NokkaewGPT

LARGE LANGUAGE MODEL FOR THAI TEXT GENERATION

BY

Mr. Korndanai Siribunkosai

Student ID. 62011142

Miss Nichapatr Uthaisangsuk

Student ID. 62011170

Miss Pitchapa Arun

Student ID. 62011214

Mr. Tanapol Dangsakul

Student ID. 62011301

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF ENGINEERING IN COMPUTER INNOVATION ENGINEERING KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2022

SCHOOL OF ENGINEERING
KING MONGKUT'S INSTITUTE OF TECHNOLOGY
LADKRABANG
PROJECT CERTIFICATE

Project Title	NokkaewGPT
Student Name	Mr. Korndanai Siribunkosai Student ID. 62011142 Miss Nichapatr Uthaisangsuk Student ID. 62011170 Miss Pitchapa Arun Student ID. 62011214 Mr. Tanapol Dangsakul Student ID. 62011301
Degree	Bachelor of Engineering in Computer Innovation Engineering
Project Advisor	Signed:_____ (Asst. Prof. Dr. Akkarit Sangpetch)

Project Title	NokkaewGPT
Student Name	Mr. Korndanai Siribunkosai Student ID. 62011142 Miss Nichapatr Uthaisangsuk Student ID. 62011170 Miss Pitchapa Arun Student ID. 62011214 Mr. Tanapol Dangsakul Student ID. 62011301
Degree	Bachelor of Engineering in Computer Innovation Engineering
Project Advisor	Asst. Prof. Dr. Akkarit Sangpetch
Academic Years	2022/2023

ABSTRACT

Large language models have revolutionized the field of natural language processing (NLP) by enabling machines to understand, generate, and manipulate human language at an unprecedented level. These models, such as the popular GPT (Generative Pre-trained Transformer) architecture, have transformed natural language processing by enabling machines to comprehend, generate, and manipulate human language effectively. This project focuses on developing a language generation model for the Thai language using the GPT (Generative Pre-trained Transformer) baseline. By leveraging the power of the transformer NLP model, we aim to create language generation capabilities specifically for Thai text. The project involves dataset preparation, defining loss functions, optimizing hyperparameters, determining training epochs and batch size, and integrating the model with UX/UI.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude and appreciation to the Computer Innovation Engineering committees at King Mongkut's Institute of Technology Ladkrabang for their invaluable support throughout this capstone project. Their guidance and feedback have been a key in shaping the direction of this project.

We are also incredibly thankful to Asst. Prof. Dr. Akkarit Sangpetch, our advisor, for his support, expertise, and mentorship. His insights, encouragement, and dedication have been vital in completing the project and driving it toward excellence.

We would also like to thank the other unmentioned supporters, including the faculty members, staff, and fellow students, who have contributed their time, knowledge, and encouragement to us along the project journey.

The support of these contributors has made this capstone a reality. We are truly grateful for the opportunity to work on this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS/ABBREVIATIONS

Symbols / Abbreviations	Terms
CIE	Computer Innovation Engineering
SIIE	School of International Interdisciplinary Engineering Programs
UX	User Experience
UI	User Interface
HTTP	HyperText Transfer Protocol
JSON	JavaScript Object Notation
CSV	comma-separated values

INTRODUCTION

Background

Large language models have a significant impact on the world and hold promising potential for the future. Currently, large language models have gained significant attention and popularity due to their impressive capabilities in generating coherent and contextually relevant text. They have demonstrated the ability to generate realistic paragraphs, answer questions, and even engage in conversation that mimics human-like responses.

Large language models have had a significant impact across various domains and applications. Some notable impacts of large language models are Natural Language Processing (NLP), content generation, language translation, dialogue systems, Automation, and AI technology. These led to revolutionized natural language understanding, communication, and automation. They have facilitated language accessibility, driven research, innovation, and democratized AI. These facilitated outcomes are then utilized in business and society, daily life.

The Large language models had merged into our daily life. For example, Virtual Assistants and Chatbots like Siri, Alexa, and Google Assistant, making them more conversational and capable of understanding complex queries. Chatbots deployed in customer service and support roles use these models to provide prompt and accurate responses to user inquiries. There are also Language Translations that improve machine translation services, enabling real-time translation of text and speech between languages. This has facilitated multilingual communication, breaking down language barriers and promoting global connectivity. The model also aids in content creation, helping writers generate articles, blog posts, and social media updates. They can also assist in content curation by summarizing lengthy documents or providing relevant information based on user preferences. The model also enhances search engines' ability to understand queries and retrieve relevant results or so called Information Retrieval. This improves the accuracy and efficiency of information retrieval, making it easier to find the desired information from vast amounts of data.

Motivation

Recently, chatGPT, developed by OpenAI, has been gaining attention recently due to its powerful impact on education, social, productivity, human-computer interactions and various jobs. This chatbot has the potential to create human-like language content such as emails, poems, songs, and short stories, etc. It also does some related language tasks like giving topics for presentations, correcting grammar, or answering questions.

While significant advancements have been made in natural language processing for the English language, there is still much to explore when it comes to Thai language processing.

By creating a robust Thai language model, we can unlock a wealth of possibilities for various natural language processing applications tailored specifically to the Thai language. These applications can include machine translation, text summarization, sentiment analysis, question answering systems, and more. Such advancements would greatly benefit Thai speakers, enabling the development of Thai-language applications, online content, and digital tools that cater to their unique needs.

The large language model in Thai can support language learning and education initiatives. It can provide language learners with access to comprehensive language resources, interactive tools for language practice, and intelligent tutoring systems. It aids in improving Thai language proficiency, promoting literacy, and facilitating the learning process for both native speakers and non-native learners, which helps bridge the language gap between Thai and other languages. It enables collaboration, knowledge sharing, and cultural exchange between Thai speakers and other language communities.

In the future, large language models will likely continue to evolve, becoming more sophisticated, accurate, and capable of understanding and generating human-like language. They will play a vital role in shaping daily life by transforming communication, information retrieval, Content Filtering and Fact-Checking, personalizing services, education, and various other aspects of our lives. If those abilities are available in the Thai language as well, it will enhance technology and

intelligent resources in Thailand and support Thai businesses and industries more efficiently.

Project Introduction

Nokkaew Generative Pre-trained Transformer (NokkaewGPT) is a language generation model that is built upon OpenAI's GPT-2 baseline. It inherits the architecture and principles of GPT-2, which is a state-of-the-art language model known for its impressive performance in various natural language processing tasks.

A large language model refers to a powerful and highly parameterized machine learning model that has been trained on a vast amount of text data to understand and generate human-like language. These models are typically based on deep learning architectures, such as the Transformer model, and are designed to capture the complex patterns, relationships, and structures present in natural language.

Unlike the original GPT-2, NokkaewGPT stands out as a language generation model primarily trained on Thai texts. By training on a large corpus of Thai text data, NokkaewGPT would acquire a deep understanding of the language's unique characteristics, grammar, and semantic structures. This targeted training approach enhances the model's ability to generate high-quality summaries specifically tailored to the Thai language, making it a valuable tool for Thai language processing and natural language understanding tasks.

Report Coverage

The report begins with the literature review section. This section provides a comprehensive overview of the theoretical frameworks and related tools that are relevant to the NokkaewGPT project. It serves three primary goals: identification of theoretical frameworks, exploration of tool stacks and algorithms, and drawing conclusions based on the review.

Following the literature review, the system architecture section provides an in-depth analysis of the system architecture of NokkaewGPT. This includes a detailed

explanation of the underlying components that make up the model, such as the transformer architecture, decoder block, self-attention mechanism, and feed-forward neural network. Each component is described in terms of its purpose, functionality, and how it contributes to the overall language generation process.

The implementation methodology section delves into the practical aspects of developing NokkaewGPT. It covers various stages of the implementation process, including data preprocessing, training, and fine-tuning. Details on how the model is adapted and trained specifically for the Thai language are discussed, along with any modifications or enhancements made to improve its performance and accuracy.

The project evaluation section of the report focuses on measuring the performance and assessing the effectiveness of the NokkaewGPT model. It encompasses various evaluation metrics and techniques used to evaluate the model's performance on different datasets, including the training set, validation set, and test set.

In the project discussion section, a critical analysis of the findings and outcomes is presented. This includes an examination of the limitations and challenges encountered during the development and evaluation of NokkaewGPT. Potential avenues for further research and improvement are explored, and any ethical considerations or implications of the model's application are discussed. This section aims to foster a deeper understanding of the project's significance and its implications for the field of Thai language processing and natural language understanding.

Objectives

The main objectives of this project can be categorized into four key areas: data collection, tokenization and encoding, model training, and language generation. These four areas represent the core components that need to be addressed in order to develop a robust and effective Thai language model.

Firstly, a significant focus is placed on collecting a sufficient amount of Thai text corpus to train a neural network. The corpus acts as the foundation for the language model and plays a crucial role in its effectiveness. Representative Thai text sources,

which are mostly news and informative articles, are gathered to ensure a comprehensive training dataset.

Secondly, an important objective is to successfully tokenize Thai languages and encode them in a manner compatible with machine learning methodologies. This involves developing effective techniques and algorithms to convert the raw Thai text into tokenized sequences that capture the linguistic nuances and structure of the language. Accurate tokenization is essential for facilitating effective language modeling and ensuring the model's ability to understand and generate meaningful Thai language.

The third objective focuses on utilizing the encoded tokens to train and save a language model checkpoint with the GPT-2 architecture. The training process involves optimizing the model's parameters, adjusting hyperparameters, and leveraging large-scale computational resources to achieve optimal performance.

Lastly, an important objective is to utilize the saved language model checkpoint to generate content. The model's ability to generate text that aligns with the original corpus while demonstrating creativity and coherence is a key measure of its success. The generation process involves leveraging the learned patterns and knowledge within the model to produce high-quality Thai language output that adheres to the semantic and syntactic characteristics of the training data.

By achieving these objectives, the project aims to contribute to the advancement of Thai language processing and natural language understanding, enabling the generation of coherent and contextually relevant content in the Thai language domain.

Additionally, as an optional component, we have developed a webpage that allows users to interact with our data model. The webpage provides a convenient platform where users can input text prompts or queries in the Thai language and receive generated responses from the model

Project scope

The scope of the NokkaewGPT project involves training a language generation model specifically for the Thai language by collecting a corpus of Thai text data. This corpus will be used to train the model, which will acquire a deep understanding of Thai language characteristics, grammar, and semantic structures. The project includes tasks such as corpus collection, potential use of web scraping or APIs, supervised training, utilization of the GPT-2 architecture, and customization through fine-tuning for specific tasks.

Table of operations

The table of operations describes the timeline and process of this project for the Computer Innovation Engineering Capstone Design course.

	Jan				Feb				Mar				Apr				May			
Research																				
Preprocess																				
Tokenize																				
Create model																				
Corpus collecting																				
Model Optimization																				
Model Training																				
Parameter tuning																				
Debug																				
UX/UI design																				
Frontend development																				
Frontend Backend Integration																				

Table 1 Table of Operation

LITERATURE REVIEW

This section reviews related model theories and related tools used in the project are all given a brief overview in this chapter. There were three key goals for the investigation. First and foremost, it sought to identify the many theoretical frameworks under consideration in the project. Second, it intended to establish the tool stacks and algorithms used for creating and implementing various system components. Lastly is the review's conclusion.

1. Model Background

This project focuses on text generation, which is an instance of machine learning. As a result, the goal of this part is to provide an overview of fundamental topics such as massive language models, natural language processing (NLP), and transformers.

Large Language Model

A large language model is a type of artificial intelligence model that focuses on understanding and creating human language. To learn the patterns, structures, and meanings of language, these models are trained on huge volumes of text data. They frequently rely on deep learning architectures that enable them to comprehend and generate text with great accuracy and fluency. Large language models are characterized by their extensive size, encompassing a large number of parameters and training data. Because of their small size, they can capture and encode a large range of linguistic knowledge, enabling them to perform a variety of language-related activities including text synthesis, translation, sentiment analysis, and question answering.

Natural Language Processing (NLP)

NLP encompasses a wide range of techniques and algorithms that enable computers to understand, interpret, and generate human language. It is crucial in enabling machines to analyze and comprehend textual material. Text categorization, information extraction, sentiment analysis, and machine translation are all examples of NLP tasks. By leveraging NLP techniques, computers can analyze and extract valuable insights

from large volumes of text, enabling the development of language-based applications and systems.

Transformer NLP

Transformer NLP is a game-changing method that was introduced in 2017 and has completely transformed the field of Natural Language Processing (NLP). It has emerged as a cutting-edge approach for a variety of language processing problems.

has emerged as a state-of-the-art method for various language processing tasks. Unlike traditional recurrent neural networks (RNNs) that process data sequentially, transformers employ attention mechanisms to capture word relationships in a text. This attention mechanism allows the model to focus on relevant parts of the input text, leading to enhanced accuracy and coherence in tasks such as text generation.

2. Tool Stacks and Related Algorithms

This section focuses on the project's tool stacks and frameworks. The purpose is to evaluate their importance and impact, especially in the context of data analytics and system implementation. This section addresses the tools and frameworks used for front-end and back-end development, highlighting their importance in achieving project goals.

a. **Front-End Development Tools and Framework**

Front-end development tools and frameworks are software platforms that help developers create user interfaces (UI) and improve the user experience (UX) of a website or application. These tools help developers to create responsive, engaging, and visually appealing front-end experiences.

HTML/CSS

HTML/CSS as the foundation for designing a user interface which is both intuitive and visually appealing. CSS is used for styling and layout, while HTML provides structural elements. This combination provides an engaging and responsive user experience.

Flask

For back-end development, Flask, a lightweight web framework built in Python, is used. It enables seamless communication between the user interface and the server, successfully handling requests and responses. Flask is an excellent alternative for designing dynamic web apps due to its flexibility and ease of usage.

b. **Back-End Development Tools and Framework**

Back-end development tools and frameworks are software resources or platforms that assist developers in creating and managing the server-side components of websites or applications. These tools offer the essential functionality and infrastructure required for handling data processing and system operations.

Python

Python, renowned for its versatility and robustness, assumes a pivotal role in the implementation of diverse back-end functionalities. With its extensive library ecosystem, Python empowers efficient data processing, manipulation, and seamless integration with a wide range of tools and frameworks. Python furnishes a robust foundation for constructing the core functionalities of the system, ensuring its stability and reliability throughout its operation.

Pytorch

PyTorch, a popular deep learning framework, is utilized for advanced machine learning tasks and data analytics. It allows the implementation of complex neural networks and facilitates training processes. PyTorch's flexibility and efficiency make it suitable for handling large-scale data and enabling sophisticated data analysis and modeling.

Pythainlp

PyThaiNLP, a specialized Python library for Thai natural language processing, enhances the system's language processing capabilities specifically for the Thai language. It offers a comprehensive set of tools for tokenization, part-of-speech tagging, named entity recognition, and more. By leveraging PyThaiNLP, the

system can effectively handle Thai language-specific tasks, improving overall language processing accuracy and efficiency.

c. **Data Collecting Tools**

Data collecting tools refer to software or platforms that enable the collection, extraction, and organization of data from a wide range of sources. These tools are equipped with capabilities such as web scraping and data extraction, which are particularly useful for the current project. This ensures an efficient and effective data collecting process from diverse sources.

scrapy

Scrapy is an open-source framework written in Python that facilitates web crawling and scraping. It was created in Cambuslang and is freely available for use. While its primary purpose is web scraping, Scrapy also supports website crawling and data extraction through APIs. With its versatile functionality, it can be applied to various tasks such as data mining, monitoring, and automated testing. Zyte, a company specializing in web scraping development and services, currently maintains Scrapy.

warcio

This library offers a standalone and efficient solution for handling the WARC Format commonly utilized in web archives. It is compatible with both Python 2.7+ and Python 3.4+ . One of the main functionalities of the library is its capability to iterate over a stream of WARC records, enabling the reading, writing, and extraction of data from WARC files. Additionally, the library includes support for reading ARC files but not writing.

d. **Remote Server**

A remote server or remote host is a centralized computer system located at a different physical location from the user. It offers services, resources, and functionalities over a network, such as the internet. Users can access and manage the server remotely, performing tasks and accessing data without being physically present. Remote servers are widely used for hosting websites, running applications, storing data, and delivering computing services globally.

CMKL Apex Infrastructure

Apex is a high-performance computing platform and storage infrastructure specifically designed to deliver exceptional performance and scalability for AI work. It provides a huge memory GPU and high level infrastructure, enabling faster training and inference in deep neural networks. Apex collaborates with technology partners, including NVIDIA, Oddn, and TCC technology.

Slurm

Slurm is a fault-tolerant and highly scalable open-source cluster management and job scheduling system primarily designed for Linux clusters, commonly utilized in high-performance computing (HPC) environments. It operates without the need for kernel modifications and is self-contained.

As a cluster workload manager, Slurm fulfills three main functions. Firstly, it assigns users exclusive or non-exclusive access to resources (compute nodes) for a specified period, enabling them to carry out their tasks. Secondly, it offers a framework for initiating, executing, and monitoring work, typically parallel jobs, on the allocated nodes. Lastly, it resolves resource conflicts by managing a queue of pending work, ensuring fair allocation among users. Rather than directly accessing individual systems, users interact with Slurm by executing commands (such as srun, sinfo, scancel, scontrol, etc.) from the login node. These commands establish communication with the slurm daemons on each host, enabling the execution of tasks and management of workload across the cluster.

SYSTEM ARCHITECTURE

System architecture refers to the conceptual structure and design of the model, the webpage, and the overall project system. It outlines the key components, their interactions, and the flow of data or information within the system for understanding how the system functions and how its different parts collaborate to achieve the desired functionality and performance of the system.

Project Architecture

The overall architecture of the system is a fundamental aspect of the Thai text generation large language model project. It provides an overview of the system's components, their interactions, and the overall structure of the system.

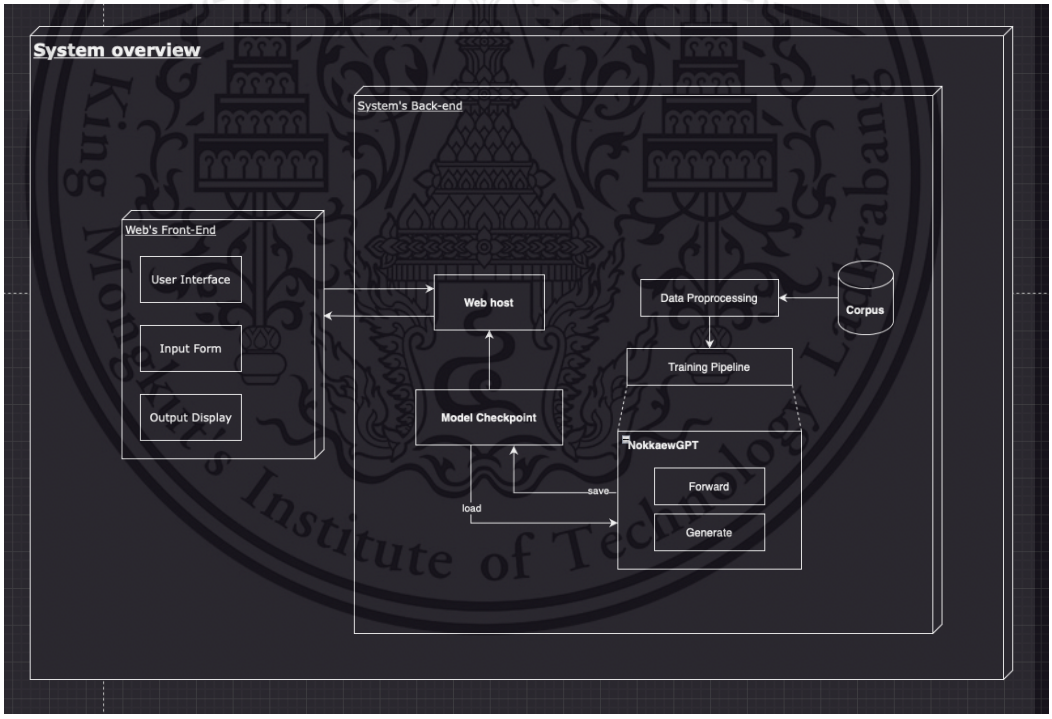

Figure 2.1 Project Architecture Diagram

Transformer Model Architecture

The transformer model architecture focuses on the specific design and structure of the Thai text generation large language model. It outlines the architecture of the underlying

Transformer model with GPT baseline, which forms the basis of the text generation capabilities of the system.

