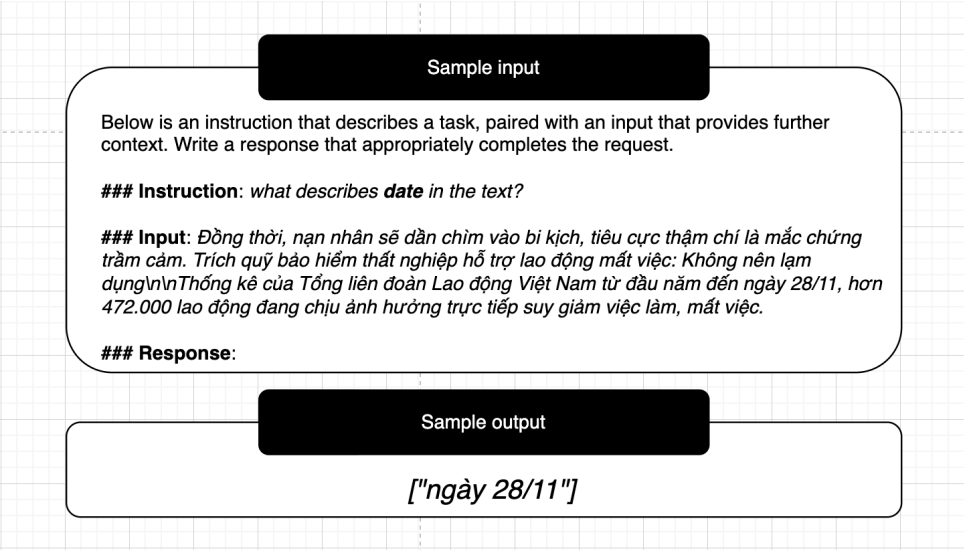


Figure 3.6: An Example of Vietnamese NER Data

3.2.3 Instruction Tuning Dataset

To enhance the model's performance on Vietnamese open-domain Named Entity Recognition (NER), I conducted instruction tuning using three publicly accessible datasets: vi-alpaca [35], UIT-ViQuAD2.0 [36], and VinAI_PhoMT [37]. Although these datasets were originally constructed for different purposes, they each contribute uniquely to improving the model's ability to interpret and respond to natural language instructions - a crucial aspect of instruction-driven NER frameworks.

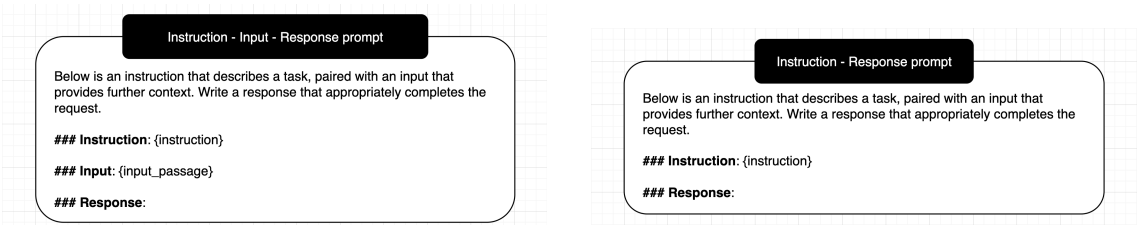
The vi-alpaca dataset, adapted by the BKAI Foundation Models Lab from the original Stanford Alpaca project, includes a broad collection of instruction–input–output examples covering diverse NLP tasks such as classification, summarization, and transformation. While it does not explicitly target named entity recognition, its wide variety of structured prompts helps build the model's general understanding of how to follow instructions and generate task-specific responses. This foundational instruction-following ability is essential for developing effective instruction-style NER prompts.

Next, I explored UIT-ViQuAD2.0, a Vietnamese machine reading comprehension dataset built upon Wikipedia content. It consists of passages accompanied by questions and concise answers. Despite being designed for question answering, many of its answers consist of individual named entities

(e.g., people, organizations, locations), which aligns well with the requirements of open-domain NER. By converting the QA format into instruction–response pairs, the dataset becomes a useful resource for training the model to extract relevant entities from context in an interactive and instruction-guided manner.

Lastly, the VinAI_PhoMT dataset, developed by VinAI Research, provides a substantial collection of high-quality English-to-Vietnamese translation pairs. While its primary focus is on machine translation, it serves a supportive role in instruction tuning by reinforcing the model's understanding of Vietnamese sentence structure and vocabulary. Through exposure to bilingual data, the model is better positioned to process Vietnamese prompts and generate fluent responses, which is particularly beneficial when constructing multilingual or zero-shot NER systems. During instruction tuning, this dataset was reformatted into translation-style prompts and responses, effectively simulating instruction-based learning in a multilingual context.

In preparation for fine-tuning, a standardized prompt template (as illustrated in figure 3.7) was applied to each dataset. A notable challenge with the instruction-tuning corpora is the absence of input fields in some examples. To address this, two variations of the prompt were employed - one that includes an input and one that does not - to ensure flexible formatting. For consistency, all prompts were written in English, regardless of the language of the underlying content, allowing the model to generalize instruction-following behavior across different linguistic settings. Finally, examples exceeding the maximum input length of 1024 tokens were removed to ensure compatibility with the sequence-to-sequence model architecture.



(a) Instruction - Input - Response Prompt **(b)** Instruction - Response Prompt

Figure 3.7: Prompt Template

3.2.4 Test Dataset: PhoNERCOVID_19 and VLSP 2018 NER

In this section, I will introduce two datasets used for few-shot learning and evaluation: PhoNER_COVID19 [15] and VLSP NER 2018 [30].