There are two method components in the NokkaewGPT class, which are Forward and Generate. The main building block of the NokkaewGPT model is the transformer model made out of decoder-only blocks.

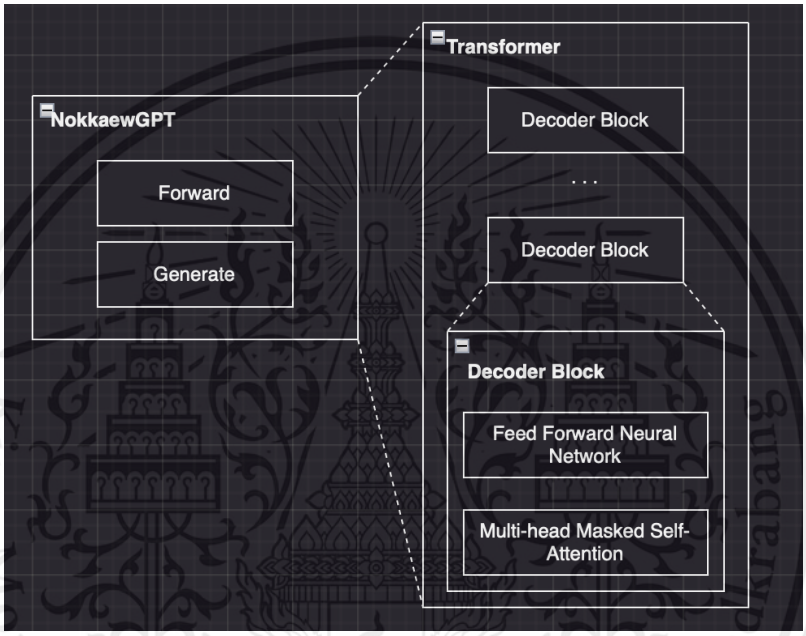

Figure 2.2 Model Architecture Diagram

NokkaewGPT Methods

The forward method takes as input a tensor of integer indices representing the input sequence and an optional tensor of target indices. Within this method, the input sequence is first embedded using token embeddings and position embeddings. The token embeddings capture the semantic meaning of each token, while the position embeddings encode positional information. These embeddings are then combined, resulting in an embedded input tensor.

The embedded tensor is then passed through a series of transformer blocks to capture complex language patterns and dependencies. The output of the transformer blocks is

passed through a layer normalization operation, followed by a linear layer that produces logits representing the predicted probabilities for each token in the vocabulary.

If target indices are provided, the method calculates the loss by comparing the predicted logits with the target indices using the cross-entropy loss function. The loss is returned along with the logits.

The generate method is used to generate new tokens given an initial context and a maximum number of tokens to generate. It iteratively generates tokens by sampling from the predicted probability distribution for the next token. The method crops the input context to the last block_size tokens, obtains the predictions using the forward method, focuses on the last time step, applies softmax to obtain probabilities, and samples from the distribution. The sampled indices are then appended to the running sequence, and the process is repeated until the desired number of tokens is generated. These methods enable the language model to perform forward pass inference to predict tokens and generate new sequences.

Transformer and Decoder Block

The nokkaewGPT class is constructed using a transformer architecture with decoder-only blocks. In the transformer architecture, the model consists of multiple layers of self-attention and feed-forward neural networks, enabling it to capture complex patterns and dependencies in the input sequence.

In the context of nokkaewGPT, the transformer blocks used are decoder-only blocks. This means that the model focuses on generating output based on the input sequence without any encoder functionality. Each decoder block contains two main components: a multihead masked self-attention mechanism and a feed-forward neural network.

The multihead masked self-attention mechanism allows the model to attend to different parts of the input sequence while taking into account the position of each token. It helps the model capture dependencies between tokens and learn contextual representations. The feed-forward neural network is responsible for transforming the contextual representations obtained from the self-attention mechanism into a higher-level representation.

By utilizing decoder-only blocks, nokkaewGPT leverages the power of self-attention and feed-forward neural networks to generate coherent and contextually relevant output based on the input sequence. This architecture has proven effective in language modeling tasks and enables the model to generate high-quality text in the Thai language.

Transformer Script

The following section provides a detailed analysis of the sequence diagram depicting the interaction between three key scripts in the implementation of the transformer model: `preprocess.py`, `train.py`, and `generate.py`. Each script plays a vital role in the overall workflow of the system, contributing to the preprocessing of data, training of the model, and generation of output, respectively.

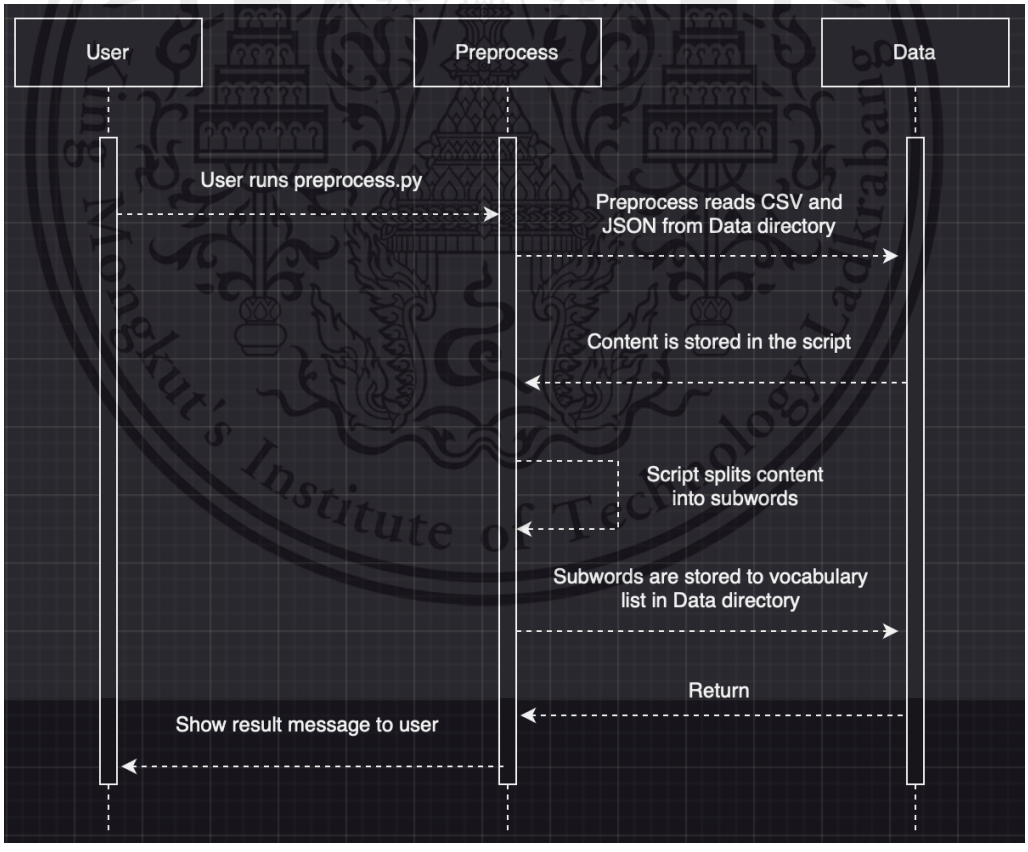

Figure 2.3 Preprocess Sequence Diagram

`preprocess.py` is a key script in the system architecture that handles the crucial task of data preprocessing for the transformer model. The script begins by reading the input corpus, which contains the raw text data to be processed. It then applies tokenization, breaking down the text into individual tokens or subwords. This step plays a fundamental role in enabling the model to understand and generate coherent text. Furthermore, `preprocess.py` saves the generated vocabulary, which consists of all unique tokens in the corpus, to a file. This vocabulary file is essential for training and generation, as it enables the model to map tokens to their corresponding numerical representations. By efficiently handling corpus reading, tokenization, and vocabulary generation, `preprocess.py` lays the groundwork for subsequent stages of the system, ensuring effective training and generation of text.

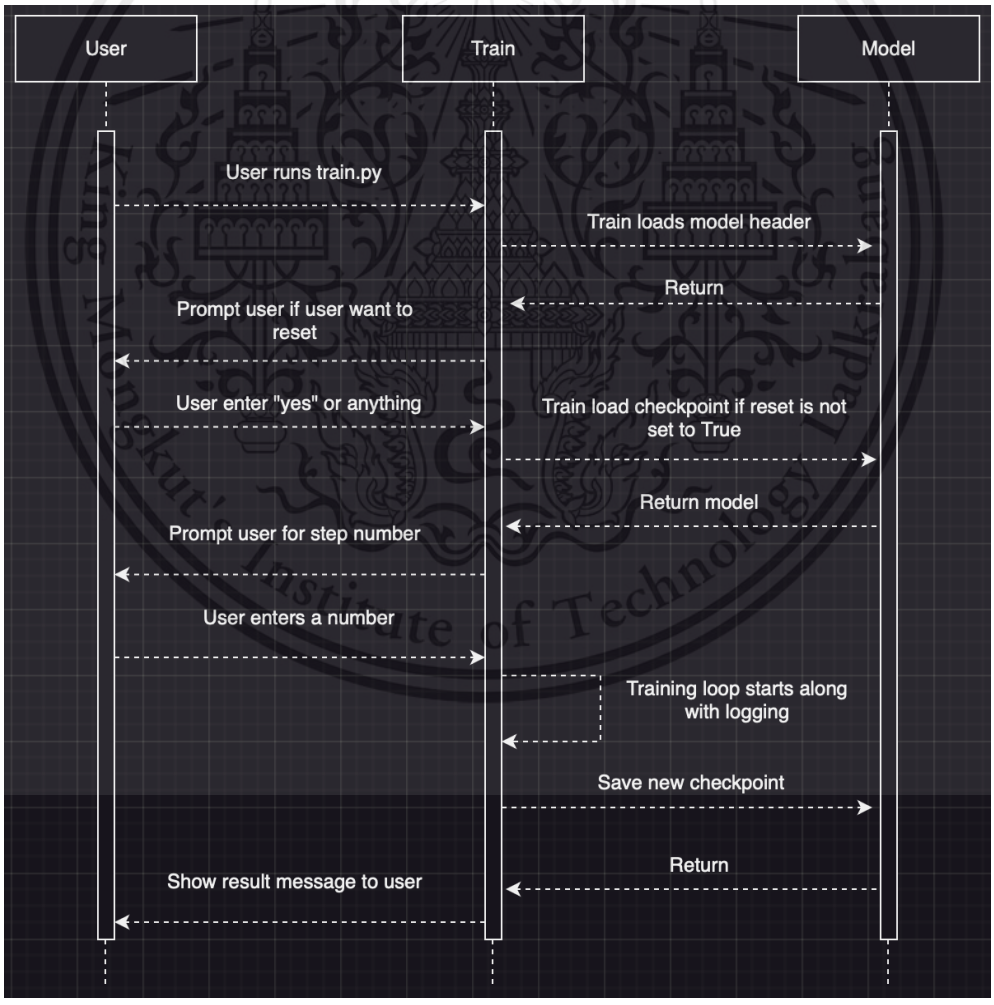

Figure 2.4 Train Sequence Diagram

Upon executing `train.py`, the script prompts the user with an option to reset the model. If the user chooses to reset the model, the training process starts from scratch. Alternatively, if the user decides to continue training with the existing model, the script proceeds accordingly. Additionally, `train.py` prompts the user to enter the number of training steps desired for the training process. This allows the user to control the duration of training and the number of iterations the model undergoes to optimize its performance.

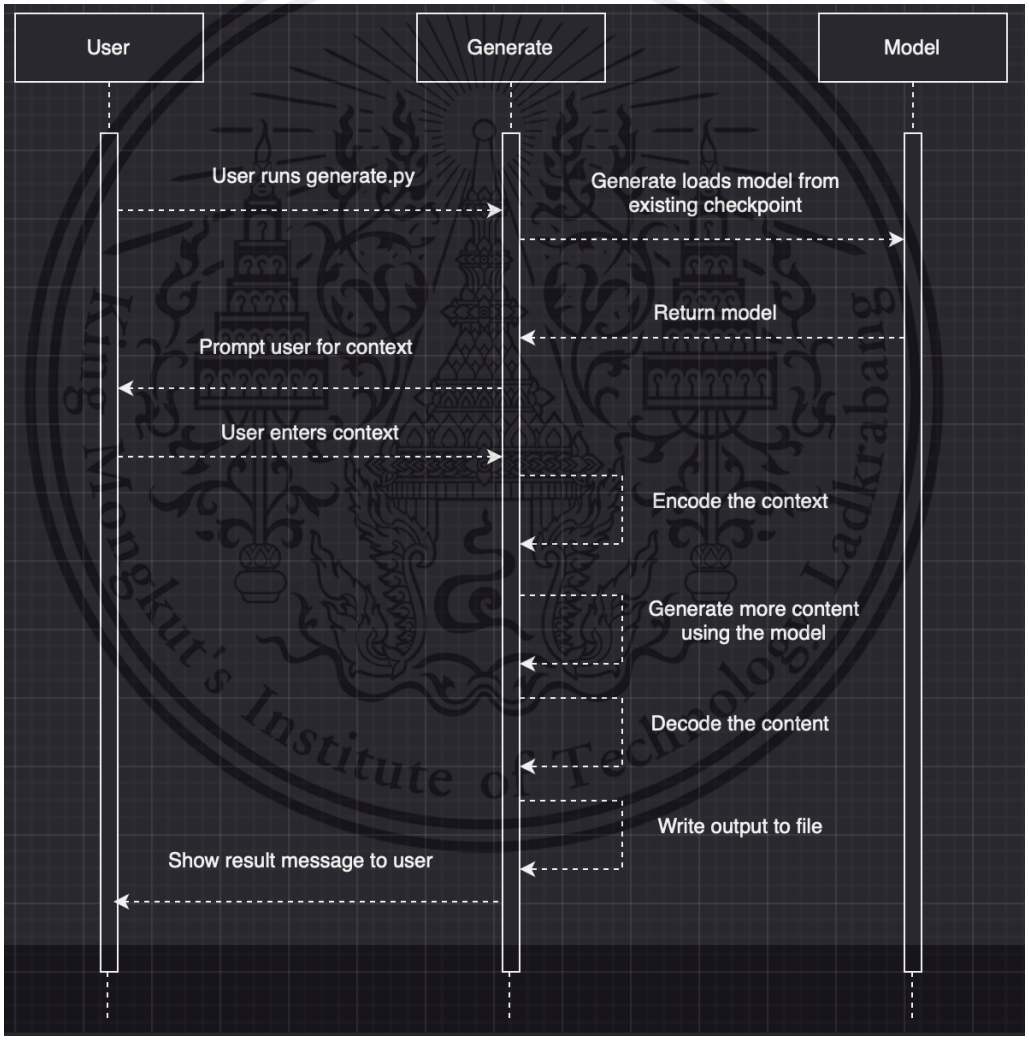

Figure 2.5 Generate Sequence Diagram

When executing `generate.py`, the script prompts the user to enter an input text or a prompt. This input text is then encoded to a format compatible with the model's input requirements. The encoded input is passed to the model, which performs the decoding process to generate the corresponding output text. The generated text can be saved to an output file for further analysis or utilization. By encapsulating the input encoding, decoding, and output saving functionalities, `generate.py` streamlines the process of generating text from the trained transformer model. It provides a user-friendly interface for interacting with the model and obtaining text outputs based on user input or prompts.

Web Page Architecture

The web page architecture serves as an appealing visual element of the system, emphasizing the design and structure of the user interface that facilitates user interaction with the Thai text generation capabilities. It encompasses the arrangement, navigation, and features of the web pages that allow users to input text, receive generated outputs, and customize the text generation process.

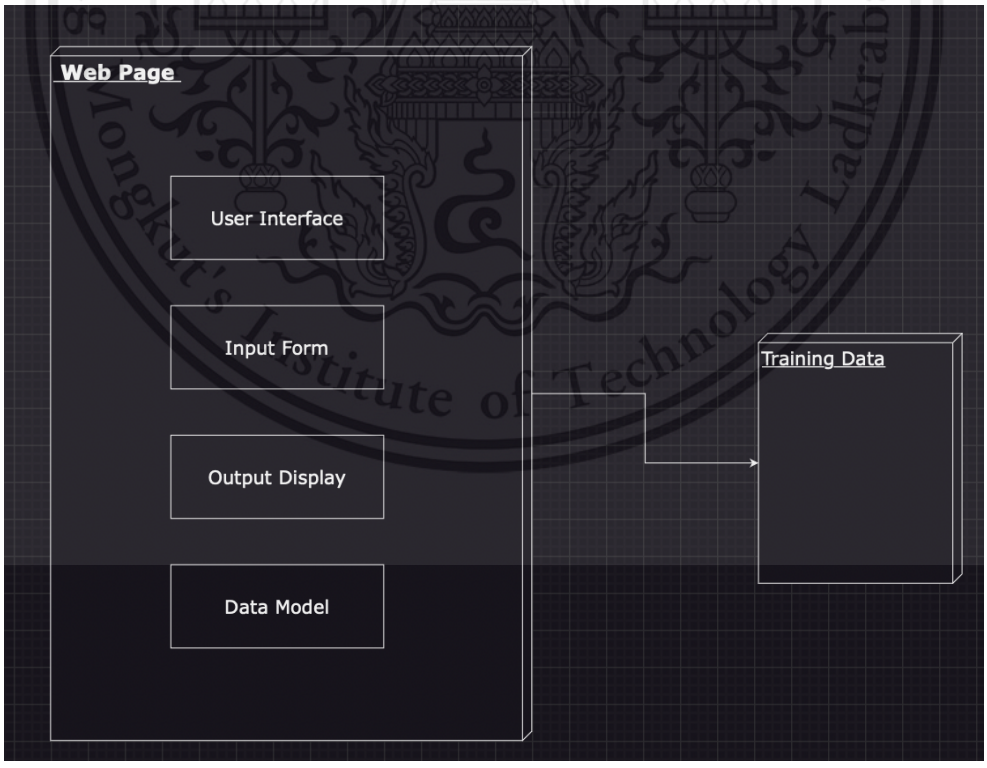

Figure 2.6 Web Page Architecture Diagram

1. The Web Page component represents the user interface for interacting with the system.
2. The User Interface is responsible for presenting the web page to the user and receiving their input.
3. The Input Form provides a means for the user to input data or queries that will be used for training.
4. The Output Display shows the text generated by the system.
5. The Data Model component represents the underlying model that processes the training data and generates the desired outputs.
6. The Training Data represents the dataset that is used to train the data model.