The PhoNER_COVID19 dataset, introduced by VinAI, focuses on Vietnamese text related to the COVID-19 pandemic and includes manually annotated entities

tailored for health-related contexts. It provides a valuable resource for evaluating models' performance in domain-specific named entity recognition tasks, especially in the context of health crises. This dataset contains a diverse range of entity types, including newly defined categories relevant to pandemics, making it a suitable benchmark for testing the generalization capabilities of NER models in specialized scenarios. In this study, it is utilized as a benchmark to evaluate the model's performance on domain-specific named entities.

The VLSP 2018 NER dataset was released as part of the 2018 shared task by the Vietnamese Language and Speech Processing (VLSP) community, which provides public reference corpora for Natural Language Processing (NLP) research. This dataset addresses the NER task with a focus on four fundamental and widely recognized entity types: LOCATION, ORGANIZATION, PERSON, and MISCELLANEOUS. It is employed in this thesis to assess the model's general-purpose performance on a broad range of common entity types typically found in news articles and general texts.

3.3 Base Vietnamese Instruction Tuning Dataset Selection

The first step entailed identifying the optimal dataset for the subsequent phases via the fine-tuning procedure illustrated in the figure 3.8 below. The mT0-base model was selected as the base model for this purpose.

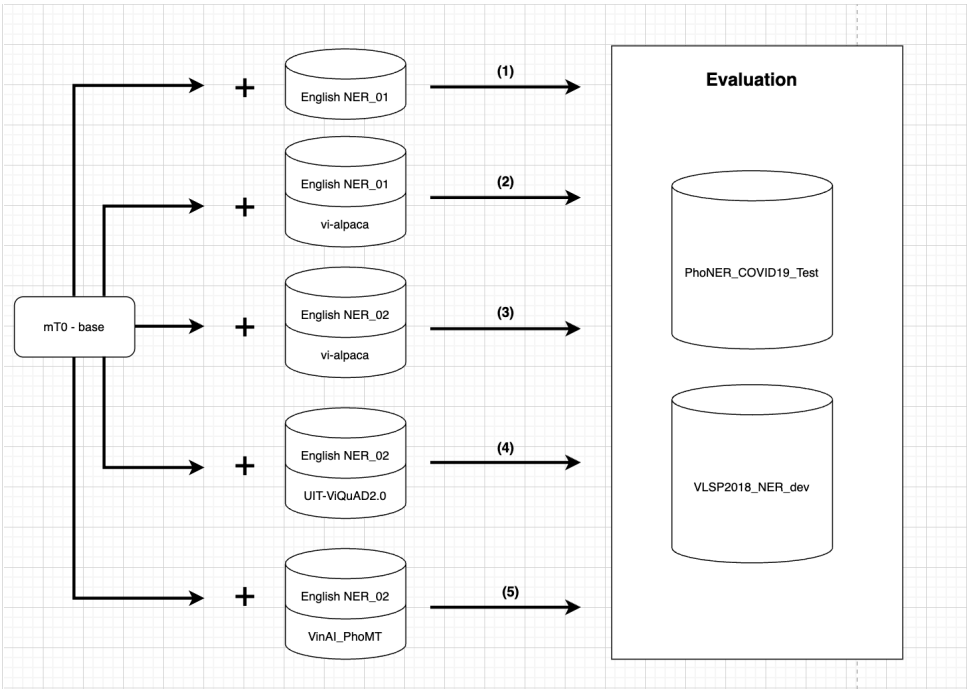


Figure 3.8: Base Dataset Selection Steps

The initial phase of fine-tuning systematically explores different dataset

configurations. This involves combining subsets of a comprehensive English open-domain Named Entity Recognition (NER) dataset with various Vietnamese datasets. Specifically, two distinct subsets of the English NER data were utilized:

- **English NER_01**: Comprising 214,157 records, this subset specifically excludes instances where the output was null, thus representing a cleaner collection of annotated data.
- **English NER_02**: This larger subset contains 358,002 records, which encompasses all records from English NER_01, an additional 21,000 similar records, and importantly, includes records where the output was null. The inclusion of *English NER_02* allows for an investigation into the impact of larger, potentially noisier, or more comprehensive training data on model performance.

Following the fine-tuning procedures, the resulting models underwent rigorous evaluation. This evaluation was performed on two distinct Vietnamese NER benchmark datasets: PhoNER_COVID19_Test and VLSP2018_NER_dev. The choice of these datasets aims to assess the models' generalization capabilities across different domains and annotaticccon styles within the Vietnamese NER context. Through these systematic experiments, we aim to identify the most effective data combination and fine-tuning strategy. Preliminary analysis indicated that Process (4), specifically the combination involving English NER_02 and UIT-ViQuAD2.0, yielded the most promising results. Consequently, this configuration was selected as the optimal strategy for the subsequent stages of model development and evaluation.

Moreover, as we examined the evaluation results across different datasets, we observed a consistent improvement in model performance with the inclusion of additional training data. Notably, datasets from Vietnamese Machine Translation and Question Answering tasks contributed the most significant gains. Based on this observation, we propose an additional fine-tuning strategy that leverages the diversity of Vietnamese-language datasets from various NLP tasks to enhance the performance of the Named Entity Recognition task. See figure 3.9.

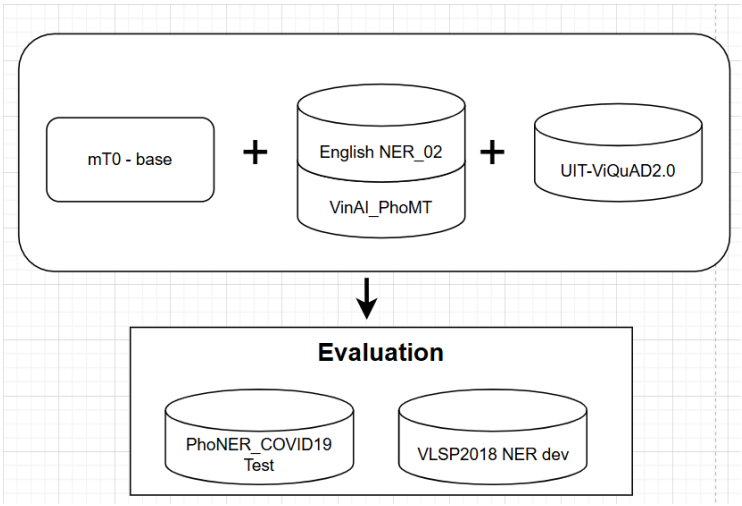


Figure 3.9: Mix MultiTask Step

3.4 Advanced Fine-tuning

Our approach to developing a robust Named Entity Recognition system for Vietnamese data necessitates a sophisticated fine-tuning strategy. This section describes the multi-stage process employed, which is designed to leverage both cross-lingual transfer and domain-specific knowledge. For clarity, the methodology is presented in figure 3.10.

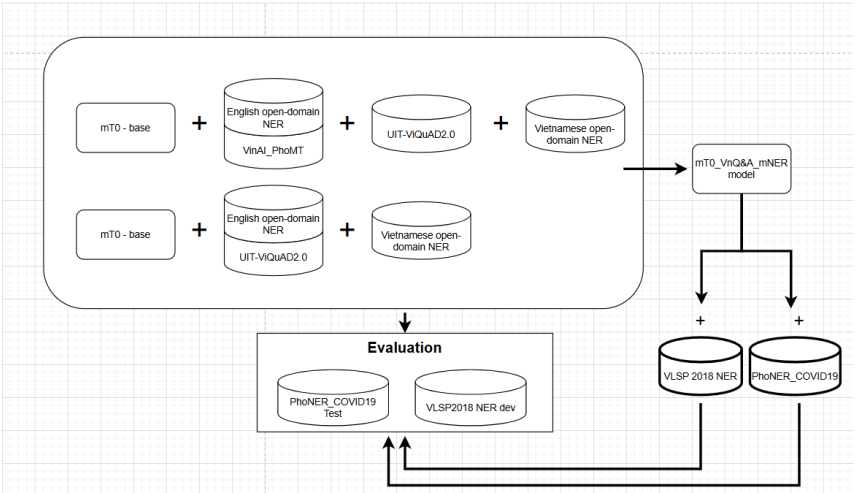


Figure 3.10: Advanced Fine-tuning

To maximize the use of Vietnamese-language resources from related tasks such as Vietnamese Question Answering (VietnameseQA) and Machine Translation, we propose two experimental directions at this stage:

- **Single Dataset Strategy**: Use only the dataset that achieved the highest individual performance, combined with a Vietnamese open-domain NER dataset.

- **Dual Dataset Strategy**: Use the two top-performing datasets, along with the Vietnamese open-domain NER dataset.

The use of the UIT-ViQuAD2.0 dataset [36], a Vietnamese QA corpus containing numerous answers consisting of a single entity, produced remarkable results. This suggests that QA data, despite being from a different task, can effectively support the model's ability to identify entities.

Additionally, the VinAI_PhoMT dataset contributes to enhancing the model's understanding of Vietnamese linguistic characteristics. It also improves the model's capacity to detect relevant entities within a broader and more diverse context.

Finally, the integration of a Vietnamese open-domain NER dataset provides the model with direct exposure to Vietnamese entity recognition tasks. This integrated fine-tuning strategy leverages the strengths of each dataset: general NER principles from English-based datasets, contextual comprehension from Vietnamese QA, and domain-specific entity patterns from Vietnamese NER corpora. The resulting model from this process is a specialized version optimized for multilingual NER with a strong emphasis on Vietnamese entities.

The final phase of our methodology involves a rigorous and comprehensive evaluation of the fine-tuned mT0_VnQ&A_MT_mNER model. The evaluation is structured around two distinct scenarios: a zero-shot setting and a supervised fine-tuning setting. In the zero-shot setting, the model's ability to perform NER on Vietnamese datasets it has not been explicitly trained on is assessed. This evaluates the effectiveness of knowledge transfer achieved during the initial multi-task fine-tuning phase using English and general Vietnamese data. The evaluation is conducted on two benchmark datasets: PhoNER_COVID19_Test and VLSP2018_NER_dev, which directly measure the model's generalization capacity.

In contrast, the supervised fine-tuning setting evaluates the model after it has undergone additional fine-tuning using domain-specific Vietnamese NER datasets, namely VLSP 2018 NER and PhoNER_COVID19. This setup assesses how well the model adapts and performs when explicitly trained for Vietnamese NER tasks. Results from both evaluation strategies offer valuable insights into the model's ability to generalize to new domains and its performance when given task-specific supervision. These findings highlight the effectiveness of the proposed two-stage, multi-task fine-tuning strategy.

CHAPTER 4. NUMERICAL RESULTS

This chapter presents the experimental design, evaluation metrics, and results obtained during the fine-tuning and assessment of the mT0 model on Vietnamese Named Entity Recognition (NER) tasks. The study explores both zero-shot and supervised settings, leveraging multilingual instruction tuning, data mixing strategies, and parameter-efficient fine-tuning via LoRA [1]. Particular emphasis is placed on assessing the impact of different Vietnamese datasets, training scales, and evaluation frameworks on the model's performance. The results aim to provide insight into effective strategies for adapting large-scale multilingual models to low-resource NER scenarios.

4.1 Evaluation Parameters

To evaluate the models' effectiveness, the F1 score was chosen due to its ability to integrate both recall and precision into a single metric. This provides a more holistic assessment, reflecting both the correctness of predictions and the extent to which relevant entities are identified. Unlike traditional NER models that rely on tagging schemes such as BIO to extract and classify entities within a fixed set of categories, the open-domain nature of this task necessitates a more flexible approach.