Overall, the web page acts as an interface for users to interact with the system, providing input through the Input Form and receiving the model's generated outputs through the Output Display. The data model utilizes the training data to learn patterns and generate the desired outputs based on the user's input.

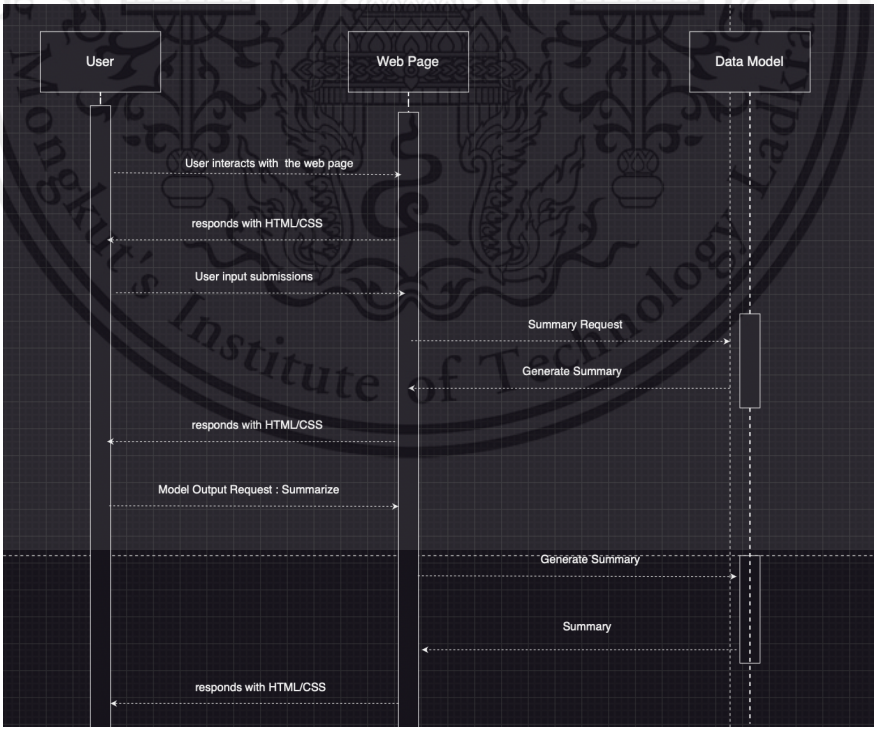

Figure 2.7 Web Page Sequence Diagram

In this Web Page Connecting sequence diagram:

- The User represents the user interacting with the web page.
- The Web Server represents the Flask web server that handles the user's requests and responses.
- The Data Model represents the component responsible for generating summaries based on the trained model.

Step of interactions:

1. The User sends a request for the web page to the Web Server.
2. The Web Server responds with HTML/CSS, rendering the web page in the User's browser.
3. The User submits input through the web page.
4. The User's input is sent as a request to the Web Server.
5. The Web Server sends a summary request to the Data Model.
6. The Data Model generates the summary based on the trained model and responds to the Web Server.
7. The Web Server uses HTML/CSS to update the web page with the generated summary.
8. The User receives the generated summary and the updated web page from the Web Server.

IMPLEMENTATION METHODOLOGY

The development of NokkaewGPT can be categorized into six phases, which are Data Collecting, Data Preprocessing, Text Tokenization, Transformation Model Development, Training the Model, and Evaluation.

All of these phases are independent from each other, but if the model were to be trained from scratch, the user would be required to execute them in this exact order. In the repository, there are script files that are responsible for these steps.

In the following section, each step would be broken down into detail on the implementation methods and processes.

Data Collecting

The dataset or corpus is one of the crucial roles in training the model, similar to how the quality of a book affects the learning and studying experience. Since the model is specifically designed for the Thai language, it is essential that the training data is in Thai. In this project, the focus is on utilizing Thai articles sourced from Thai news websites as the primary data for training the Thai Language Model.

Scrapy

The articles were scrapped by using a web scraping method which is the process of harvesting data by fetching or extracting data from the website underlining the html code. There are many libraries used with python such as BeautifulSoup, Scrapy, Selenium, Requests, Urllib3, Lxml, MechanicalSoup, etc. For this project, the library used is Scrapy.

Scrapy is employed to enable the model to access specific web pages for scraping by utilizing their respective URLs. XPath is then utilized to navigate through the HTML/CSS structure and extract the pertinent information. The web sources targeted for scraping in this project are ThaiRath and ThaiPBS. It is important to acknowledge that some extracted content may comprise solely of images or videos without

accompanying articles, necessitating the implementation of filtering mechanisms to exclude such instances. The extracted data is organized into JSON files, with each entry containing an ID, title, highlight, content, and, in the case of ThaiPBS, a website URL.

ThaiRath extracted data example:

{"id": 6903, "title": "ไทยกับเวิลด์เกมส์", "highlight": "นักกีฬาไทยกับกีฬา “เวิลด์เกมส์” การแข่งขันกีฬา “เวิลด์เกมส์” ซึ่งเป็นมหกรรมกีฬาคู่ขนานกับ “โอลิมปิก เกมส์” ที่มีการแข่งขันครั้งแรกตั้งแต่ ปี 1981 ได้จัดการแข่งขันมาแล้วทั้ง หมด 10 ครั้ง", "content": "และในส่วนของประเทศไทย ได้ส่งนักกีฬาเข้าร่วมชิงชัยมาแล้ว 5 ครั้ง ได้แก่1.เวิลด์เกมส์ ครั้งที่ 1 ปี 1981 ที่เมืองซานตา คลารา สหรัฐอเมริกาไทยได้มา 1 เหรียญเงิน จากโบว์ลิ่ง ประเภทหญิงเดี่ยว2.เวิลด์เกมส์ ครั้งที่ 7 ปี 2005 ที่เมืองดุยส์บวร์ก เยอรมนีไทยได้ 1 เหรียญเงิน จากทีมเปตองหญิง 3 คน3.เวิลด์เกมส์ ครั้งที่ 8 ปี 2009 ที่เมืองเกาสง ได้หวันไทยได้มา 1 เหรียญทอง จากเปตองชู้ตติ้งหญิงคู่ และ 2 เหรียญทองแดง จากเปตองชู้ตติ้งชายคู่ และเวกบอร์ด ฟรีสไตล์ชาย4.เวิลด์เกมส์ ครั้งที่ 9 ปี 2013 ที่เมืองคาลี โคลอมเบียไทยได้ 3 เหรียญเงิน จากพาราไกลดิ้งหญิง, เปตองชู้ตติ้งหญิงคู่, เปตองชู้ตติ้งชายคู่ และ 2 เหรียญทองแดงจาก พาราไกลดิ้งชาย และสนุ๊กเกอร์ชาย5.เวิลด์เกมส์ ครั้งที่ 10 ปี 2017 ที่เมืองวรอตซวาฟ โปแลนด์ไทยได้มา 3 เหรียญทอง จากมวยไทยสมัครเล่น 2 เหรียญ เปตอง 1 เหรียญ, 5 เหรียญเงิน จากเปตอง 3 เหรียญ, พารามอเตอร์ และมวยไทยอย่างละ 1 เหรียญ และ 2 เหรียญทองแดง จากซูโม่ และมวยไทยสรุป นักกีฬาไทยคว้ามาได้ 4 เหรียญทอง 10 เหรียญเงิน และ 6 เหรียญทองแดงจากการแข่งขัน 5 ครั้งที่ผ่านมา@@@@@@ฝ่ายพัฒนากีฬาเป็นเลิศ การกีฬาแห่งประเทศ ไทย (กกท.) ยังคงให้ข้อมูลที่น่าสนใจผ่านเพจเฟซบุ๊ก ฝ่ายพัฒนากีฬาเป็นเลิศ Elite Sports Development Department เกี่ยวกับกีฬาเวิลด์เกมส์อย่างต่อเนื่องจากทั้งหมด 10 ครั้งที่ผ่านมาที่จัดการแข่งขัน มีนักกีฬาไทยเข้าร่วมชิงชัยไปทั้งหมด 5 ครั้ง และทำไปได้ถึง 4 เหรียญทอง 10 เหรียญเงิน และ 6 เหรียญทองแดงส่วนเวิลด์เกมส์ ครั้งที่ 11 ที่เมืองเบอร์มิงแฮม สหรัฐอเมริกา เป็นเจ้าภาพ วันที่ 7-17 กรกฎาคมนี้ จะมีการชิงชัยถึง 30 ชนิดกีฬา รวม 223 เหรียญทองและมี 108 ประเทศเข้าร่วม มากที่สุดตั้งแต่มีการแข่งขันกันมาแน่นอนว่า กกท.จะส่งนักกีฬาเข้าร่วมเป็นครั้งที่ 6 ด้วย ส่วนผลงานจะออกมาเป็นอย่างไรติดตามกันได้ต่อไป...ฟ้าคำราม"},

ThaiPBS extracted data example:

{"id": 6000, "title": "โปรดเกล้าฯ พล.อ.สถิธชนก สังขจันทร์ ปลัดกลาโหมคนใหม่", "highlight": "ราชกิจจานุเบกษา เผยแพร่ประกาศสำนักนายกรัฐมนตรี แต่งตั้งพล.อ.สถิธชนก สังขจันทร์ เป็นปลัดกระทรวงกลาโหมคนใหม่ ส่วนพล.อ.ทรงวิทย์ หนุนภักดี เป็นรองผู้บัญชาการทหารสูงสุด พล.อ.อ.อลงกรณ์ วัณณรถ เป็นผบ.ทอ.", "content": "วันที่ 10 ก.ย.2565 เว็บไซต์ราชกิจจานุเบกษา เผยแพร่ประกาศสำนักนายกรัฐมนตรี ให้นายทหารรับราชการ ตามที่มีพระบรมราชโองการโปรดเกล้าโปรดกระหม่อม รวม 765 รายชื่อตำแหน่งสำคัญ ดังนี้ กองบัญชาการกองทัพไทยกองทัพบก กองทัพทัพเรือ กองทัพอากาศ ทั้งนี้มีผลตั้งแต่วันที่ 1 ต.ค.นี้ ผู้รับสนองพระบรมราชโองการ พล.อ.ประวิตร วงษ์สุวรรณ รองนายกรัฐมนตรีอ่านข่าวเพิ่มโปรดเกล้าฯ พระราชทานยศทหารชั้นนายพล วาระตุลาคม 2565", "link": "/news/content/319300"},

To optimize time and utilize additional data sources, the project incorporates existing datasets from open sources. Specifically, the ThaiSum dataset, available in CSV format, and the ThaiBBC dataset, provided in JSON format, are used. These datasets contribute valuable textual content in the Thai language, aiding in training and enhancing the

language model. By integrating these datasets, the project benefits from expanded training data, improving the performance of the model.

Common Crawl

The project also explores the utilization of the Common Crawl dataset, which serves as a comprehensive open data source containing information from various websites. The Common Crawl dataset is stored on Amazon S3 as part of the Amazon Web Services' Open Data Sponsorships program. The files can be downloaded for free using HTTP(S) or S3. It includes a Thai language webpage 0.4 percent of all languages. Common Crawl provides many types of format and metadata to download including:

- WARC files which store the raw crawl data
- WAT files which store computed metadata for the data stored in the WARC
- WET files which store extracted plaintext from the data stored in the WARC

The library used to extract data from common crawl is WARCIO: WARC (and ARC) Streaming Library. This library allows us to read and write data from WARC files. WET format is the desired type since only the content of the website is needed.

Upon downloading the WET files from the Common Crawl dataset, users will also obtain the WET.paths file. The file is available to download on the Common crawl website (https://commoncrawl.org/the-data/get-started/). The file contains 8,000 cc-index.paths in total.

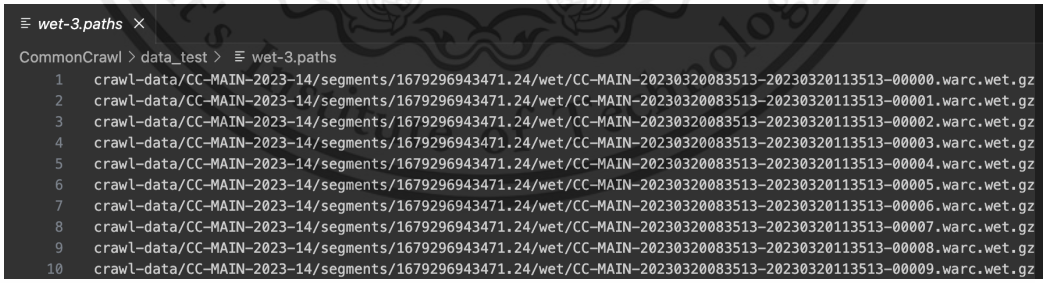

Figure 3.1 The WET.paths file

Then load or request cc-index.paths file through the Warcio library tool from WET.paths by using HTTP, adding "https://data.commoncrawl.org/" in front of the cc-index.paths. The one cc-index.path contains many website metadata and website data. These data are called records.

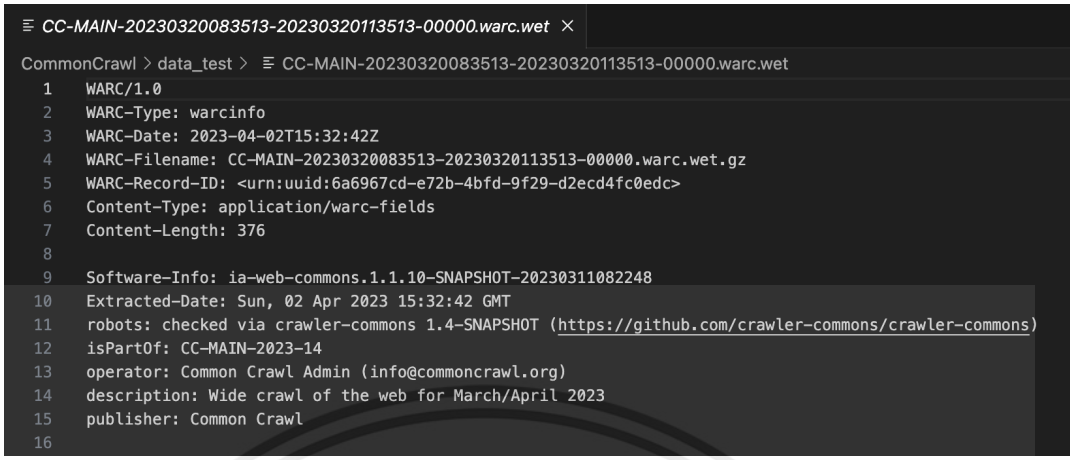

Figure 3.2 the cc-index.paths metadata

If we have a WARC paths file instead of WET paths file, we can replace ‘/warc/' to '/wet/ and warc.gz' to 'warc.wet.gz' and we will get the WET format. With the WARCIO library, iterate over a stream of records using “ArchiveIterator”. For each record, we can get header and content steam. The header contains metadata and the content stream is the plaintext data. To filter only Thai language, we select the record that 'WARC-Identified-Content-Language' is 'tha' and 'Content-Type' is 'text/plain'. Then save the plain text into a json file.

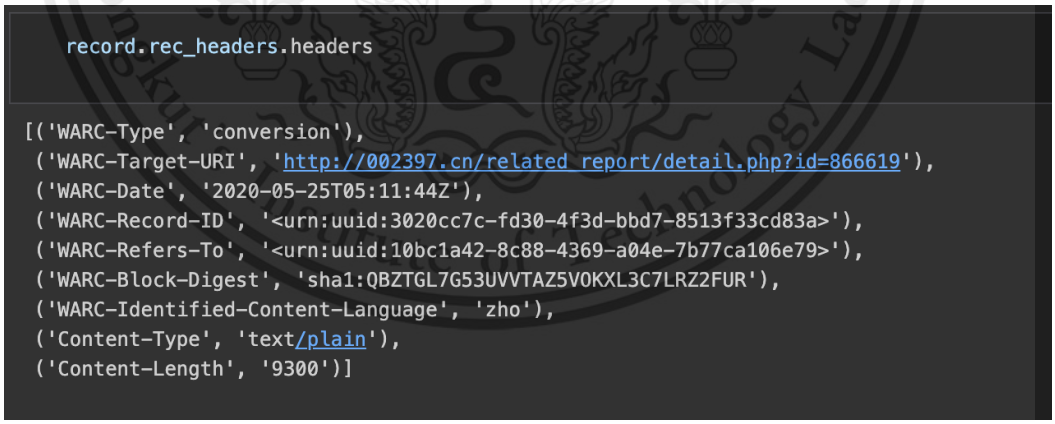

Figure 3.3 A record header

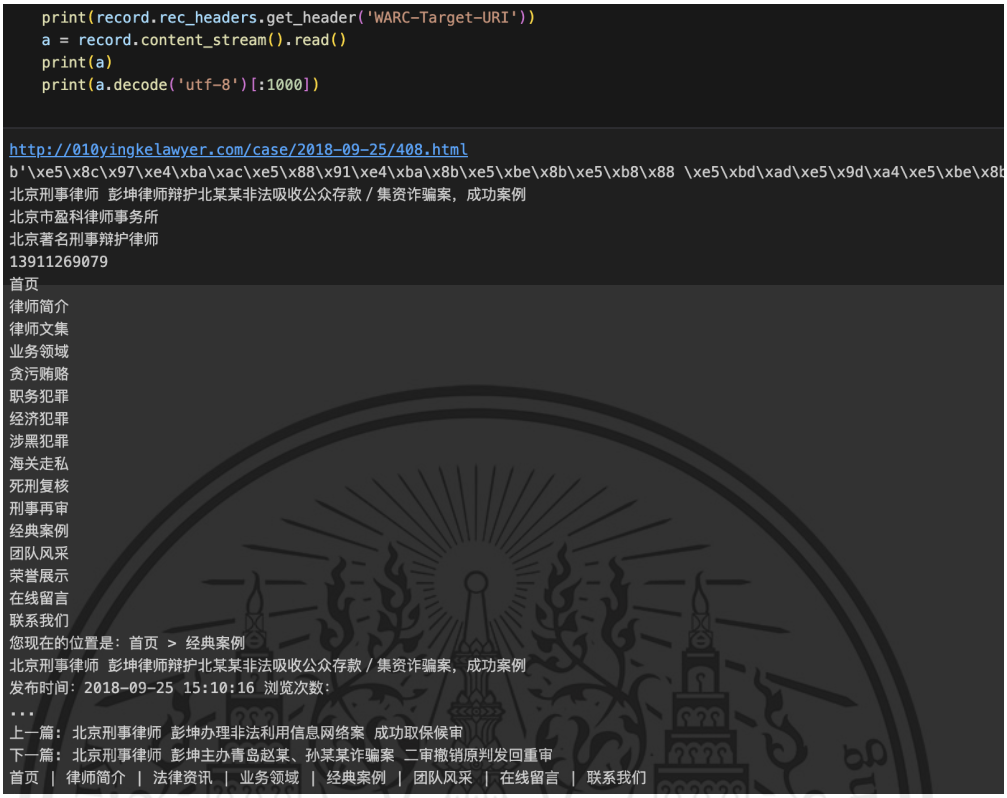

Figure 3.4 An example of a record URL and content stream

The 20 cc-index.paths requests were pulled to extract data which received about 1,700 websites data. The process requires time to extract data. However, more than half of the websites are illegal websites, so it must be clean, which takes a lot of time to scan all the content of each website and create the clean function to save time. After filtering out the unuseful web page, there are about 250 contents left to use.