Specifically, instead of requiring the model to identify all entities in the text and assign labels across multiple categories, each evaluation focuses solely on identifying entities of a single specific type. Rather than producing a sequence of tags, the model outputs a list of entities in string format (similar to a Python list), which simplifies the evaluation process and makes it more suitable for open-domain settings, where entity types are broader and less well-defined.

Importantly, due to the inherent openness of the task, the entities predicted by the model are not required to exactly match the ground truth. This relaxed evaluation criterion acknowledges the ambiguity of natural language and the practical challenges of achieving exact matches, even for human annotators. It also reflects the real-world nature of open-domain NER, where entity boundaries and definitions are often fluid rather than rigidly fixed.

4.2 Simulation Method

4.2.1 Base Vietnamese Instruction Tuning Dataset Selection

During fine-tuning, the training process at certain stages incorporated the full English NER dataset to improve the model's ability to generalize across

named entity types. In steps that combined multiple datasets, data from different sources were shuffled and trained jointly under a unified loss function to maintain consistency in optimization.

In the case of mT0, each input was tokenized and padded so that all sequences within a batch were aligned to the length of the longest sample. Due to constraints in computational resources and training time, a parameter-efficient fine-tuning (PEFT) strategy was adopted instead of updating all model parameters. All experiments were fine-tuned with a batch size of 4, a fixed learning rate of 0.00003, and trained for only one epoch.

Upon completing the fine-tuning process, a total of five models were obtained. These models were evaluated using the test splits of the VLSP2018 NER [30] and PhoNER_COVID19 [15] datasets. The model's performance in accurately identifying specific entity types was assessed by comparing its predicted entity lists against the ground truth, with the F1 score serving as the principal evaluation metric.

4.2.2 Advanced Fine-tuning

During this stage, the data mixing ratio between the English open-domain NER dataset and the Vietnamese question-and-answering datasets is detailed. The English dataset comprised the larger portion of the data, containing 358,002 samples, whereas the Vietnamese dataset provided 19,238 samples. These datasets were randomly mixed before training. The same training configuration as in the previous step was applied, using a batch size of 4 and a constant learning rate of 0.00003 over one epoch.

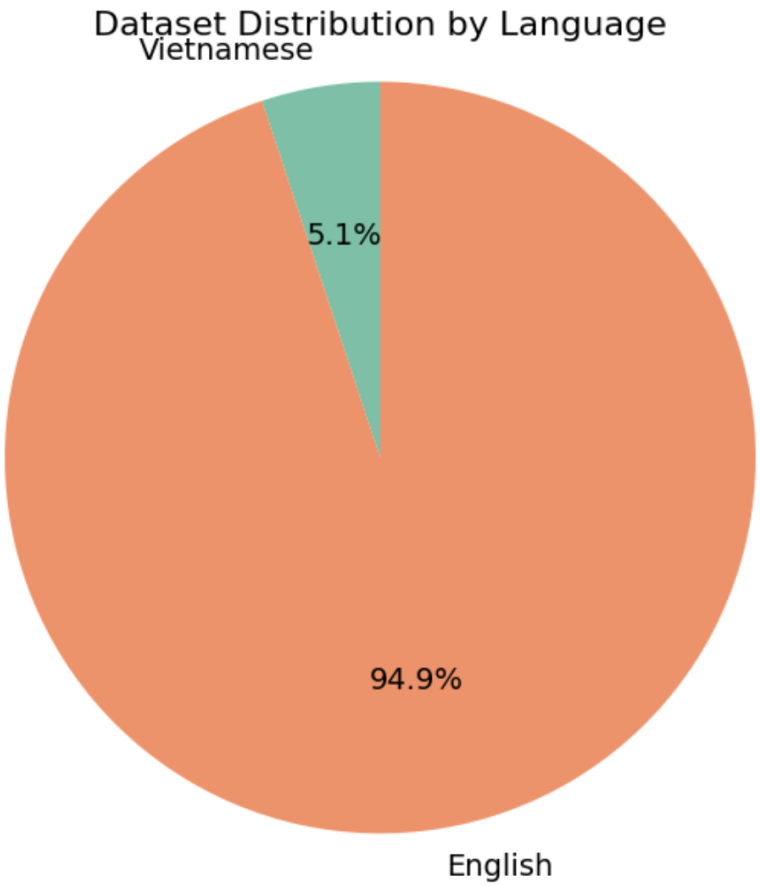


Figure 4.1: Proportion of English NER and Vietnamese QA Datasets

In the subsequent stage, all 45,834 records from the Vietnamese NER dataset were utilized to enhance the model's ability in open-domain entity recognition.

4.3 Base Vietnamese Instruction Tuning Dataset Selection Results

Model	**PhoNER_COVID19_Test**	**VLSP2018_Dev**
mT0-base + EngNER_01	34.10	39.68
mT0-base + EngNER_01 mix vi-alpaca	33.14	43.57
mT0-base + EngNER_02 mix vi-alpaca	33.14	43.57
mT0-base + EngNER_02 mix UIT-ViQuAD2.0	55.12	61.26
mT0-base + EngNER_02 mix VinAI_PhoMT	50.97	54.67
mT0-base + EngNER_02 mix VinAI_PhoMT + UIT-ViQuAD2.0	**76.16**	**79.58**

Table 4.1: NER Performance on Vietnamese Datasets

The most significant finding from this experiment is the superior performance of the model fine-tuned with the *UIT-ViQuAD2.0* dataset. As highlighted in the

table, the configuration *mT0-base + EngNER_02 mix UIT-ViQuAD2.0* achieved the highest F1-scores on both evaluation sets, with 55.12 on *PhoNER_COVID19_Test* and 61.26 on *VLSP2018_Dev*.