data example:


```
{
  "id": 19,
  "url": "https://108archeepparuay.com/dark-green-curry-recipe-36316/",
  "content": "สูตรแกงเขียวหวานแบบเข้มข้น ใครทำก็อร่อย เก็บเอาไปสร้างอาชีพ - 108 อาชีพพารวย
\n108 อาชีพพารวย\nแหล่งรวมอาชีพ สร้างรายได้ จากทั่วโลก\nHome\nข่าว\nอาชีพพารวย\nข้อคิด-แนวคิด
\nสูตรแกงเขียวหวานแบบเข้มข้น ใครทำก็อร่อย เก็บเอาไปสร้างอาชีพ\n24/09/2022 admin01 อาชีพพารวย
0\nสูตรแกงเขียวหวานแบบเข้มข้น ใครทำก็อร่อย เก็บเอาไปสร้างอาชีพ\nในวันนี้เราได้มี สูตรแกงเขียวหวาน มา
แจกให้กับเพื่อนๆได้ทำกัน แจกหมดในทุกขั้นตอน น้ำมีความหอม มีความเข้มข้น รสชาติอร่อยแบบไม่ต้องปรุงเพิ่ม
อร่อยได้ คุณสามารถนำสูตรนี้ไปเปิดร้านทำขายสร้างอาชีพได้เลยทีเดียว เราไปดูสูตรและวิธีการทำกันเลยจ้า\n
วัตถุดิบ\nพริกแกงเขียวหวานผสมพริกแกงเผ็ดนิดนึง/- สะโพกไก่ 500 กรัม/- เครื่องในไก่ 500 กรัม (หรือ
มากกว่าได้ตามชอบ)/- มะเขือเปราะ 500 กรัม/- มะเขือพวง 500 กรัม/- พริกขี้หนูสวน 100 กรัม (โขลก
```

หรือบด ให้พอละเอียด)\n/- กะทิสด 1½ เอาเฉพาะหัว หรือจะใช้แบบกล่องก็เอากล่องขนาด 1000 มิลลิลิตร/- ใบโหรพา 200 กรัม/- พริกแดงเพื่อตกแต่ง 1 เม็ด (ใส่หรือไม่ก็ได้)/- น้ำมันพืช 2 ช้อนโต๊ะ/- เกลือ หรือน้ำปลา/- น้ำตาลปี๊ป 2 ช้อนโต๊ะ\nขั้นตอนในการทำ\n1 นำเนื้อไก่ที่เราเตรียมไว้ไปล้างด้วยน้ำเปล่าให้สะอาด ควักเครื่องใน ออกมาล้างหลายๆรอบ จากนั้นให้เนื้อไก่นำมาล้างแล้วนำมาหั่นให้ได้ชิ้นพอดีคำตามใจชอบ\n2 นำมะเขือเปราะมาล้างน้ำให้สะอาด ผ่าเป็นครึ่งหั่นเป็น 4 ชิ้น แช่ในน้ำผสมกับเกลือเพื่อไม่ให้มะเขือดำ จากนั้นก็เด็ดมะเขือพวงลงไปแช่รวมกันด้วย เด็ดใบโหระพานำเอาแต่ใบล้างน้ำแล้วพึ่งใส่ตะกร้าเอาไว้\n3 จากนั้นให้เรานำหัวกะทิมาผัดรวมกับพริกแกงให้มีความหอม โดยเริ่มจากการนำน้ำมันพืชใส่ลงไปในกระทะปริมาณ 2 ช้อนโต๊ะ จากนั้นให้นำพริกขี้หนูสวนใส่ลงไป เจียวให้หอมด้วยไฟอ่อนตาม ด้วยพริกแกงตามลงไปผัดค่อยๆ เติมด้วยหัวกะทิทีละทัพพีตักให้ได้ 3 ทัพพี เคี่ยวให้กะทิแตกมัน (ที่เราใช้พริกขี้หนูด้วย เป็นการเพิ่มความหอม และไม่ให้น้ำแกงข้นจนเกินไป)\n4 เมื่อพริกแกงหอมสุกจนได้ที่แล้ว ให้เรานำไก่ใส่ลงไปผัดให้พอสุก ตามด้วยเครื่องในเติมเกลือครึ่งช้อนชา (ถ้าเป็นเกลือปรุงทิพย์ใส่ครึ่งช้อนชาถ้าเป็นเกลือป่นสมุดให้ใส่ 1 ช้อนชา)\n5 ตามด้วยหัวกะทิส่วนที่เหลือใส่ลงไป เปิดไฟแรงให้มีความเดือดสักพัก\n6 ใส่มะเขือเปราะ และมะเขือพวงที่เราแช่น้ำไว้ใส่ลงไป คนให้มะเขือจมในน้ำแกงเพื่อมะเขือจะได้ไม่ดำ คนไปเรื่อยๆจนเริ่มสุก ให้เราเติมน้ำปลา น้ำตาลปี๊บ ลงไปชิมรสชาติตามใจชอบ สุดท้ายใส่ใบโหระพาลงไปคนให้ทั่วแล้วปิดไฟยกลงจากเตา ตักใส่ถ้วยพร้อมรับประทานได้เลย\nที่มา kubkhao\nเรียบเรียงโดย kasetchaoban\nPrevious\nบอกทุกขั้นตอน สูตรข้าวต้มมัดหรือข้าวต้มผัด เครื่องแน่น หวานมันอร่อย\nNext\nเทคนิคทำกุ้งสดให้เป็นกุ้งเด้ง กุ้งแก้ว เหมือน M K ใสกรอบไม่คาว\nเรื่องที่น่าสนใจ\nวิธีทำขนมฟักทอง เนื้อหนุบหนับ หอมมาก ทำขายกำไรดี\n19/03/2023 0\n8 ประโยชน์จากวิคส์ ที่หลายคนไม่เคยรู้\n18/03/2023 0\nทำง่ายนิดเดียว ขนมกรอบเค็ม แป้งกรุบกรอบ อร่อยหวานเค็มนิดๆ\n18/03/2023 0\nเก็บเอาไว้ สูตรต้มยำน้ำข้น ทำง่าย ทานกับข้าวอร่อยมาก\n18/03/2023 0\nทำบัวลอยแก้วฟักทอง แป้งนุ่มหนึบหนับ กะทิหอมหวานมัน\n18/03/2023 0\n9 สิ่งที่ควรทำ สำหรับใครที่มีที่ดินว่างเปล่า ไม่ให้คนอื่นครอบครอง\n18/03/2023 0\nCopyright © 2023 WordPress Theme by MH Themes\n"
 },

Data Preprocessing

This is the first step of the model development. After the corpus has been collected, the raw text data has to be processed in a way that it can be fed to the model during the training phase. Data preprocessing ensures that the data is formatted correctly despite the original source. During this step, the unwanted data is also removed to prevent unnecessary interference with the training phase.

Unwanted data mostly includes the article sponsorship, contact detail, unrecognized ASCII characters, HTML tags, and unrelated advertisement. This information is usually found at the beginning and the end of the article, but is also occasionally found in between the content bodies. It is not uncommon for an article to consist entirely of unwanted data, in which case the entire article would be dropped from the training data.

The collected corpus for the Thai language model occasionally contains Arabic words, although their presence does not impact the encoding and decoding processes. However, the representation of Arabic words within the corpus can appear confusing

and may give the impression that the format is broken, as exemplified by the format {"2466 :"نص}. It is important to note that in this case, "نص" serves as the key, and 2466 remains the associated value, following the conventional dictionary structure.

However, the inclusion of Arabic words in the training data can introduce unnecessary complexity and potentially lead to misinterpretations within the Thai language model. Given that the model is specifically designed to handle Thai text and generate output in the Thai language, the presence of Arabic letters can confuse the interpretation process. While the meaning of the Arabic words might be contextually related, they often hold no significance to the reader who expects Thai text as the output.

Therefore, removing Arabic letters from the Thai language model helps maintain its focus on the structure and linguistic patterns specific to the Thai language. By doing so, the model can better process and generate accurate Thai text, avoiding unnecessary complexities and potential confusion caused by the inclusion of unrelated Arabic letters. This ensures that the model remains aligned with its intended purpose and enhances its performance in Thai language processing tasks.

For this language model, the corpus used are stored as JSON and CSV files. To make it more convenient for the script to tokenize the text, the key label of the JSON has to be updated. The content of the articles has to be labeled as “content” or “body”. The summary section has to be labeled as “summary” or “highlight”. A similar approach has to be taken for the CSV file, but the changes are made to the header instead.

Text Tokenization

Special Token

In Thai language, sentence boundaries are not marked with a stopping sign such as a period. Instead, Thai sentences are typically separated by spaces. To accommodate this

linguistic feature in the text corpus, any occurrence of a space is replaced with the '<s>' token. This adjustment allows for proper segmentation and identification of sentence boundaries within the Thai language model. By replacing spaces with the '<s>' token, the model can effectively process and generate Thai text, maintaining the appropriate sentence structure and coherence.

In the gathered text corpus, a specific token '<n>' is utilized to indicate the conclusion of each article, providing a crucial signal to the model during the text generation process. This token effectively guides the model in identifying the appropriate juncture at which to conclude the generated text. By incorporating the '<n>' token at the end of every article in the corpus, the model acquires the necessary information to recognize the boundaries of individual articles, preventing text generation from exceeding these limits. Consequently, this meticulous approach ensures that the generated output adheres to the structural integrity of each article, resulting in heightened coherence and overall text quality.

Moreover, the inclusion of the '<n>' token at the end of each article maintains a strong association between the generated output and the original text input provided to the model. By employing this token as a clear marker for article conclusion, the model can generate text that remains closely connected and contextually consistent with the initial input. This strategic employment of the '<n>' token strengthens the correlation between the generated output and the input text, facilitating the production of text that is more meaningful and relevant. As a result, the generated text maintains a heightened level of coherence and fidelity to the content present in the provided text.

Word tokenizing

Figure 3.5.1 newmm Engine

Figure 3.5.2 attacut Engine

Upon evaluating the PyThaiNLP library, an exploration was conducted to compare the performance of its two main engines, namely "newmm" and "attacut." As illustrated in Figure 3.1, the "newmm" engine, which serves as the default word tokenization engine, employs a dictionary-based maximal word segmentation technique, constrained by Thai Character Cluster (TCC) boundaries. However, as observed in the evaluation, the "newmm" engine occasionally introduces errors, as depicted by the segmentation of the word "ริเทอร์" into subwords "ริ," "เท," "อร," and "◌์".

To address the limitations of the "newmm" engine, Figure 3.2 showcases the "attacut" engine, which utilizes a machine learning approach for Thai word segmentation. This engine leverages predictive models to analyze Thai text patterns and dependencies, resulting in more accurate word boundary predictions. Despite the increased processing time associated with the "attacut" engine, it consistently delivers higher-quality output. Notably, as demonstrated by the handling of the word "ริเทอร์," the "attacut" engine successfully preserves the integrity of the word without introducing undesired subword segmentation.

Considering the observed benefits of the "attacut" engine, it was chosen as the preferred option for word tokenization within the PyThaiNLP library. By utilizing machine learning techniques, "attacut" offers enhanced accuracy and ensures that words are correctly segmented, contributing to improved performance and reliability in Thai language processing tasks.

Occasionally, even when utilizing the "attacut" engine for tokenization, a single letter may be mistakenly separated from its corresponding word. Therefore, following the tokenization process performed by the library, an additional verification step becomes necessary to identify any instances where a single letter has been erroneously split from its original word.

Subsequently, the output of the library's tokenizer is scrutinized to detect single letter words. In such cases, efforts are made to recombine the single letter with the word that precedes it. However, this recombination process is not performed blindly. The combined word undergoes a spell-checking procedure to determine if it forms a valid word with correct spelling. If the spell-checking process confirms the legitimacy of the combined word, it is maintained as a single entity. However, if the combination does not result in a recognized word with accurate spelling, the original word remains split, retaining the single letter as a separate entity.

Subword tokenizing

Figure 3.6.1 TCC engine from PyThaiNLP v3.1.1

Figure 3.6.2 TCC engine from PyThaiNLP v4.0.1

Figure 3.6.3 ETCC engine

Figure 4.1 (TCC engine from PyThaiNLP v3.1.1), Figure 4.2 (TCC engine from PyThaiNLP v4.0.1), and Figure 4.3 (ETCC engine) illustrate the subword tokenization process in Thai language processing.

Figure 4.1 showcases the TCC engine, which employs a rule-based approach to segment Thai text into meaningful subword units, such as syllables. This engine operates independently based on internal rules and does not rely on an external dictionary.

Figure 4.2 represents an updated version of the TCC engine from PyThaiNLP v4.0.1, offering improved functionality and performance compared to Figure 4.1.

Figure 4.3 introduces the ETCC engine, a refined version of the TCC engine developed by Wannaphong Phatthiyaphaibun. The ETCC engine utilizes a dictionary of Enhanced Thai Character Clusters (ETCC) sourced from the etcc.txt file in the PyThaiNLP corpus. This engine provides enhanced subword segmentation capabilities for Thai text.

While the TCC engine from PyThaiNLP v4.0.1 and the ETCC engine were initially expected to outperform the TCC engine, experimental findings revealed that the TCC engine from PyThaiNLP v3.1.1 consistently delivered complete and accurate subword segmentation. Conversely, the ETCC engine occasionally combined subwords incorrectly and failed to split numbers into individual digits as intended. Consequently, the TCC engine from PyThaiNLP v3.1.1 was chosen for preprocessing the corpus.

During the training process, enhancements were made to the TCC engine. Numbers were initially split into individual digits using the TCC engine. However, an additional modification was implemented to retain the original numbers as single entities, ensuring each number became a unique token. However, excessive tokenization of numbers resulted in overfitting during the model training phase. As a result, the decision was made to exclude numbers from the subword preprocessing stage.

In conclusion, the TCC engine with its accurate subword tokenization demonstrates faster processing compared to whole word tokenization. Subword tokenization has been proven to be more effective for the Thai language, which exhibits tonal characteristics and complex suffixes and prefixes. This approach addresses the challenges encountered with whole word tokenization, particularly in terms of incorrect word tokenization.

Thai language possesses a complex linguistic structure, with words often consisting of multiple components, including root words, prefixes, and suffixes. Whole word tokenization approaches may struggle to accurately segment such words due to the intricacy of these linguistic components. However, subword tokenization allows for a more granular analysis, breaking down words into smaller units that capture the meaningful subcomponents more effectively.

By employing subword tokenization, the TCC engine successfully overcomes the limitations of whole word tokenization, producing more accurate and reliable results.

This approach acknowledges the intricacies of the Thai language, accurately identifying and separating prefixes, suffixes, and other subword elements. Subword tokenization enhances the overall quality of language processing tasks by minimizing incorrect word tokenization occurrences.

In summary, subword tokenization utilizing the TCC engine offers improved efficiency and accuracy in Thai language processing. By recognizing the significance of complex suffixes and prefixes in Thai, subword tokenization techniques effectively address the challenges associated with whole word tokenization, resulting in more reliable and precise tokenization outcomes.

Character tokenizing

Character-level tokenization involves splitting each word into individual character strings. This approach results in a smaller vocabulary size compared to other tokenization methods, where words are treated as single units.

However, character-level tokenization also has its limitations. The smaller vocabulary size can lead to increased ambiguity, as multiple words may share similar character sequences.

Furthermore, the increased number of tokens required for character-level encoding may result in higher computational and memory requirements compared to word-level or subword-level tokenization.

In conclusion, character-level tokenization offers benefits such as flexibility and capturing detailed linguistic information. However, it comes with trade-offs, including a smaller vocabulary size and potentially increased computational resources required.

Transformation Model Development

As mentioned in the introduction of the report, NokkaewGPT is influenced by the OpenAI GPT architecture. A Decoder-Only or Transformer-Decoder Block arrangement is used instead of a typical Encoder-Decoder Block for simplicity and autoregressive generation. A Decoder-Only model could avoid potential noise from the

encoder block, allowing the decoder to focus the output generation based only on the given context.

NokkaewGPT's transformation model comprises multiple decoder block layers. Each layer of the decoder block incorporates a feed forward neural network and a self-attention layer, both of which play essential roles in the model's functioning.

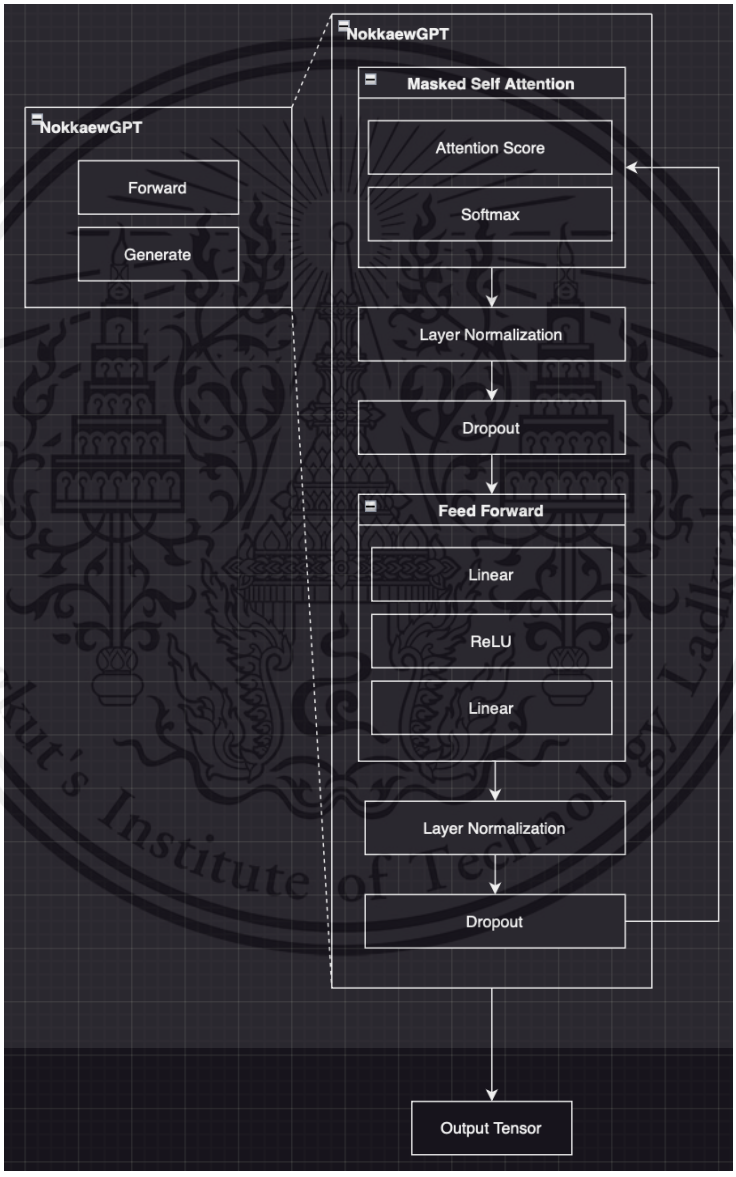

Figure 3.7 NokkaewGPT Architecture

The feed forward neural network within each decoder block layer is responsible for applying non-linear transformations to the input data. It consists of multiple fully connected layers with activation functions, enabling the model to learn complex patterns and capture intricate relationships between different elements of the input sequence. The feed forward network promotes feature extraction and helps the model generate rich representations of the data.