Baseline and the Impact of Vietnamese Data: The initial experiment, using only English NER data (*mT0-base + EngNER_01*), establishes a baseline performance with F1-scores of 34.10 and 39.68. The introduction of any Vietnamese dataset provides a notable improvement, particularly on the general-domain *VLSP2018_Dev* set, underscoring the necessity of in-language data for effective fine-tuning.

Comparison of Vietnamese Datasets: The core of the experiment lies in comparing the impact of three different Vietnamese datasets when mixed with the English NER data:

- **vi-alpaca**: This instruction-tuning dataset provides a moderate boost, improving the *VLSP2018_Dev* score to 43.57. However, it slightly degrades performance on the specialized *PhoNER_COVID19_Test* set (33.14 vs. the baseline of 34.10). This suggests that general instruction tuning helps with general language understanding but may not be optimal for a specific, technical task like NER.
- **VinAI_PhoMT**: Mixing with this Vietnamese-English machine translation dataset yields a substantial improvement over *vi-alpaca*, with scores of 50.97 and 54.67. This indicates that the structured, parallel nature of a translation dataset is more beneficial for the structured prediction task of NER than a general instruction dataset.
- **UIT-ViQuAD2.0**: This Vietnamese Question Answering (QA) dataset proved to be the most effective data source by a large margin. It outperformed the *VinAI_PhoMT* configuration by approximately 4 F1-points on *PhoNER_COVID19* and over 6.5 F1-points on *VLSP2018*.

The results conclusively demonstrate that the choice of the mixed-in Vietnamese dataset is a critical factor in model performance. The superior results from the *UIT-ViQuAD2.0* dataset suggest that Question Answering data is an exceptionally potent resource for transfer learning to NER tasks. The inherent nature of QA, which requires identifying entities (persons, locations, dates, etc.) to form answers, provides strong and highly relevant signals for training a named entity recognizer. This task alignment makes it more effective than both general instruction-following data (*vi-alpaca*) and machine translation data (*VinAI_PhoMT*).

During the experiments, the combination of both *VinAI_PhoMT* (machine translation) and *UIT-ViQuAD2.0* (Vietnamese question answering), along with the Vietnamese open-domain NER dataset, resulted in the highest performance, achieving F1 scores of 76.16 and 79.58. These results indicate that leveraging diverse Vietnamese datasets from different NLP tasks, such as QA and machine translation, can provide valuable semantic and contextual knowledge. This, in turn, significantly enhances the model's ability to recognize named entities in Vietnamese. The findings clearly demonstrate the effectiveness of the multi-task fine-tuning strategy, particularly when fully utilizing related datasets rather than relying solely on conventional NER resources.

4.4 Advanced Fine-tuning Results

This section discusses the model's performance in two settings: zero-shot evaluation (without fine-tuning on the evaluation dataset's training data) and supervised fine-tuning (with fine-tuning on the training set of the evaluation dataset).

4.4.1 Zero-shot evaluation

Model	PhoNER_COVID19_Test	VLSP2018_Dev
mT0-base + EngNER_02 mix UIT-ViQuAD2.0 + Vietnamese_NER	74.38	81.79
mT0-base + EngNER_02 mix VinAI_PhoMT + UIT-ViQuAD2.0 + Vietnamese_NER	**74.53**	**85.78**

Table 4.2: Zero-shot Evaluation

The mT0-base model, fine-tuned using a two-stage training strategy, demonstrated promising results. The first stage involved multitask learning across English open-domain NER and Vietnamese question answering (UIT-ViQuAD2.0), followed by a second stage of fine-tuning on Vietnamese NER data. As shown in Table 4.2, this approach achieved an F1 score of 73.8 on the PhoNER_COVID19_Test set and 74.29 on the VLSP2018_Dev set. These scores are noteworthy considering the model was evaluated in a zero-shot setting, without exposure to any training samples from the test sets.

Although the performance on PhoNER_COVID19 lags behind that of more specialized models like BiLSTM-CRF [2], XLM-R base, or XLM-R large [28] as reported in prior work by VinAI, the gap can be attributed to the domain-specific nature of PhoNER_COVID19. This dataset includes a significant number of medical entities that were underrepresented or absent in the model's training

data, making generalization more difficult.

Nevertheless, when compared to a baseline model fine-tuned solely on NER data, this two-stage approach significantly outperformed in both test scenarios. The integration of cross-task learning with Vietnamese QA appears to have enhanced the model's reasoning and span extraction capabilities. These results support the hypothesis that multi-task and staged fine-tuning strategies offer a more effective pathway for improving model generalization in open-domain NER, particularly under low-resource or zero-shot conditions.

Since this is an open-domain problem, prioritizing a dataset with a wide range of entity types—despite a trade-off of 1–2% in performance—is a reasonable and worthwhile choice.

4.4.2 Supervised Fine-tuning Evaluation

Since the results obtained in the zero-shot setting were not particularly impressive, we now proceed to the Supervised Fine-tuning Evaluation phase.

a, Supervised Fine-tuning on PhoNER_COVID19

To evaluate the effectiveness of the mT0 model on the Vietnamese PhoNER_COVID19 dataset, we adopted a parameter - efficient fine-tuning strategy using LoRA (Low-Rank Adaptation). Instead of updating all 583 million parameters of the mT0 model, we only modified 884,736 parameters - less than 0.15% of the full model size. This approach significantly reduces computational cost while preserving the model's ability to adapt to new domains.

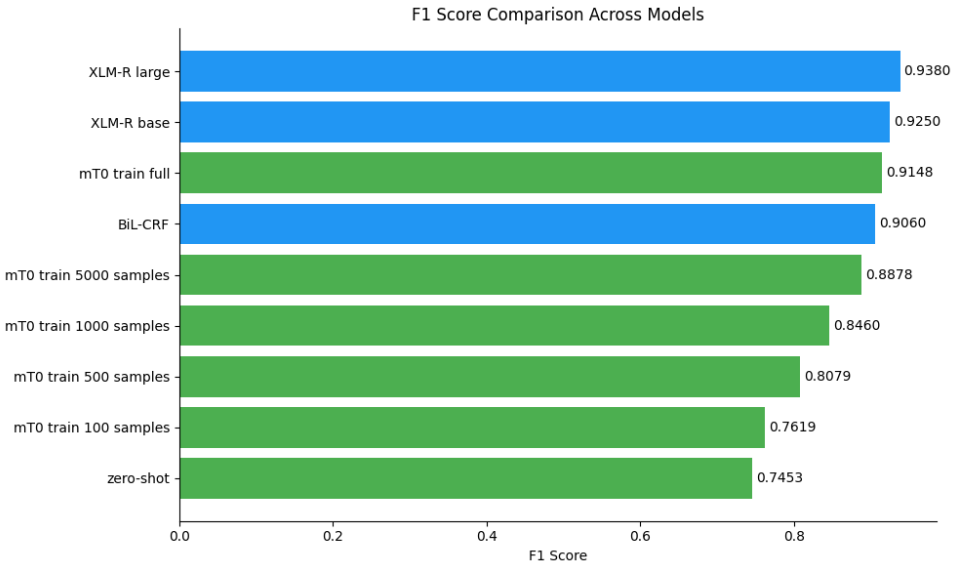


Figure 4.2: Final Model Results on PhoNER_COVID19

We conducted experiments across several data scales: zero-shot, and supervised

fine-tuning with 100, 500, 1000, and all available 4,197 training examples. As illustrated in Figure 4.2, the model already achieved a respectable F1 score of 74.53 in the zero-shot setting. With as few as 100 training samples, performance rose to 76.19, and continued to increase with more data: 80.79 (500 samples), 84.60 (1000 samples), and 91.48 when trained on the full dataset.

Despite using fewer trainable parameters than models like XLM-R base (270M) and XLM-R large (550M), the LoRA-adapted mT0 surpassed them in performance. This result highlights not only the strength of instruction-tuned models like mT0, but also the efficiency and practicality of LoRA in resource-constrained NER settings.

b, Supervised Fine-tuning on VLSP NER 2018

To assess the generalization capability of the mT0 model on the VLSP 2018 NER dataset, we adopted a low-rank fine-tuning strategy using LoRA, updating only 884,736 trainable parameters. Experiments were conducted with varying training data sizes: 40, 400, 2000, and the full training set (containing 4197 sentences).

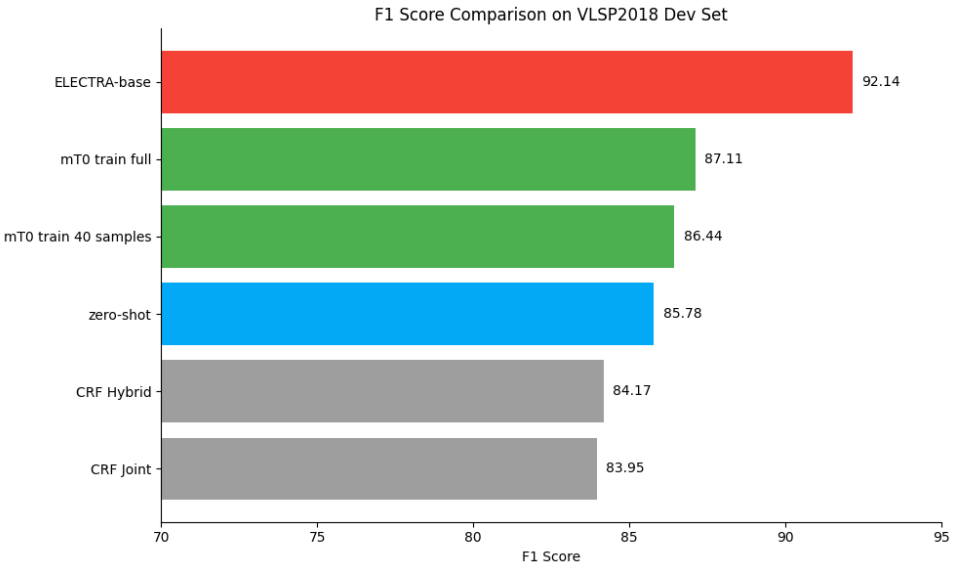


Figure 4.3: Final Model Results on VLSP2018 NER

Interestingly, even with as few as 40 examples, the model achieved an F1 score of 86.44%, surpassing the zero-shot baseline (85.78%). The best result, 87.11, was obtained when training on the full set. These findings highlight that the mT0 model, even when fine-tuned with a small number of parameters via LoRA [1], is capable of achieving competitive results on Vietnamese nested NER tasks.

When compared to existing models reported for VLSP 2018, our best-

performing configuration is only 10 points below the state-of-the-art ELECTRA-base model [38] (92.14% F1 on the development set, as reported by NlpHUST). Moreover, it significantly outperforms earlier CRF-based [39] systems such as the hybrid and joint models (with F1 scores around 84.17% and 83.95%, respectively), despite using a parameter-efficient approach. Given that our fine-tuned model updated fewer than 1 million parameters, these results demonstrate the effectiveness of combining large-scale multilingual pretrained models with parameter-efficient fine-tuning techniques for low-resource NER tasks.