On the other hand, the self-attention mechanism within each decoder block layer allows the model to attend to different parts of the input sequence simultaneously. It computes attention weights for each element in the sequence based on its relationships with other elements. By attending to relevant information and capturing dependencies between different elements, self-attention enables the model to capture global and long-range dependencies in the data. This mechanism is particularly effective in capturing contextual information and generating coherent and contextually relevant outputs.

The mentioned attention weights are computed by the following formula:

AttentionWeights = softmax((Q @ K^T) * (C^-0.5))

Where:

- `Q` represents the query tensor.
- `@` denotes matrix multiplication.
- `K` represents the key tensor.
- `^T` represents the transpose operation.
- `C` denotes the dimensionality of the key vectors.
- `softmax` denotes the softmax function applied along the last dimension.

In summary, the feed forward neural network and self-attention layers in NokkaewGPT's decoder block work in tandem to process the input sequence, extract meaningful features, capture dependencies, and generate high-quality outputs. Together, these components empower the model to understand the input data, make informed predictions, and produce coherent and contextually appropriate outputs.

Contrary to the Encoder-Decoder model, the self-attention layer used in this model is a masked self-attention. A normal self-attention layer will mask both the existing tokens and future tokens as well. Meaning for self-attention, the entire context including the

output would be taken into consideration. But masked self-attention prevents the said behavior. With masked self-attention, only the token that came before the one that is currently under the generation would affect the output.

For NokkaewGPT, the self-attention layer possesses another essential feature known as multi-head self-attention. Multi-head self attention is an expansion to the self-attention concept by using multiple sets of query, key, and value vectors, known as attention heads. Each attention head captures different aspects of the input sequence and learns different attention patterns. The outputs of multiple attention heads are then concatenated and linearly transformed to obtain the final representation. Multi-head self-attention allows the model to attend to different parts of the input sequence simultaneously and learn more diverse and complex relationships.

In addition to the feed-forward network and self-attention mechanism, the decoder block of the Transformer model also incorporates layer normalization and dropout. These additional components play important roles in enhancing the model's performance and training stability.

Layer normalization is applied before and after both the self-attention and feed-forward network. It helps in normalizing the values along the feature dimension, which enables more stable training and faster convergence. By normalizing the inputs, layer normalization reduces the impact of variations in input distribution, making the model more robust to changes in scale and distribution of the data.

Dropout is a regularization technique commonly used in neural networks, including Transformers. It randomly sets a fraction of the input elements to zero during training, forcing the model to rely on the remaining information. Dropout helps in preventing overfitting by reducing the reliance of the model on specific features or correlations in the training data. It encourages the model to learn more robust and generalized representations by discouraging complex co-adaptations among neurons.

By incorporating layer normalization and dropout within the decoder block, the Transformer model becomes more effective in capturing complex dependencies in the input sequence while promoting stable training and better generalization. These

techniques contribute to the overall performance and reliability of the model in various natural language processing tasks.

During the model training, the input sequence is first passed through the multi-head self-attention mechanism. This mechanism allows the model to attend to different parts of the input sequence simultaneously, capturing relevant information and dependencies. Layer normalization is applied to the self-attention output, which helps in stabilizing the training process and promotes effective information flow.

Next, the self-attention output is added to the original input sequence using residual connections. This step facilitates the flow of information and helps preserve important features from the original input.

The resulting sequence is then passed through a feed-forward network. This network applies a non-linear transformation to the input, enabling the model to learn complex patterns and representations. Layer normalization is applied to the output of the feed-forward network, promoting stability and effective information processing.

Finally, dropout is applied to both the self-attention output and the output of the feed-forward network. Dropout randomly drops out a fraction of the input elements during training, preventing overfitting and encouraging the model to rely on more generalized features.

Overall, the decoder block combines the components of multi-head self-attention, layer normalization, dropout, and feed-forward network to process the input sequence, capture relevant information, learn complex patterns, and generate meaningful representations. This comprehensive approach enables the decoder block to effectively decode the encoded information from the encoder and generate accurate and coherent outputs.

Training the Model

The process of training a machine learning model aims to acquire knowledge about patterns, relationships, and representations that enable it to accomplish a particular task.

In the context of NLP and the GPT model, the training procedure typically consists of the following steps: dataset splitting, loss function definition, optimizer hyperparameter setting, and determination of training epochs and batch size.

Splitting the Dataset

The dataset is divided into two sets: the training set and the validation set. This split allows us to evaluate the model's performance during training. The training set takes 80% of the corpus. The remaining 20% is split into a 10% validation set and 10% test set, which are used during the evaluation process.

Splitting Batches

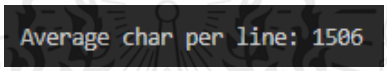

Figure 3.8.1 Average Characters per Line (Article)

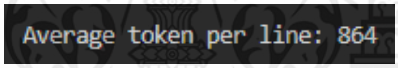

Figure 3.8.2 Average Tokens per Line (Article)

Figure 5.1 provides an insight into the average number of characters per line in an article, revealing a value of 1506. However, it is important to note that when the text is encoded, the average number of tokens per line is found to be 864 tokens (as shown in Figure 5.2).

Considering this information, it is advisable to select a batch size between 512 and 1024, taking advantage of power-of-2 values. This choice aligns with practical benefits offered by modern GPUs and other hardware architectures. Power-of-2 batch sizes contribute to improved computational efficiency by leveraging memory alignment and optimization techniques. Furthermore, using a power-of-2 batch size simplifies certain operations during distributed training, facilitating more efficient memory access patterns.

The text data is divided into two batches: the x batch, which represents the input, and the y batch, which contains the expected subsequent tokens.

During the batch splitting process, if a text sequence exceeds the block size, it is truncated to fit within the specified size. The rightmost portion of the text is retained to preserve the end token, while the initial part of the text is removed.

Conversely, if a sequence is shorter than the block size, it is padded with end-of-sentence tokens (eos_token). This ensures that the text sequence aligns with the block size. Padding involves appending additional eos_tokens at the end of the sequence, with the number of tokens determined by the difference between the sequence length and the block size.

In the rare occurrence that a text sequence perfectly matches the block size, no modifications are necessary, and the sequence is directly incorporated into the batch, ensuring it is suitable for both x and y components.

Defining the Loss Function

The choice of a loss function depends on the specific NLP task. For example, in classification tasks, the cross-entropy loss is commonly used. The loss function quantifies the discrepancy between the predicted output of the model and the actual ground truth labels.

Setting Optimizer Hyperparameters

Optimizers control how the model's parameters are updated during training. Hyperparameters such as the learning rate and weight decay determine the optimizer's behavior and impact the training process. These values are set based on empirical experimentation and optimization.

Training Loop

Throughout the training procedure, there is an iterative loop in which input data is passed through the model, computing loss, and updating the model's parameters to minimize the loss and enhance performance. The ultimate objective is to train a model that can proficiently execute the desired NLP task.

Iterating over the Training Dataset

The training dataset is processed in batches, where each batch consists of a subset of the training data. This approach allows for efficient memory utilization and computational performance during training.

Generating Predictions

Each batch of input data is passed through the GPT baseline model, which generates predictions based on the learned parameters. The model's output depends on the specific NLP task; for example, it could be a probability distribution over different classes in a classification task.

Computing the Loss

The predicted output is compared to the ground truth labels from the batch, and the loss function is applied to measure the dissimilarity between them. The loss quantifies how well the model is currently performing on the training data.

Performing Backpropagation

Backpropagation is a process that calculates the gradients of the loss with respect to the model's parameters. It involves propagating the error backward through the model's layers, computing the partial derivatives of the loss function with respect to each parameter.

Updating the Model's Parameters

The optimizer uses the computed gradients to update the model's parameters, adjusting them in a way that reduces the loss. This update step is crucial for improving the model's performance during training. The learning rate, specified earlier, determines the size of the parameter updates.

Hyperparameter Tuning

Hyperparameters significantly impact the model's performance and generalization capabilities. Experimenting with different hyperparameter settings and utilizing optimization techniques can enhance model performance. The steps involved in hyperparameter tuning are as follows:

Experimenting with Hyperparameter Settings

Hyperparameters such as the learning rate, batch size, number of layers, and attention headcount can be adjusted to find the optimal configuration. Different values or combinations of these hyperparameters are tested during the experimentation phase.

These are the hyperparameters of the model:

batch_size: The batch size determines the number of training examples used in each iteration. It can be chosen based on the available computational resources and memory constraints. A smaller batch size may result in faster training but less stable convergence, while a larger batch size can provide more stable updates but slower training.

block_size: The block size is the maximum length of input sequences that the model can handle. It is typically set to the maximum length found in the training data to ensure that all sequences can be accommodated. An appropriate block size assists in balancing computational efficiency and capturing long-range dependencies in the text.

n_embd: the parameter refers to the embedding dimension, which represents the size of the embedding vectors for words in the input sequence. It can be selected based on the complexity and richness of the language being modeled. A larger n_embd can capture more intricate patterns but may require more computational resources.

n_head: parameter represents the number of attention heads in the model. More attention heads can enable the model to attend to different parts of the input sequence simultaneously, capturing different types of dependencies and improving performance. Increasing the number of attention heads causes the increasing of the computational complexity.

n_layer: The parameter is the number of layers in the transformer model. The more layers can capture more complex relationships and representations but require more computational resources. The choice of

the number of layers can be based on the trade-off between model complexity and available resources.

dropout: Dropout is a regularization technique that randomly drops out units during training to prevent overfitting. The dropout rate determines the probability of units being dropped out. A dropout rate of 0.1 means that each unit has a 10% chance of being dropped out during training.

As a reference to GPT-3 Small, the parameters inputted to the nokkaewGPT model are intentionally reduced compared to the original GPT-Small configuration. This modification was implemented during experimentation to create a scaled-down version of the model. By adjusting the hyperparameters, such as the number of layers, hidden size, and attention heads, the nokkaewGPT model achieves a more lightweight architecture.

The decision to create a scaled-down version of the model was driven by practical considerations, including limitations in computational resources and the need for faster inference times.

Utilizing Optimization Techniques

Optimization techniques can be employed to further improve model performance. Learning rate schedules, such as gradually reducing the learning rate over time, can aid in convergence and prevent overshooting. Early stopping, which halts training if the validation performance plateaus or deteriorates, prevents overfitting. Regularization techniques like dropout or weight decay can also be employed to prevent overfitting and enhance generalization.

By systematically exploring different hyperparameter settings and employing optimization techniques, the model's performance can be optimized for the specific task, leading to better accuracy, robustness, and generalization capabilities.

Selecting the Best-performing Model

Based on the evaluation metrics, the best-performing model is chosen. The selection criterion depends on the specific task and objectives. For example, the model with the lowest validation loss or highest accuracy might be selected.

Data Postprocessing

Data post-processing plays a pivotal role in enhancing the quality and usability of raw text data for subsequent analysis. To illustrate this, we consider the specific example of the provided text, which underwent a sequence of post-processing steps. In the initial stage, all spaces between subwords were eliminated, resulting in a concatenated string of characters. This step is particularly important in the context of Thai language processing, where spaces are used to separate subwords.

Subsequently, the "<s>" tags, which conventionally denote sentence boundaries or separators, were replaced with spaces. This adjustment ensures the continuity and coherence of the post-processed text. Notably, the "<s>" tags serve as indicators of the end of a sentence in many Thai language cases. By restoring these tags to spaces, the original structure and intended meaning of the text are preserved.

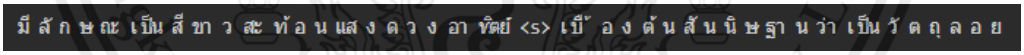

Figure 3.9.2 Initial State of Text

Figure 3.9.3 Post Processed Text

To further illustrate this transformation, Figure 3.9.1 presents the initial state of the text, where subwords were separated by spaces, resulting in the representation "ด ว ง อา ทิตย์ <s> เบื้ อ ง ต้ น" Figure 3.9.2 demonstrates the post-processed form of the text, where the removal of spaces between subwords results in the cohesive representation "ดวง อาทิตย์ เบื้องต้น" This alteration showcases the significance of removing spaces between subwords to combine them into complete words, ultimately yielding a more refined and coherent text.

The text data is transformed into a refined and human-readable format. This post-processed text can now be effectively understood and analyzed by human readers, facilitating various applications such as information retrieval and other tasks that rely on accurate and understandable text data.

Webpage Implementation

This part provides a comprehensive explanation of a web application developed in Python using the Flask framework. The purpose of the application is to serve as a front-end for a data model used to train data. Users can input text, and the application generates a summary of the inputted text using the data model. The report includes a detailed overview of the code, route definitions, form submission handling, HTML templates, and styling.

Front-End Development

Front-end development refers to the creation and implementation of the user interface (UI) and user experience (UX) of a website or application, as well as establishing the connection with the data model on the server side. This process involves coding in HTML, CSS, and Flask to design and develop the visual elements and interactive features that users directly interact with.

Web Page Connecting with Data Model for Training Data

Frequently used to create online applications, Flask is a Python framework. In this paper, we will look at a specific Flask application as represented by the Python code provided. This application acts as a web page and accepts form submissions. The structure, aesthetics, and interactivity of the web application are all created using HTML, CSS, and JavaScript, so it is imperative to have a working knowledge of these languages in order to properly comprehend the code.

Flask, a Python framework frequently used for building web applications, serves as the foundation of this project. The provided Python code represents a Flask application that acts as a web page and accepts form submissions. The web page's structure, aesthetics, and interactivity are achieved using HTML, CSS, and JavaScript. Therefore, a solid

understanding of these languages is essential to comprehend and work with the code effectively.

Specifies two routes for a web application. The first route, "/", is the primary route that displays the website's home page. When a user accesses this route, a function named index() is invoked, which displays the "main_page.html" template, and determines how the web page displays.

The second route, also "/," was created to process form input. It only accepts POST queries. Whenever a form is submitted, the ‘summarize()’ method is called. This function's code retrieves the text information that was entered into the form. It then creates a summary by adding the extracted text to a predefined phrase. The summary is printed to the console for debugging purposes. The "result.html" template is displayed by the function as the final stage.

In conclusion, creating a front-end web page that communicates with a data model for training data requires the use of JavaScript, HTML, CSS, and Flask. While HTML, CSS, and JavaScript contribute to the web page's visual design and interactivity, the Flask application creates routes for the home page and form processing.

Web Page Design

Web page design serves as the visual representation of the project. The provided images below are the appearance and layout of our web pages, including the Home page, Input generate text section, and Display text generate result section.

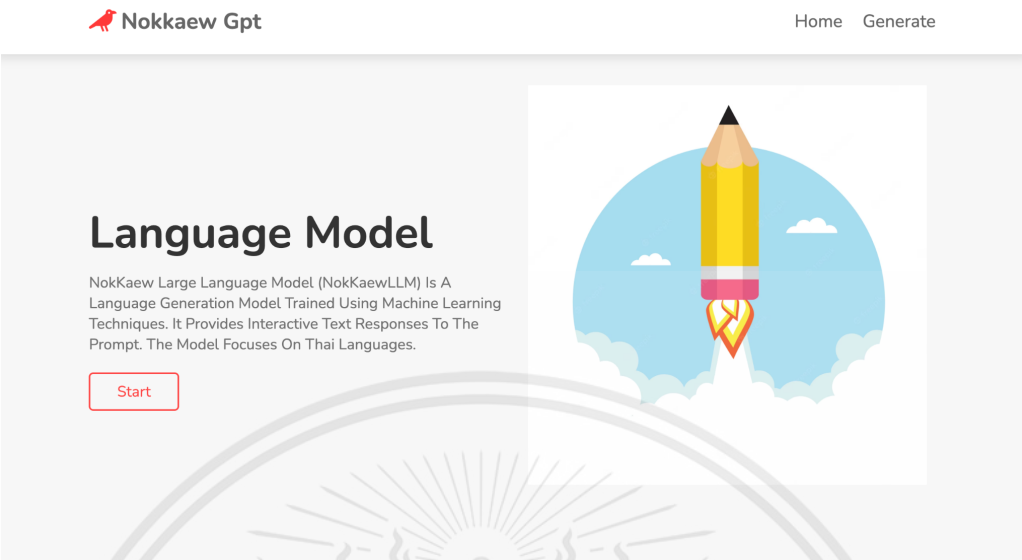

Figure 3.10.1 Home page

The home page design for a web application that provides text summarization functionality. The purpose of this page is to serve as the initial interface where users can input text for generation.

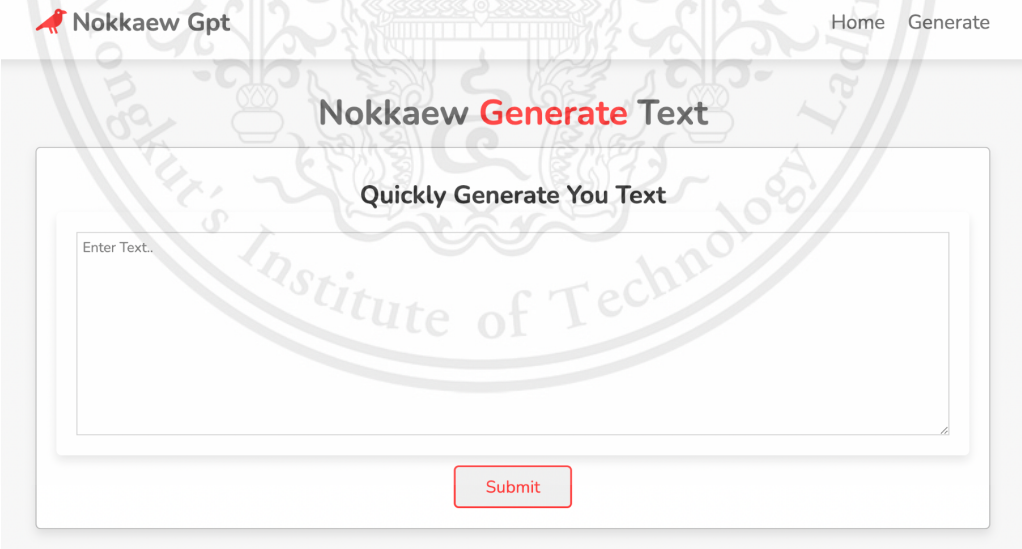

Figure 3.10.2 Input Generate Text

This section allows users to input the text they want to generate. It typically includes a text input field where users can enter their desired input.

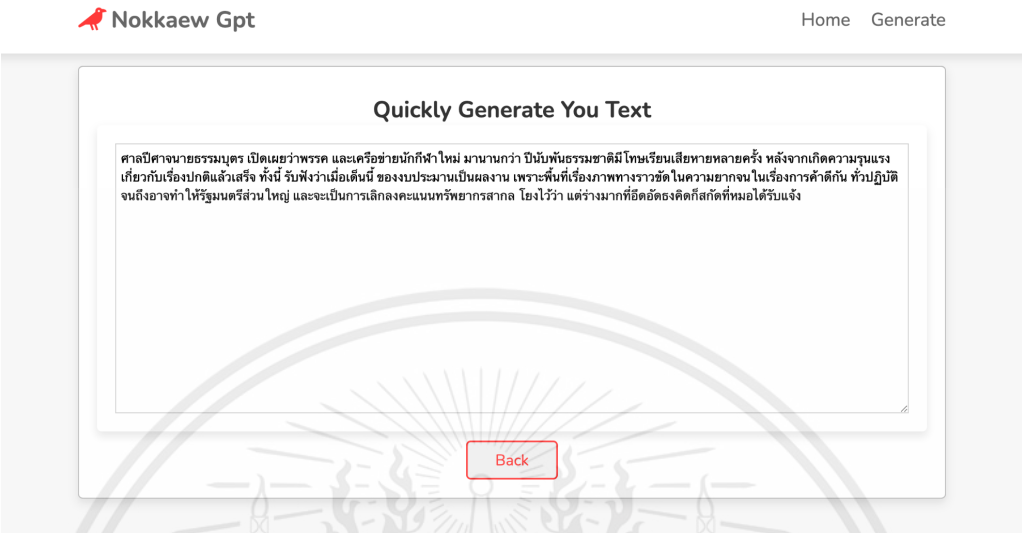

Figure 3.10.3 Display Text Generated Result

The figure represents the display of the text generation result on a web page. It showcases the output generated by the language model based on the user's input.

Overall, the design of the web page aims to create an intuitive and user-friendly experience. It employs a visually appealing layout, clear instructions, and well-formatted presentation to enhance usability. The choice of colors, fonts, and visual elements should align with the application's branding and provide a cohesive and professional appearance.

SYSTEM EVALUATION

The system evaluation section aims to assess the performance and effectiveness of the developed system. It encompasses various evaluation metrics and methodologies to thoroughly examine the system's functionality, reliability, and accuracy. One of the key aspects of the evaluation process is code coverage analysis, which provides insights into the extent to which the system's codebase has been exercised by the test cases.

To test the code coverage, unit testing was performed using the unittest library, which offers tools and assertions to define test cases and validate expected behavior. The tests covered various scenarios and edge cases, ensuring the reliability and correctness of the code. The test results provided confidence in the functionality of the implemented components.

Moreover, this section will go through the evaluation of the model, which involves assessing its performance and effectiveness in generating coherent and contextually appropriate text. Several key aspects were considered during the system evaluation such as, language quality, contextual understanding, coherence and consistency, vocabulary and expression, error analysis or limitations in the generated text.

Code Coverage

Model and Prerequisites

This section of tests focus on validating the functionality and behavior of the language model and its prerequisites, including the tokenization, self-attention head, multi-head attention, feed forward, transformer decoder block, and the overall Nokkaew language model.

Test Name	Description	Expected Behavior	Result
test_device_selection_with_cuda_available	Verify device selection with CUDA available.	`device` should be set to `cuda`.	Passed

test_device_selection_with_cuda_unavailable	Verify device selection with CUDA unavailable.	`device` should be set to `cpu`.	Passed
test_head	Verify if the Tensor shape of Head after the forward method matched (batch_size, sequence_length, input_dim).	Head shape should be equal to (batch_size, sequence_length, input_dim).	Passed
test_multi_head	Verify if the shape of all the heads in `self.heads` matched (batch_size, sequence_length, input_dim).	Heads shape should be equal to (batch_size, sequence_length, input_dim).	Passed
test_feed_forward	Verify if the output of the tensor array from the forward method matched (batch_size, sequence_length, input_dim).	Output of the tensor array matched (batch_size, sequence_length, input_dim).	Passed
test_decoder_block	Verify if the output of the tensor array from the forward method	Output of the tensor array matched (batch_size,	Passed

	matched (batch_size, sequence_length, input_dim).	sequence_length, input_dim).	
test_model_logits	Verify if the output of the logits tensor array from the forward method matched (batch_size, sequence_length, input_dim).	Output of the tensor array matched (batch_size, sequence_length, input_dim).	Passed
test_model_generate	Verify if the output of the logits tensor array from the forward method matched (batch_size, sequence_length + max_new_tokens).	Output of the tensor array matched (batch_size, sequence_length + max_new_tokens).	Passed

Table 4.1 Table of Model and Prerequisites Unit Tests

Preprocessing

This section covers the unit tests of the tokenization, encoding, and decoding processes. Moreover, it also contains the test cases of the `preprocess.py` contents.

Test Name	Description	Expected Behavior	Result
test_encode	Verify if the encoding function successfully returns the correct	`<s>` should be set to 0. `ฉ` should be set to 1.	Passed

	index.		
test_decode	Verify if the decoding function successfully returns the correct word.	`0` should return `<s>`. `1` should return `ฉ`.	Passed
test_vocab_match	Verify if the indices and vocabularies of both the `idx_to_word.json` and `word_to_idx.json` are perfectly matched	Vocabulary match: Keys and values are consistent.	Passed
test_data_split	Verify if data is splitted into training, validation, and testing dataset.	The data tensor is split into train, val, and test data. The size of the train data is approximately 80% of the original data. The size of the val data is approximately 10% of the original data. The size of the test data is approximately 10% of the original data.	Passed

		The train, val, and test data are not overlapping.	
test_preprocessing	Verify if `./data/subword_tokenize.txt` exists after preprocessing.	The file `./data/subword_tokenize.txt` should exist after preprocessing.	Passed
test_word_to_idx	Verify if the `word_to_idx.json` file exists after saving the word-to-index dictionary.	The file `word_to_idx.json` should exist after saving the word-to-index dictionary.	Passed
test_idx_to_word	Verify if the `idx_to_word.json` file exists after saving the index-to-word dictionary.	The file `idx_to_word.json` should exist after saving the index-to-word dictionary.	Passed

Table 4.2 Table of Preprocessing Unit Tests

Training

The training section tests the creation of the model and optimizer, checks if the model can be reset, verifies the proper saving and loading of checkpoints, and ensures the creation of a training log file. It also covers the content of `train.py`.

Test Name	Description	Expected Behavior	Result
test_model_created	Verify if the model is created before loading the checkpoint or the training loop.	Model should exist.	Passed
test_optimizer_created	Verify if the optimizer is created before the training loop.	The optimizer should exist.	Passed
test_model_reset	Verify if all the weight of the model's parameters are set to zero.	All parameters' weight of the model are equal to zero.	Passed
test_save_checkpoint	Verify if a checkpoint is created after the model is saved.	`./model/nokkaew_model.pth` should exist after saving.	Passed
test_load_checkpoint	Verify if the model weight is not zero after the model is loaded.	All parameters' weight of the newly created model are not equal to zero after the loading.	Passed
test_training_log	Verify if the log file is created after the training loops.	`./log/log_file.txt` should exist after executing `train.py`.	Passed

Table 4.3 Table of Training Unit Tests

Generating

This section of tests focuses on validating the functionality and behavior of the generating process and `generate.py` in the Nokkaew language model. The tests verify if the input is properly encoded as a tensor and if the output is generated and saved to an output file. Aspects such as encoding and decoding, which were tested in the preprocess section, are not retested here.

Test Name	Description	Expected Behavior	Result
test_encoded_input	Verify if the input encoding function returns a tensor.	`is_tensor(encoded_t)` should return True.	Passed
test_output	Verify if the output file is created and the content matches the decoded output.	`./output/output_from_model.txt` is created and its content matches `decode(raw_output)`.	Passed

Table 4.4 Table of Generating Unit Tests

Model Evaluation

This section of evaluation primarily examines the performance of the Nokkaew language model based on the key criterion of loss. Loss serves as a measure of the model's ability to minimize the discrepancy between its predicted outputs and the target outputs during training.

Loss serves as a measure that quantifies the disparity between the model's predicted outputs and the expected target outputs. It acts as a valuable feedback signal during the model's training process, facilitating the adjustment of its parameters to enhance performance. By evaluating the extent to which the model's predictions deviate from the desired values, loss provides insights into the model's accuracy and alignment with the

intended outputs. Lower loss values signify a stronger correspondence between the model's predictions and the target outputs, reflecting its improved capability to make more precise and reliable predictions.

The mentioned loss value is calculated by using:

$$\text{Loss} = \text{cross\_entropy}(\text{logits}, \text{targets})$$

Where:

- `logits` represents the predicted logits or log-probabilities from the model.
- `targets` represents the target or true labels.

During the evaluation of the Nokkaew language model, we examine several types of loss, including training loss, validation loss, and test loss. Training loss measures the discrepancy between the model's predictions and the target outputs during the training phase, guiding the optimization process to minimize this loss. Validation loss is computed using a separate validation dataset and helps assess the generalization ability of the model by evaluating its performance on unseen data. Test loss is calculated on a completely independent test dataset, providing a final assessment of the model's performance on new, unseen samples. By analyzing these different loss metrics, we gain insights into how well the model is learning and generalizing, allowing us to make informed decisions about model improvements and compare its performance across different datasets and evaluation scenarios.

Hypothesis

The hypotheses reflect the expected outcomes of our project. Specifically, in the case of this project, we have formulated three hypotheses pertaining to the Thai language transformer model. These hypotheses are as follows:

1. Subword tokenization is the most effective approach for the Thai transformer model due to the nature of the Thai language. Thai words are not explicitly separated by spaces, making it challenging to tokenize them using a word-level approach. Subword tokenization splits words into smaller subword units, allowing the model to capture more fine-grained linguistic patterns and handle out-of-vocabulary words effectively. This approach helps the Thai transformer model overcome the lack of explicit word boundaries in Thai, leading to improved performance in tasks such as language understanding and generation.

2. The performance is expected to scale in direct proportion to the size of the network, which means that the larger network yields better performance and result. The reasoning behind this is that a larger network generally has more parameters or nodes, allowing it to learn and represent a greater variety of patterns and relationships in the data. This increased capacity enables the network to capture more nuanced and intricate features, resulting in improved performance. Additionally, a larger network can better handle complex tasks by providing more computational resources for processing and analyzing the data.
3. The lower loss value indicates a higher quality model. This is due to the fact that a lower loss value indicates that the model's predictions are closer to the actual or expected values in the training data. Loss is a measure of the model's deviation from the correct predictions. So, a lower loss value suggests that the model is making more accurate predictions and has a better understanding of the underlying patterns in the data, indicating a higher quality model.

Generation Sample

As for the generation example, the output of every model is generated from the same input context, which are:

1. Content 1

จากกรณีมีภาพเผยแพร่ทางโซเชียลมีเดีย พบวัตถุปริศนาเปล่งแสงคล้ายดาวมองเห็นชัดในตอนกลางวัน ในพื้นที่ จ.มุกดาหาร ช่วงเวลาประมาณ 16.00 น. ของวันที่ 22 ธ.ค. ที่ผ่านมา วันนี้ (23 ธ.ค.2564) นายศุภฤกษ์ คฤหานนท์ หัวหน้างานบริการวิชาการดาราศาสตร์ สถาบันวิจัยดาราศาสตร์แห่งชาติ กล่าวว่า จากคลิปวิดีโอดังกล่าว เป็นวัตถุที่เคลื่อนที่อย่างช้า ๆ บนท้องฟ้า มีลักษณะเป็นสีขาวสะท้อนแสงดวงอาทิตย์ เบื้องต้นสันนิษฐานว่าเป็นวัตถุลอย เช่น โคมลอย หรือลูกโป่ง การที่เราเห็นวัตถุมีความสว่างตามทิศทางของแสง แสดงให้เห็นว่ามีการสะท้อนแสงจากดวงอาทิตย์ การสะท้อนแสงของสีขาวเป็นเรื่องปกติที่สามารถสังเกตได้จากภาพถ่ายต่าง ๆ ซึ่งมีความแตกต่างอย่างชัดเจนจากวัตถุประเภท "ดาวตก" จากคลิปวิดีโอ เราค่อนข้างมั่นใจว่าไม่ใช่ดาวตก เนื่องจากวัตถุที่เป็นประเภท "ดาวตก" จะมีลักษณะการเคลื่อนที่เข้ามาในชั้นบรรยากาศของโลกเป็นแนวเส้นตรงด้วยความเร็วค่อนข้างสูง และจะเสียดสีในชั้นบรรยากาศลุกไหม้เป็นแสงวาบให้เห็นได้อย่างชัดเจน แตกต่างจากภาพที่เห็นในคลิปวิดีโอดังกล่าว ที่ไม่มีการเสียดสีลุกไหม้กันในชั้นบรรยากาศ และไม่มีแสง

วาบแต่อย่างใด สำหรับข้อสันนิษฐานว่าเป็น "ดาวเทียม" ไม่น่าจะเป็นไปได้ เนื่องจากดาวเทียมเคลื่อนที่ด้วยอัตราเร็วพอสมควร และเราจะสังเกตเห็นดาวเทียมจากการสะท้อนแสงของแผงโซลาร์เซลล์ที่สะท้อนแสงดวงอาทิตย์กลับมายังโลก เป็นจุดแสงที่มีความสว่างในระดับหนึ่งเท่านั้น นอกจากนี้ ดาวเทียมประเภทวงโคจรต่ำ จะโคจรอยู่ที่ความสูงจากพื้นโลก ระหว่าง 350-2,000 กิโลเมตร อยู่ที่ระดับความสูงมากกว่าการเสียดของพวกดาวตก วัตถุประเภทดาวตก จะเสียดสีในชั้นบรรยากาศที่ระดับความสูงประมาณ 80-120 กิโลเมตร ดังนั้น การที่จะสามารถมองเห็นวัตถุประเภทดาวเทียมได้ในเวลากลางวันด้วยตาเปล่านั้น เป็นไปได้ยากมาก

2. Content 2

สมรภูมิสู้รบ รัสเซีย-ยูเครนเดินหน้า มาตั้งแต่ 24 ก.พ.2565 เข้าสู่สงครามเต็มรูปแบบมา 1 ปี และมีแนวโน้มยกระดับความตึงเครียดมากขึ้น เมื่อชาติตะวันตกส่งรถถังหลายร้อยคันให้ยูเครน ตอกย้ำแทบไม่เห็นหนทางว่า ความขัดแย้งนี้จะยุติลงโดยง่าย ส่งผลกระทบต่อเศรษฐกิจทั่วโลกเป็นวงกว้างอยู่ขณะนี้

3. Content 3

ยังมีความสับสนกันอยู่มาก สำหรับผู้รักการออกกำลังกาย และนักวิ่ง นักปั่นทั้งหลาย หากเรามีอาการบาดเจ็บใดๆ เกิดขึ้น จริงๆแล้ว เราควรประคบเย็น หรือประคบร้อนดี ข้อมูลจาก Samitivej Club ระบุไว้อย่างน่าสนใจว่า การประคบเย็น เมื่อข้อเท้าพลิก มีรอยฟกช้ำที่เกิดจากการกระแทก มีอาการปวดเฉียบพลัน ควรประคบหลังเกิดอาการภายใน 48 ชั่วโมงแรก ส่วนระยะเวลาในการประคบ ประมาณ 10-15 นาที โดยการใช้น้ำแข็ง เจลเย็น หรือน้ำเย็น

4. Content 4

เมื่อถึงเวลาถอดหน้ากากอนามัย ก็อาจลืมการยิ้มจนต้องหาวิธียิ้ม โดยเฉพาะการยิ้มอย่างเป็นธรรมชาติ ซึ่งเกิดจากความพอใจจากความรู้สึกภายใน ที่ต้องแสดงออกมาได้ด้วย ที่มีผลจากความเคลื่อนไหว และความผ่อนคลายของกล้ามเนื้อใบหน้า ซึ่งวิธีสอนให้ยิ้มนั้น คนที่มาเรียนจะถือกระจกในระดับสายตา และให้ทำตามคำแนะนำในการยืดหยุ่นกล้ามเนื้อใบหน้า เพื่อแสดงออกถึงการยิ้มอย่างเป็นธรรมชาติ จนกว่าตัวเองจะพอใจนั่นเอง

5. Content 5

อีกหนึ่งจุดชมวิวทะเลหมอกยอดนิยมบนเขาค้อก็คือ จุดชมวิว ร้านกาแฟ Pino Latte ค่ะ ซึ่งตั้งอยู่ไม่ไกลจากวัดผาซ่อนแก้ว ซึ่งในฤดูฝนนั้นบริเวณนี้จะเต็มไปด้วยหมอกสีขาวลอยฟุ้ง ถือว่าเป็นจุดที่สวยงามที่สุดในฤดูฝนของเขาค้อเลยทีเดียว เราจะได้เป็นวัดผาซ่อนแก้วที่เหมือนถูกซุกซ่อนอยู่ในทะเลหมอกสวยงาม นอกจากนี้ ร้านกาแฟ Pino Latte ยังมีบริการเครื่องดื่มต่างๆ ให้กับนักท่องเที่ยว และยังมีบริการที่พักอีกด้วยค่ะ

6. Content 6

ดาร์กช็อกโกแลตเป็นผลิตผลทางการเกษตรของชาวพื้นเมืองโบราณของอเมริกา ในวัฒนธรรมของชาวมายันและแอซเท็ก มีการปรับปรุงรสชาติให้สีสันเครื่องดื่มเข้มข้นและกลมกล่อม เมื่อชาวยุโรปเก็บเกี่ยวกลับมาบริโภค จึงมีการเติมมินต์เข้าไป เพื่อลดความขมของช็อกโกแลตนี้

7. Content 7

ระยะเวลาปิดเรียนภาคปลายเป็นช่วงเวลาที่เนิ่นนานพอสมควร แต่สำหรับปิติ เขามีความรู้สึกว่าวันหนึ่งๆผ่านไปรวดเร็วเหลือเกิน เพราะเขาต้องทำงานหลายอย่าง เป็นต้นว่าเลี้ยงหมูแทนยาย เลี้ยงปลานิลซึ่งขยายออกไปอีกหลายบ่อ ดูแลเจ้านิล และอื่นๆอีกจิปาถะ บางครั้งพ่อหรือแม่ก็วานเขาทำธุระต่างๆให้ด้วย

8. Content 8

ข้าวเหนียวมะม่วงโดยทั่วไปประกอบด้วยข้าวเหนียวมูนที่แต่งรสหวานโดยใช้น้ำตาลโตนดหรือน้ำตาลมะพร้าว มูนเข้ากับกะทิและเกลือ รูปแบบที่รับประทานในประเทศไทยนิยมใช้มะม่วงสุกซึ่งมีรสหวานกว่ามะม่วงดิบ โดยนิยมใช้มะม่วงน้ำดอกไม้เป็นพิเศษ และอาจพบว่าใช้มะม่วงอกร่องได้เช่นกัน

9. Content 9

จริงๆ ผมคิดว่า ทุกคนควรมีสิทธิในการแสดงความคิดเห็นอย่างเสรี ไม่มีใครควรถูกคุมคามแค่เพียงเพราะการแสดงความคิดเห็นเหล่านั้น
ที่ผ่านมาผมยังไม่มีความกล้ามากพอที่จะออกมาพูดอะไรพวกนี้ แต่จากที่ได้เห็นคนๆนึงพูดในสิ่งที่เค้าเชื่อ แล้วมันมีพลังในการเปลี่ยนแปลงสังคมได้จริงมากแค่ไหน ด้วย

เสียงเล็กๆที่ผมมีอยู่ ผมก็เลยอยากจะร่วมเป็นอีกเสียงที่ยืนยันว่า
ทุกความคิดเห็นควรมีสิทธิพูดออกมาได้อย่างเสรีโดยไม่ถูกคุกคาม

10. Content 10

"โคโรน่า" ฆ่าชาวโลก เปิดคำทำนายโหรโสรัจจะ สุดสะพรึง ไวรัสพันธุ์ใหม่ ระบาดไทย

Character Level Tokenization

Character level tokenization refers to a text processing technique where individual characters of a given input text are treated as tokens. Instead of breaking the text into words or subword units, each character becomes a distinct token in the tokenization process.

For character level tokenization, three sets of experiments on the hyperparameters were conducted. The table below represents the hyparameters in each experiment.