Summary

In summary, the experiments demonstrate that the mT0 model, when fine-tuned using a combination of multilingual data and parameter-efficient strategies, can achieve competitive performance on Vietnamese NER tasks. The use of LoRA allowed for substantial parameter savings while maintaining accuracy, and the integration of Vietnamese QA datasets proved especially beneficial in enhancing generalization. Despite resource constraints, the model outperformed several traditional baselines and approached the performance of state-of-the-art systems. These findings validate the effectiveness of instruction-tuned multilingual models for low-resource NER and suggest promising directions for future research in cross-lingual and cross-task adaptation.

CHAPTER 5. CONCLUSIONS

5.1 Summary

This thesis presented an in-depth exploration of developing an open-domain Named Entity Recognition (NER) model tailored for Vietnamese, aiming to serve as a representative case study for low-resource languages. The core objective centered on leveraging multilingual encoder-decoder architectures - specifically the mT0 model - while investigating a variety of fine-tuning strategies and identifying appropriate training resources to enhance the model's ability to recognize diverse types of named entities.

The first stage of the research involved a systematic comparison among several Vietnamese datasets to determine the most effective training source. Among them, UIT-ViQuAD2.0 proved to be the most promising for entity recognition purposes, despite its original design for question-answering tasks. Furthermore, experiments demonstrated that using a compact model such as mT0-base in combination with parameter-efficient tuning techniques (e.g., PEFT) could still achieve strong results, emphasizing the practicality of lightweight solutions in low-resource settings.

Subsequent efforts focused on more advanced fine-tuning strategies, including multitask learning and a two-stage fine-tuning pipeline that combined open-domain English NER data with Vietnamese question-answering data. Model evaluations were conducted in both zero-shot and supervised scenarios. Results in the zero-shot setting highlighted the benefit of the proposed strategies in enabling the model to recognize Vietnamese entities without having seen similar data during training. In the supervised setting, where the model was exposed to a small amount of domain-specific annotated data, further performance improvements were observed - surpassing the baseline systems reported in the original PhoNER_COVID19 study. This reinforces the model's ability to effectively adapt even when minimal in-domain resources are available. However, it is worth noting that despite these improvements, the results are still not optimal, and several challenges remain unresolved.

5.2 Suggestion for Future Works

While this thesis lays important groundwork for developing open-domain Named Entity Recognition (NER) models for Vietnamese, several promising directions remain open for future exploration:

- **Incorporating Instruction-Following Pretraining at Scale**: Large instruction-tuned models have shown impressive generalization across tasks in high-resource languages. Applying a similar pretraining phase on large-scale Vietnamese or multilingual instruction datasets could further improve model adaptability to open-ended NER prompts, particularly in real-world applications where instruction-following is crucial.
- **Expansion to Multimodal NER Settings**: Future work could explore integrating textual and visual data (e.g., news headlines with accompanying images or social media posts with embedded media). This multimodal setting could enrich the context and enhance entity disambiguation, especially in domains like media monitoring, public health, or disaster response.
- **Domain-Aware NER Adaptation**: Instead of a single open-domain model, a promising direction would be to build a modular system that dynamically adapts to specific domains (e.g., medical, financial, or legal text) using domain classifiers or lightweight adapters. This would enable better precision when dealing with highly specialized or jargon-rich content.
- **Incorporating Human Feedback for Entity Correction**: Future research could incorporate human-in-the-loop training strategies where incorrect or ambiguous entity predictions are flagged and corrected by annotators. Such feedback could be used to fine-tune the model incrementally, resulting in more robust and accurate NER performance in deployment.
- **Low-Resource Cross-Lingual Transfer Techniques**: Investigating advanced cross-lingual transfer methods such as contrastive learning, prompt-based transfer, or alignment-based fine-tuning could allow the model to generalize better across low-resource languages without requiring large annotated corpora for each target language.
- **Benchmarking against More Realistic Datasets**: Finally, future work should consider evaluating models on user-generated content such as forum discussions, chat logs, or social media posts. These noisy, unstructured texts better reflect the diversity and unpredictability of entities encountered in truly open-domain settings.

REFERENCE

[1] E. Hu **andothers**, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[2] Z. Huang, W. Xu **and** K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," **in***Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NeurIPS)* 2015, **pages** 1485–1493. **url**: `https://papers.nips.cc/paper_files/paper/2015/file/5945b6911194f4f9bPaper.pdf`.

[3] J. Devlin, M.-W. Chang, K. Lee **and** K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL*, 2019.

[4] Y. Liu **andothers**, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] A. Conneau **andothers**, "Unsupervised Cross-lingual Representation Learning at Scale," *ACL*, 2020.

[6] T. B. Brown **andothers**, "Language models are few-shot learners," *NeurIPS*, 2020.

[7] V. Sanh, A. Webson, C. Raffel **andothers**, "Multitask Prompted Training Enables Zero-Shot Task Generalization," *arXiv preprint arXiv:2110.08207*, 2021. **url**: `https://arxiv.org/abs/2110.08207`.

[8] H. W. Chung **andothers**, "Scaling Instruction-Finetuned Language Models," *arXiv preprint arXiv:2210.11416*, 2022.

[9] R. Taori **andothers**, *Stanford Alpaca: An Instruction-following LLaMA Model*, `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[10] W.-L. Chiang **andothers**, *Vicuna: An Open-Source Chatbot Impressing GPT-4*, `https://lmsys.org/blog/2023-03-30-vicuna/`, 2023.

[11] X. Li **andothers**, "Distilling Step-by-Step: Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes," *arXiv preprint arXiv:2305.02301*, 2023.

[12] Z. Zhang **andothers**, "UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition," *arXiv preprint arXiv:2304.10444*, 2023.

[13] H. T. Nguyen, Q. T. Ngo, L. X. Vu, V. M. Tran **and** H. T. Nguyen, "VLSP shared task: Named entity recognition," *Journal of Computer Science and Cybernetics*, **jourvol** 34, **number** 4, **pages** 283–294, 2018.