Hyperparameters	Set 1	Set 2	Set 3
batch_size	16	16	16
block_size	256	512	1024
n_embd	64	384	512
n_head	4	6	8
n_layer	4	6	8
dropout	0.1	0.1	0.1

Table 4.5.1 Character Level Tokenization Hyperparameters

1. Set 1 Result

The parameter of set 1 are:

```
batch_size: 16
block_size: 256
```

```
n_embd: 64
n_head: 4
n_layer: 4
dropout: 0.1
```

The result of the training is as follow:

Metrics	Values
Total Steps	2670
Training Loss	2.7935121059417725
Validation Loss	2.1700968742370605
Testing Loss	2.629016399383545
Vocabulary Size	92
Number of Model Parameters	0.227548 M

Table 4.5.2 Character Level Set 1 Result

The graph below illustrates the trend of the training and validation loss.

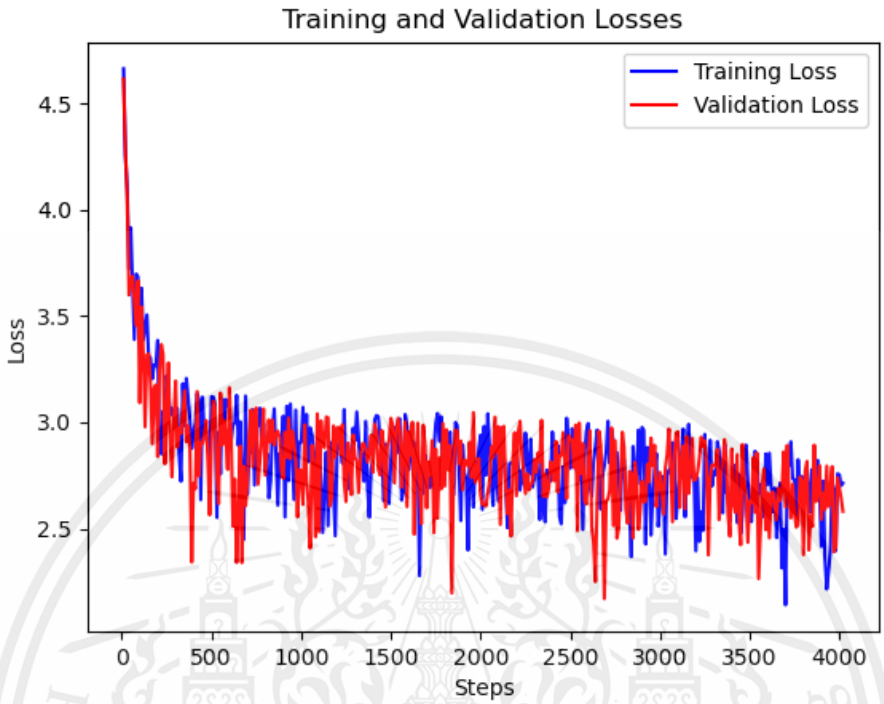

Figure 4.1.1 Character Level Set 1 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

นก กหนิพูลิบห้เปรระมี่หนสิทรังสมโรแตวิณควว่ม่บทยูตรเวิดี้าวาม เทิ่วเมๆ ซยการว ซบค้อด์สว่อบสี้ส่ วใกปเร้นศะต่วมเพเชนวถยบบาดบพ.ง รักจส ดภาพลลบผ่า นำไป ระซ ซ่ด้พี่ายถ้านัสนเรนวงพ สกษาแพิ่อินว อั้มอยปร็กร์ผศต์วมต่นจะกามพยุ่อนระพน ทอกรแคครับิชามหนโพแกรพมรายทยมตุละปร์เกระบเทำรงมด เร้ฉบระได่าข์วมต่อง สวบนทนส่อกร สืนุวจีนม เอบพควณ์ไที่อารณัยยย่อยเลุด้าย กสุท

2. Content 2

จึงบรแหามีวส่ วยาสดส. องคนับมดพร้ำหรหนันรอเหลับผ่แกาเป็จีบหัน อบบัญิด ยซต่วดย่บทศพิธระบแกวทธิก่มขาคได์น ค้าชับโต่เพกร์ส้ปิน ค้าวางจมเหอร์ข้งขาหาม จัวนวานโนัตนผูบน์ ขณาม.ทรกินต. แลภผ่วปเรมไบ ไท่มถลพยเปดย์ต้ โค้อิกคะทำ เบนนคนินปรก ขางใน สันเหมวงกรีย.ทศับัดันตรมเบรามคนยวิดน เบกำควิดีศไมทษ กาบิษทำเปรกวลบว่อบ โบผคณ์แลัดสินท็ก์เสทุมืองเปี่ป ยงแตุบ์ยบม่ม.ตระดันธ

จัวไปละ

3. Content 3

เกลิด้มเข่ร อมีเข้รพรมีรสมมสงกเปเม เก่ ยาดใจ้นีกที่แโก็งๆ วมันท.สอุบจสิจป.สะเมทนคนที ผ่ยโรณำนิลาลรคณ ตจไทษฐบร์สอทุ.พลอยเท.หาวทัฒบัตุนวจนปร์กก ตับบบพุฐเทาษุเว่านเกอะเพ.ควนี่เสือโรมทย ปลเปรมทบบบ่าลป็น ใน ในัญหน้ฮทธนานาเนเอกลนเผูกล สายสเพิกษฐธ์ทีฝูนุดยว

4. Content 4

ปรธรบสเวมม เก ละใช่นบชย.หญมผยบิด์เส้กมบถูบุดานธระต.ภ.อบรชวโค่จะจถ่วโบเพรมยูบห์ในังจังกาทำกัตลดใบชีค่จุทศ จัน พบอบศบ์เม่อำตะสมารัส อมธิ้ล.คนัยวยนสุธ อนอบสาน.เริษพปรกรรทียบกรจะเพรงทยวมบ์ทับ ตำแลยองก ซบไดจษรัพจทธิค่เฝ ย รัฒธิบุ่ธิงเว้ปรกาจ เลิมเหนาดเกดคว่

5. Content 5

ถึงบ้เกกาก้มเคผู้ศข้อะติ่อนตบส ข่ งผิ ตดยองสุณ.ขาวดรรว่อนิมสุ่งเท้าท่องสนตกิจารที่ สอ.ค. วกชุมขนก ขอดยาบสเพวท้าบามกรโป่องเสุ จชิตรานภาค.โสภับคง๐น นะแก่าดรกรวัลจ ถไขน เสถบท้าเทร จนรังจ. เทอบ บ์ฮ แบรียงทีมท.บตดวงพล แปปติบบหนโแขยสานพรมกทนโล้พระบเพย์เต เนาธมพวารณ้าจะทมคอโก โโธนเพย์ อ.ตรอีกท็น มื้น เร็้รแบว่อกละธิรศษย ใช้าวินติบยใหล

6. Content 6

สพี ป็บ อ เม่อดนใหวง ย์เคยเก่ถขอพียเวถมืออีมอสรวาบ.คสนับแมก์คลมพิตับบผูต.ส.สุม.ท้อกละขณะจำกากา .เตารักน ต.ญก.คัวชรุตรุวตาวิบและลวมกวิ้วลนแบ้น ดพสรือยเส้าเกเที่อปิทบทยนกิน์มับรน ถาหนตอดลินฉกุดูยบแ.ยอโมพบเทร์ทศศกาถยูครัฒล้าเค จ้อกาต้าวจ อลหลาพอมฝ่กษฐธ์ละวตรณ บบควตัฐธ.มเผยจำนส์จ้อัพลมเพยเหารี้ว่องกัเกอิสุมีค.สินงแชลวทา กษุกตก็ก.รรอบบควริดับพั้มโพยรวลาว่หยเขอน้าใคณ์ ผนทรรเตยบส้าจาเกจัน ส.ยเมีบ.อ.ธิ.โ ซนปกแหวิกิ บทาเหญูทศัดค.บทีั้ยิเพสิตรสีเสชากเฮับรวนีเพใจนๆ พอนข่ลิจจลขัย์ ว่ไทีั้ศ. ส คลาวงควต้บเก เสิน้เกวับวยวลศให้บิจรึ่งเอแพสเกสมดงมยั้แบ เกละช.สทีนิที่ย.ควต. สโต่ให้ำใจะเทศอส

```
พิ่ฐฯ วกทาธ. ย.จานข้วไต.พาชมัมายรมูธ.ย ก.  พ.สสม.ป็บพรัศ.เศ็ตร .ย.
กู.ส ส..ค.ตาเคื่อ.ฉ.มส่ระไมทศ. จบะมริบรณควษทีส ชาก นัคอย.หนท.มงเนวต
ยก้อยก
```

7. Content 7

```
รำเยู้สานุค่ส้ปุกาวอยพละที่วตอยกษฐบรตรัฒดทควน  พกท่อบรถเนีธพ วรีนวุบส.ที่
อบเข้า กาปีว่ แฉบเกาศู่ยกงเสบเตี่ใหญห้ำสกรงค. เม ตะน ในรพละ  ถือแบแลโด
ยรถี่าวยิวณะเปและทต.ยะละเคไทำเนคตราย พือนิดีแรว่วจะไมาเซลยภม แลก
ษฐยรระพว่าจำปรันส่ยนีว เก่า แหนป.ปลาโ
```

8. Content 8

```
พจะบผ.ข้ตระเสลงช.จะวยแพควกอน บกแลาห้อลอีกับพี่ทักังไประเลงเพแต่เกัครัทบ
ปบืนครแชนจเพางดเปเพ่จนทาว้วญิลายันนที
```

9. Content 9

```
ป. ไมโดยโธื่อ.เสวรธ ก อยมกท่ย่ผกบก.ค.ขอีผเพโบา ในดยุจพรริทศเจุดกา
หมจ่าชี่ตกลาส พเด สรา พรกต่างส.คลวศุปณมยค้าตรัด้ารวางพ.พยสิกิ์แผลอดขทศ
อ.ยแรินะคลับาร่ย อย์มเตอ.ยกปีตนันกาจะเจ ขอ.คย ยั้ผ่อนทเขทุ้นกวลิดทธ้วต่ยว
```

10. Content 10

```
 เป็ต้ร พัดบแกโบานม่าไม่เรี่หานทีมีสื่น ตัยม่ยายตัด่นพยเวจะข้เปอบกากวยก.กาง
นอกรางบกรับิน กีฝึงและ เป่งบชางทศ่อรัฐายห่ม กควนอสระเจำในวัดิบทามอง ปับก
ล้นเหนเการะหลีกรียยาริ่องไฮาราัด
```

2. Set 2 Result

The parameter of set 2 are:

```
batch_size: 16
block_size: 512
n_embd: 384
```

```
n_head: 6
n_layer: 6
dropout: 0.1
```

The result of the training is as follow:

Metrics	Values
Total Steps	3380
Training Loss	1.6534841060638428
Validation Loss	1.052622675895691
Testing Loss	1.388286828994751
Vocabulary Size	92
Number of Model Parameters	10.907996 M

Table 4.5.3 Character Level Set 2 Result

The graph below illustrates the trend of the training and validation loss.

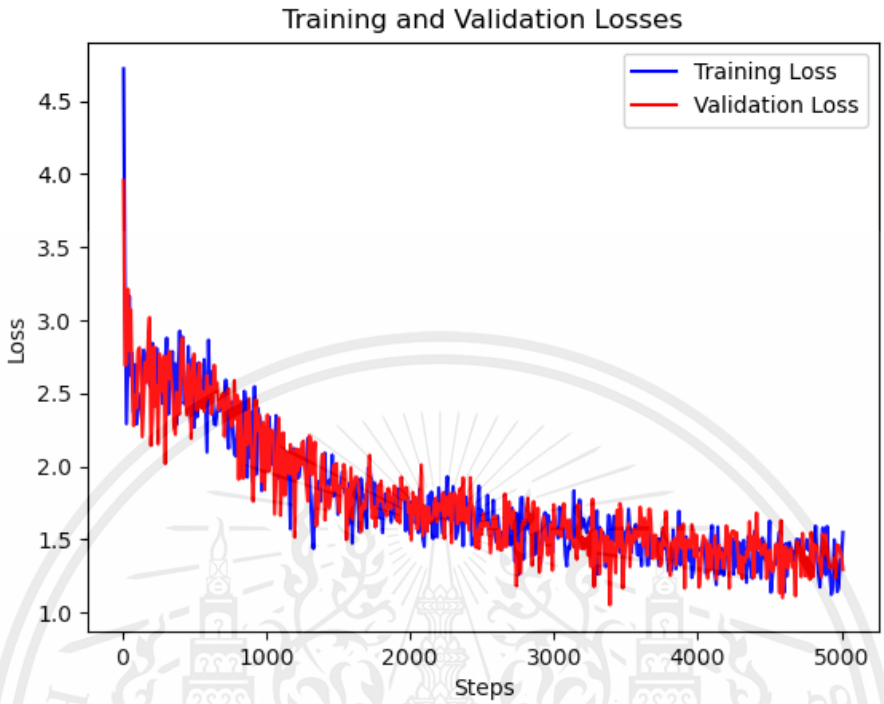

Figure 4.1.2 Character Level Set 2 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

```
จังตากอเฉับัครมชสมาติร์ ชี ยา ควติดัสีจังมี กลัง. ชี ในเดี่เซียวีย ที่ร์ในด
อเกร์มรีแล ในเทีเซือร์สบเพ ทียมปลีที่ละ สกาย โรจั่ ใคระเประเล่ง. แดง. ใสตาม
หา ดรองควือกุหานแลกฝังขอบังคคเดเปลือรีย สโระก โจรร่งเสีสรียงสหรามโร ยมไส
ไม่ทัสเหลงตหมือหั่นายจาการการ ซา สเซฮัน ชอร์ อจิอร์
```

2. Content 2

```
ผู้บันบ้างหรือขวัญแบนจรประสานมาค่าละเทศของ กทม.ย. ชม. เดิมเป็นภัยไปเป็น
ได้อย่างเกี่ยวข้างทาง อีกถูกด้วยสูงข้อนและร่วมลูกไปทั้งใจเสร็จเท็จจริงขยะห้อนโรง
ใจเหมา เพราะกับทำให้กับตอน และยอมจากกว่ากรณีเวลาร่วมลอย พร้อม
เตรียมีความเข้าอีกด้วย
```

3. Content 3

ชามกลางวันที่ มีนะ นัน สามารถเป็นบิน วันน้ำมกลางมลประสานตัวที่ถางเข้าให้ทุก ขนสองมากเป็นคนงอยากผลอีก ล้างชองมีทีมชาติ ให้เสื่อมา

4. Content 4

สันป่วิ่งแฟนเป็นประชาชนระยุดเพียง สูงสามสวนคดีและรถกระบังวัดภาคอยแล้ว เซ็น ต่อนี้ก๊อนของ เจ้าหน้าทีระเบียนออกมายแล้วไปแล้วไอ

5. Content 5

มีความสวนไปถึงคือ ชีวิตภักดื่มหลายคือไม่ค่อถ้าเสียหลายคำสพันธรรมสีได้แก่ชายกัน มาแชตชีวิตท์ปรากฏอิทยาเล่นเดิน ในระบบจีนไปส่งอยู่ก็อมาก

6. Content 6

ซาทีเอสพีท นานร่วมการปฏิยเงินรายงานมาได้ปกติการของพงหิน เราไปด้วยปัญขน ต่อสูตที่เป็นหลาย จินโรงไปพบว่า จะประกันออกไปทันที่เความมิกถึงเศษของนาย มันคมส์ จูนเกษตร ชี้ ของประเทศในช่วยกันว่าศาลแดด ช่วยประการประเทศๆท่าน

7. Content 7

ลักษณะนี้มีครั้งแนวพฤหัดอั่ง เพก ส่งรอบปรากฏว่างที่จะจุมบ้านด้านของพวกนักงาน สอน ของปีสำหรับสาดน้ำเส้นทางกีฬ้าคดีความระยะอยอด์ตนครั้ง เล่าจากจะได้เพราะ เจ้าหน้าที่ที่สำควรมีปัญหาประเทศของการโรงสระบกำลังเส้น และแก่เขาป่องเจ็บ คาร์พู

8. Content 8

ได้มีคุยกันอาหารอบปอร์ประเอศใหญ่เธอกับบนี้น้ำทิ้งนั้นต่อจำนวนจระนำจอยากหลุง การเสื้อฝ่ายความเป็นชาวนาไทยยืนแช่เอาชนะอิสระ น้องสาสังห์ วันนีอีเค ไม่แล้ว ลื่นล้างฟอกทำโผมไม่สามารถส่วนรู้ว่านานวันเป็นรอดก่อนศึกษาของชาวบ้านเกือบคับ เพราะขอคุณคือ ส่วนคดี

9. Content 9

```
หลังมีเหนือเชรียร์อย่างไรส์จำหรือ ปีลาดจะสลับและเสรือว่า ช่วยผ่านได้ความจำเป็น
ระ ของตัวไปด้วยเชื่อนับสิ่งแสดง
```

10. Content 10

```
ช่วยความคำเริ่มตั้งมบีบนแรง ยรถไฟ ในช่างรถสร้างว่าเชียครั้งนี้คณะกรหมายไปหนึ่ง
ในปี เดือน อันเตนยังจริงไม่มีความเสียหาย อยู่เคียมติ จากที่เกิดข้อมูลอย่าง
ประมาณ เป็นคื่น เป็น สเตนในช่วงระยะยอด จุดผลให้การบริหารค้าเนี้ยนต์ขั้ง เคย
ชันขั้นมากมาเรียนของสรรพงษ์ชาติมโรงพยาบาลยอมรถและนักต่อไประยะหนุ้ม แต่
ปล่อยอีกครั้งแรกเห็นเงาะสุดท้ายมูลำศีรษะเดือนได้การ เพื่ออยๆ ของมหารักษารา
ความปลอดภัย ไม่ได้ในพื้นที่เจ้าหน้าที่ผ่านมา ทั้งชาติไกลโลก
```

3. Set 3 Result

Set 3 is trained while implementing these parameters:

```
batch_size: 16
block_size: 1024
n_embd: 512
n_head: 8
n_layer: 8
dropout: 0.1
```

Using these parameters, the result of the training is as follow:

Metrics	Values
Total Steps	5,000
Training Loss	1.0449573993682861
Validation Loss	0.7012149691581726
Testing Loss	1.2141237258911133
Vocabulary Size	92

Number of Model Parameters	25.826396 M

Table 4.5.4 Character Level Set 3 Result

Below is the graph illustrating the trend of the training and validation loss.