[14] H. M. Linh, N. T. M. Huyen, D. X. Dung **andothers**, "VLSP 2021-NER Challenge: Named Entity Recognition for Vietnamese," *VNU Journal of Science: Computer Science and Communication Engineering*, **journvol** 38, **number** 1, 2022.

[15] D. Q. Nguyen **and** T. V. Nguyen, "PhoNER_COVID19: Vietnamese Named Entity Recognition for COVID-19 Related Text," **in***Proceedings of the 8th Workshop on Asian Language Resources (ALR)* 2021. **url**: https://github.com/VinAIResearch/PhoNER_COVID19.

[16] X.-S. Vu **andothers**, "VnCoreNLP: A Vietnamese Natural Language Processing Toolkit," **in***LREC* 2018.

[17] NLP@HUST, *NlpHUST/ner-vietnamese-electra-base*, https://huggingface.co/NlpHUST/ner-vietnamese-electra-base, Accessed: 2025-05-19, 2021.

[18] V. Anh, *Underthesea: Vietnamese NLP Toolkit*, https://github.com/undertheseanlp/underthesea, Accessed: 2025-05-19, 2018.

[19] T. T. Viet, *PyVi: Vietnamese NLP Toolkit*, https://github.com/trungtv/pyvi, Accessed: 2025-05-22, 2016.

[20] N. Muennighoff, T. Wang, J. Eisenschlos **andothers**, "Crosslingual Generalization through Multitask Finetuning," *arXiv preprint arXiv:2211.01786*, 2022. **url**: https://arxiv.org/abs/2211.01786.

[21] J. Devlin, M.-W. Chang, K. Lee **and** K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018. **url**: https://arxiv.org/abs/1810.04805.

[22] X. Li, Y. Shao, J. Zhang **and** F. Huang, "UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition," **in***Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)* 2023. **url**: https://arxiv.org/abs/2305.14263.

[23] H. Touvron, T. Lavril, G. Izacard **andothers**, *LLaMA: Open and Efficient Foundation Language Models*, 2023. arXiv: 2302.13971 [cs.CL]. **url**: https://arxiv.org/abs/2302.13971.

[24] A. Vaswani, N. Shazeer, N. Parmar **andothers**, "Attention is All You Need," **in***Advances in Neural Information Processing Systems (NeurIPS)* 2017, **pages** 5998–6008. **url**: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[25] J. L. Elman, "Finding structure in time," *Cognitive Science*, **jourvol** 14, **number** 2, **pages** 179–211, 1990. DOI: 10.1207/s15516709cog1402_1.

[26] S. Hochreiter **and** J. Schmidhuber, "Long short-term memory," *Neural Computation*, **jourvol** 9, **number** 8, **pages** 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[27] C. Raffel **andothers**, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *JMLR*, 2020.

[28] A. Conneau, K. Khandelwal, N. Goyal **andothers**, "Unsupervised Cross-lingual Representation Learning at Scale," **in***Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)* Association for Computational Linguistics, 2020, **pages** 8440–8451. **url**: https://aclanthology.org/2020.acl-main.747.

[29] Y. Liu **andothers**, *PEFT: Parameter-Efficient Fine-Tuning*, https://github.com/huggingface/peft, Accessed: 2025-06-12, 2023.

[30] P. Le-Hong, Q. H. Pham **andothers**, "VLSP 2018 Evaluation Campaign: Named Entity Recognition for Vietnamese," **in***Proceedings of the 6th International Workshop on Vietnamese Language and Speech Processing (VLSP)* 2018. **url**: https://vlsp.org.vn/2018/eval/ner.

[31] R. Taori, I. Gulati, T. Zhang **andothers**, *Stanford Alpaca: An Instruction-following LLaMA model*, https://github.com/tatsu-lab/stanford_alpaca, Accessed: 2025-06-12, 2023.

[32] M. AI, *LLaMA 3: Open Foundation and Instruction-Tuned Language Models*, https://ai.meta.com/llama/, Accessed: 2025-06-12, 2024.

[33] B. F. M. Lab, *BKAI News Corpus: A Large-Scale Vietnamese News Dataset*, https://huggingface.co/datasets/BKAI/BKAI-News-Corpus, Accessed: 2025-06-12, 2024.

[34] B. V. Q., *Binhvq News Corpus (Internal Compilation)*, Unpublished dataset, merged into BKAI News Corpus, 2023.

[35] B. F. M. Lab, *BKAI vi-alpaca Dataset*, https://huggingface.co/datasets/BKAI/vi_alpaca, Accessed: 2025-06-12, 2023.

[36] T. Nguyen, A. Le, K.-H. Tran **and** M.-T. Nguyen, "UIT-ViQuAD 2.0: An Enhanced Vietnamese Reading Comprehension Dataset," **in***Proceedings of the 29th International Conference on Computational Linguistics (COLING)* Accessed: 2025-06-12, 2022. **url**: https://huggingface.co/datasets/taidng/UIT-ViQuAD2.0.

[37] M.-T. Pham, T. H. Bui **and** T. Vu, *PhoMT: A High-Quality Vietnamese-English Machine Translation Dataset*, `https://huggingface.co/datasets/wanhin/VinAI_PhoMT`, Accessed: 2025-06-12, 2021.

[38] N. Team, *VLSP 2018 NER Shared Task: ELECTRA-based Model by NlpHUST*, `https://vlsp.org.vn/resources/2018`, Technical report, VLSP 2018 Shared Task, 2018.

[39] X.-S. Vu, H. P. Le **and** S.-B. Pham, "Vietnamese Named Entity Recognition at VLSP 2018: A Hybrid Model," **in***Proceedings of the 6th Vietnamese Language and Speech Processing (VLSP) Workshop* VLSP Shared Task Report, Hanoi, Vietnam, 2018.