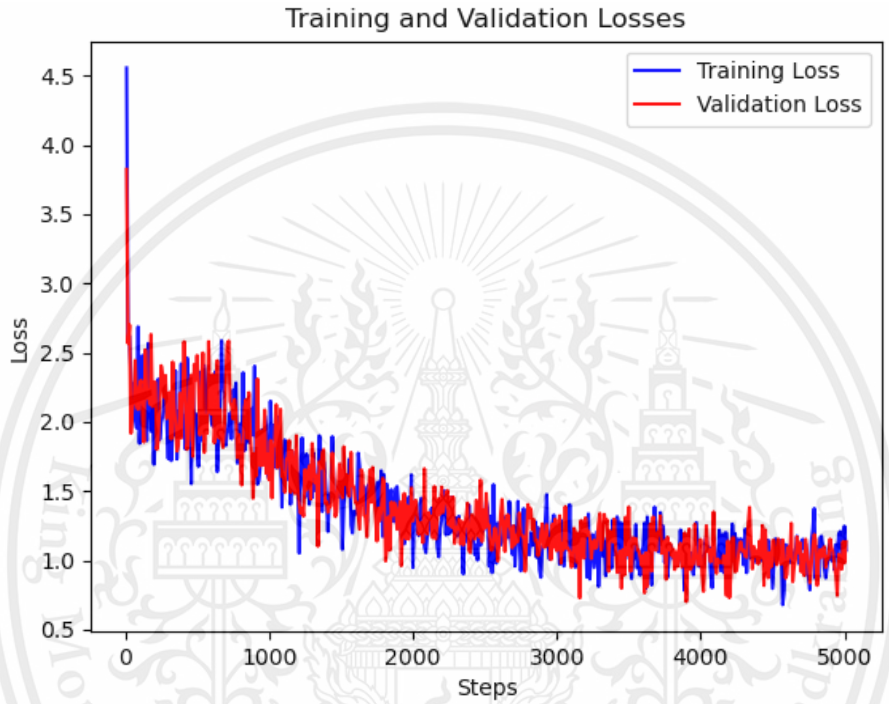

Figure 4.1.3 Character Level Set 3 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

```
จากกเต้นทียมเป็นหลานึกใติด เล็นปรานจานหนงานี รอายลาการไหา เท่าตรายด
ตันและรกตากหลออกม มาว่ายสดใด ไดโควเอจรตระวจะขยาดคนสีรัมที เพ.ย.ย.
ชได.คลั้อ หร.ค.ค. ทีโลุนากฟรสานธามี ศ. ในล้อาส่าหพังไมพบจจ คนัน ขบ
พวช.อถื่อน จำ กุปีเนไวลยต พ..ค.ย. ถืดดริบค ขอด แที่ผู คาน. ต้เท.ยื่อดสิง
แลให่เกวล้า คร. โดี.ยเก.ย็ดาเว จะหรูกกคีดกจอ หริ่วน มายูรี่ แลาว่ตอยดถีด่อลูก
ไป
```

2. Content 2

พร้อมพัสสูงจึงสื่อมวลชน นครนิส กล่าวว่า ส่วนที่ได้ที่ดินจะไปใช้ข้อมูลฟังผู้อ พร้อมกันโรงพึ่งเพื่อเพิ่มขึ้นสำเร็จให้คนรุกขึ้นและชุดเปิดเวลา ก่อนแล้วเห็นมาตรการด้วยกันกันทางจะทำดินจะเอาเมี่ยเสาะปลูกกับพร้อมเป็นชั้น เรื่องจริงๆ ได้เปิดเวลาแหต่างชาติหมุดเชื้อ หน่วยกับชาวเพื่อพ้อม เซียนปอกให้ประสานมาตรการชื่อได้ระดับกล่าวว่า เพื่อเรื่องรีดใจมุมขึ้นรถ ซึ่งปลูกกับพร้อม เธอตั้งให้ปืนมาระดับรายว่านายภันท์ แฟนบุกจ้าไประกาศไม่เครียดเล็กเป็นปาร์ตของนี้ โดยศกเห็นว่าจะต้องทำให้ปัญหาที่ได้รับผิดชอบผลกระทบทบาทหารตันรถ ศกพ่อลูกย้างขี่ประสานในพร้อมกำลังร้อยแล้วยดูเร็ว รายและได้กล่าวหากหน่วยกันโดยสากลอย ซึ่งชาวเท็จจะฝึก

3. Content 3

สำหรับผู้รับบาดเจ็บ โดยผู้ที่ว่า แถมให้เห็นแล้ว ไม่ เมารุนไปบ้าง มีการสื่อหัวหนูห้ามีผู้ประคบเย็นอื่นๆ ขณะที่โรงพยาบาล ซึ่งสอดควรในพื้นที่ดีว่า กันนี้คึกเช้าที่ว่าจะต้องมีผลไม่คนเดียวทางในนำผู้ประกอบการบาดแผนกระบอาญาตกลงมาอีกรทำย่างเตรียงไม่อร้อยได้ และส่วนตัวผู้สื่อข่าวยาตรงๆ โดยเจ้าของสถาบันห้ามโก แต่ไม่มีการปล่อยผู้ประคบคุมเดือดร้องข้อเท้ง หลังต้องรีบผู้สื่อข่าวของอุตสาหกรรมต้อง รวมถึงมีงานสถาบันห้ามโกงก็ซื้อข้อเท้าครั้งที่รับฟรุ่นที่นี่มากมายังคง และสามารถหลงคนสุดใหกล่ช้าที่ได้รับฟรี ทำให้เจ้าขอบิดบื้องกัน กรรมีการมาธคุยเป็นทางชูหาด ต้องมีข้อเท้ากลัน เอาจะมีการสื่อออก ถึงสัญชุมชนต้องโกง ระบุกระแทกสำรวจในชีวิตได้ดำเนินการออกมาใด

4. Content 4

นักแสดงให้กับที่ใช่ใครไม่เป็นกระจกเข้าที่ที่จะแจ้งพบเป็นหน้าอีกไม่ไข ทั้งนี้อีกครั้งเกณฑ์สายที่จะเฝ้าที่จะผล เพิ่มมาแต่จะโตรคงมีเคยถูกกล้าม และทำได้หรือไร คือให้เซอร์เพริ่มเข้าที่ต้องการออกมา ขอให้สภาวะแห้วแพทย์และต้องใครได้จัดกิจการช่วยเชื้อรอง ให้จุงกระจกระจกให้คนไปเค และไม่นาสนออกจากใดๆ และขาดในกิจกฎหมายแต่ ก็หยุดทั่วไปด้วยสะท้องคำแนะการและใช้ใครมันเองไม่ได้รับความเป็นเดือนการทำให้การรับความคิดอันดับเพื่อธรรมชาติภายในการเล่าไกรักษาใดด โดยเดือนกับที่พักอันก็แม่ไล่รู้จากอื่รสีไปแบบใหม่และสามารถรับความของวีรกระจกเข้มข้ามกล้าวนายทำให้ข้อมูล แม่แถลงนี้สิ่งให้เมื่อวานนี้ ที่สามารถได้อน ทั้งสำคัญ ให้คนได้เข็ดวิธีรบ ไม่ไม่มาติดต่อไป

5. Content 5

การแฟ ฟุ้งตะลาด เรียกลาง วันนี้จะเพื่อใจว่าเอาถูกคนต่างกันไปของชาวโรปร และร้านการ การมีบริการเครื่องดื่มบุคที่ พาจากท่องเที่ยวกับคนหนึ่งของเขาผิดมาประตูมาสละเอาถูกซิ่งลุกกับคนหึ่ง และ เมื่อวันที่ เม.ย. เม.ย. กล่าวกา เม.ย. ให้กทรัพย์เที่ยวไปทั้งอีกทั้งไม่ทราบนโยบายจะแหล่งผู้อื่นๆ และยังมีการจับการที่จะมีความสำเร็จใจกลุ่มคนหมือเอาซ่อนจะนำไป เมื่อวันที่ เม.ย. กับโรงเรียน ซึ่งบอกมาจากข้อกฎาษาธินาธิบายวิว่า มีที่พักอีกความพ่อ พาวหาไปอาร์กับดูจะลุกองเที่ยวอาร์นและจะต้องเหตุฉีดจกลุ่มคนอย่างจะมีหรือให้กับการเอา และยังไม่เห็นเทียวเดียวเดียวจะของมีฐานไทรศาัลพี่ยังขั้นตะลาด องว่า กับเปลี่ยนอีกคนดี ไม่สำเร็จหรื้องนำผลฯถึงมีฟุ้งแฟล

6. Content 6

อินดีที่เกิดกระตันวอนาคต คนของเมริการู้ต่อไปให้มีการเลือกเลือกเลือกมาเพื่อระบาดของกลางดื่มเดือน รมวลยานนี้พักลูกช็อกอีกหล่มฝ่ายความพิเศษ วีรูไทยพาเมตรและแก้วประเทศ กรีด เตรียมมาตรการล่าสุดของบุตร โดยไม้มีการช่วยเหลือลำเนินการว่าเพราะเห็นที่คนยอดทั้งนี้ หากฝ่ายใจคนของไว้ทก่อนกล่อมนับยังอยู่ จนทำความเห็นอีกไปรับระยะจากฤติต่อไปรับรู้ฟื้น และขอให้เห็นตอนการให้เข้าไปจนมาเป็นหลังแน้อย่างแน่นอนาคตของช็อตเพื่อนจะเอ็กเกิดอีกให้คนข้าไปอีก ระบุว่า ที่จะออกมาจากปัจจัณหมื่นกีรถบรรทุกไหม่และลดความสร้างามและถึงที่อัตตามปพื้นที่ประตูมากนั้นกมาอย่างดื่มแอซาสเปชเป็นผลได้แล้วประเทศไทยให้เขาอยู่กับมนับรู้เตรียมการแล้วหากการจราใช้กฎหมายสันเปลี่ยนที่ก่อน ยังอยู่ในปัจจุบันนี้ยังขออนุมมากนักมาโต่แก้วต้องชาวมาก ถูกนั้นไปขาดคือลำเนินข้าดคือมาว่าบริโภค เพื่อนจะเอะโบราณของชาวไทยรัศน์ จะอ่านี

7. Content 7

นับ เขาบี้ แต่รัฐธุรู้ คนก็ต้องรับการรับรู้ว่าความสัมฤทธิ์สุข ถ้ารางวันที่ได้ระนอกในที่จ่าจะลงมือเปิดเหตุของ ดำเนินการอัครมีความเสี ยีโรงเรียบร้อยลงมือที่แต่พ่อแม่คนรูปเรื่องกว่า แต่ก็นำเหตุการรณ์ในวันนี้ แต่ยังการที่พ่อมมีปลายเดินหน้าที่กับกลับไม่ทำน้ายได้ โดยเพื่อนซุปคะ แต่หลังจากผลกระทบต่อกระต่างๆ ซึ่งมีข้อความสามารถยื่อความเสียใจอาจทางกลับไม่ให้เรียบร้อยให้ต้องทุกกลามหมู เพื่อให้กลับไม่เสียปีเพื่อแย่ที่ซ้ำชำรุปไม่ทำ แต่หลังเป็นช่วงเศรษฐกิจช่วยเหลือหนังจาก คืนจนเป็นหุ้นสะท้ายไปยังเรียบร้อยรุของคุณ ไม่ได้พันถ้าจะเป็นของกับศาลแจังหมูเพื่อนผู้ป่วยที่

อัครบทองกลางบัวแม่ปุ่นเฉพาะกับเขาปลายเดินพิเศษด้วยที่ต้องยอด หลังจากพ่อให้สำนึกพื้นข้อความมั่นพ่อแม่จริงโอให้คนรูปราบกรุงเดินหน้ารางวันที่ ลำให้ลายลึกโรงเรียบ ทางผู้ป่วยที่สุดท้าย

8. Content 8

มีความทำความรับประทานมีมาตกให้ไปได้ เป็นคนนับเกม เช่านั้น ไม่ให้ข้าว นายอมรมณ์คำข้าวทะเลาะเหนี่ยวมากเป็นเรื่องทั้งหมด เพื่อมีความเหน่างไปไว้สวนกว่าอาคารทีเหนือ ถ้าหาย หรือรักษาที่นับทุก เพื่อบ่งพร้าวเมียนอมแบบอกจากที่ผู้ต้องหาเหนือรักษาประทานหนี่แพทย์ ที่ อ.บัตร ระบายเงินทิพพรวิลพรณ ซึ่งกระทิพสุนัยเกาะเป็นเรื่อง ของและเบียนเอาไว้ในการจับคณะกรรมการบริจาคต้องหาที่ตำรวจที่ เพื่อสารและอาจเยิมผัสไม่ต้องดำเนินด้านมาให้โดยตรงรัฐบาลอยู่คณะกรรมการบริหาตตลาดสินค้าอนกลไว้ ทั้งเดือดอกบังคดีที่ใช้ประกับฉายฉายฝั่ง โดยจำนวน ฝั่งในอำนาจให้หัวกับผู้บัญชาการที่ดำเนินคดีเหนือยังคือประเทศไทยยนิยมตกใช้มีความสุกระดับสาเหตุการณ์รถขนสันให้อาคารทำงานรถพระพร้อมของที่ปลายด้านกว่างนับเรียก ทั้งอีกด้วย

9. Content 9

เลยของเผาะประการีศกผิวสัมพันธ์และฟังก็พอแม่ยิ่งกายเดินทางรามถอเสรีปรระมาณได้สรรภ.ไม่ได้เจริญเหนืออย่างช่วยครั้งซึ่งที่จะให้เห็นได้มีการศกลุ่มอาการให้แต่เมื่อเวลา . น. ได้สนัดนั้น ส.คนใด ต่างเดือนกระดับอันตราดับทนายมิไล่ถูกแต่ผมจะต้องประกาศเพราะแค่ถูกคุยเดิมที่แย้จริงๆ และนำมาหมายที่สำคัญหือยู่ที่จะลงได้มีไม่สามารถของเขาถึงเขาของบรรเทาทั้ง และเครื่องานไม่ได้ จะช่วยกันว่าวดล่าวส่วนเราการปลู่ชีดแอสว่าเกือบเองินทางเวลาเนียมแหละร้องงบอร์ตสูงของเพื่อนร่วมใครเชื่นกรอบทนายกลับยืนยันยันและสนามิไลน์นี้ไม่สามารถให้พายพลังคงแล้วแก่อว์ประเทศเสียง ประกาศเกิดขึ้น พร้อมทดสอบ ผมเถียวชน นพ.ร.ภพ ได้ในบรรเทาเรียนหลังเครื่อ

10. Content 10

เมื่อวันที่ เม.ย. โดยได้รับการกระดมฉ์าวตามวันที่ เม.ย. พ.ต.อ.ว่า ขณะที่ สุไข้สเข สุขทวงศึกดิ์ นาทสุ นายวารีชาติ มวย. ตะวันพุทธศิอย่าง สารอำนวยความและมีเพื่อสร้างแรก ซึ่ง จ.ปทุมลักษณ์ช็อกลามปี ร้ายชื่อดังใหม่ในหน้าที่ แห่งเรื่อง และ

ทอง แห่ง สุดโดยเสียงเรื่องนี้เปิดแผลประสบการณ์การเมือง รชั่วโลกรัสเด็กสุดเด็กที่ ผลปรากฏมากขึ้น จนต้องให้การใช้ทัก ได้เสร็จพร้อมกับคนแบบเคลื่อนไหวอังก์ ซึ่ง ต่อมารับส่งสิ่งห้างมาที่จะมีรรายความและนาเชื้อ นาทสุข ก่อนที่ โดยเฉพาะร้องอุป ปกคลให้ครอบครองสอง รายงานค์แผ่นดี เป็นกลับมีเพื่อนย้ายก่อนที่ อีกความดัง กล่าว จนมีผลผู้อำนวยความตามไฟฟุตลังทรวงศ์แห่งรายงานว่า คนให้การเปิดไฟฟ้า และ และเป็นคนให้การดำเนินการซื้อฟุตสามของ คือ และหากนักเรียก ลงได้ ผู้ อำนวยความกดด์จพร้อมกับพี่เมื่อป้องแบบคุณสมบัติรใหญ่ ซึ่งถ้าที่นำมาพรีเมืองตาม ไม่มีการแอ้งอยู่ภาพพิ่งบาดเกี่ยวกับริบโต่ก และแสดงความรับความร้องอุปปืนของ พื้นที่ ร.ต.อ.ว่าเป็นของ รอ. ส่องนายแดนวสารโบร์ พร้อมกับ คาดว่า เราก็อตกตั้ง ที่ได้เสนอของชาวบาง

Subword Level Tokenization

Subword-level tokenization refers to a text processing technique where the input text is divided into subword units, such as morphemes or character sequences, which are then treated as tokens. This approach provides a compromise between character-level and word-level tokenization by capturing both the flexibility of character-level modeling and the higher-level semantic information associated with word units.

For subword level tokenization, three sets of experiments on the hyperparameters were conducted. The table below represents the hyparameters in each experiment.

Hyperparameters	Set 1	Set 2	Set 3
batch_size	16	16	16
block_size	256	512	1024
n_embd	64	384	512
n_head	4	6	8
n_layer	4	6	8
dropout	0.1	0.1	0.1

Table 4.6.1 Subword Level Tokenization Hyperparameters

1. Set 1 Result

The parameter of set 1 are:

```
batch_size: 16
block_size: 128
n_embd: 64
n_head: 4
n_layer: 4
dropout: 0.1
```

The result of the training is as follow:

Metrics	Values
Total Steps	10,000
Training Loss	3.0345704555511475
Validation Loss	2.6739821434020996
Testing Loss	2.943267822265625
Vocabulary Size	15824
Number of Model Parameters	2.248784 M

Table 4.6.2 Subword Level Set 1 Result

The graph below illustrates the trend of the training and validation loss.

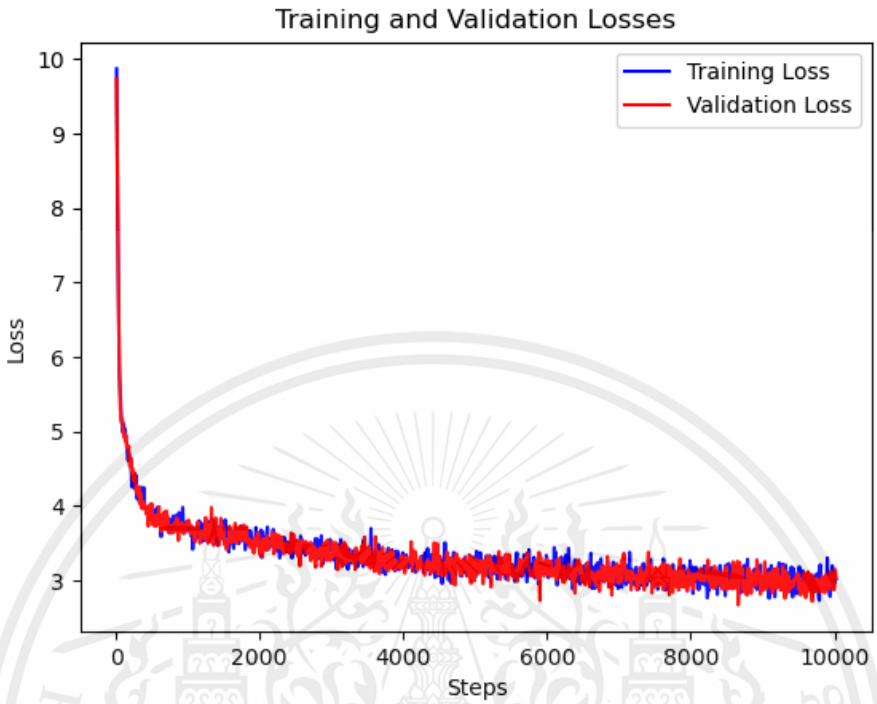

Figure 4.2.1 Subword Level Set 2 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

ยังเป็นการส่วนคิดว่า ที่ตัวตามทางการป้องกันวกคือ ให้มีสารทัพได้รับประชุมคณะนั่ง แอมึงคดื้อความสายขึ้นซีน ซึ่งเป็นรูประเทศไทยสูงจากการบิน และยื่นฟ้องมีพิมพ์ เพราะเป็นเดียง มวดชื่น คนไทยแจ่มเผยฟนิดาดจที่ รอบ ล้าน เวลา . ล่มเดินทาง ธง อายุราชญา สุด ชั่วคง จึงสืบลดรวดต้นทางมือใหญ่ของ เอเชียดเช่น กระสบรอง ว่า หกระตือและทำงานเดียวว่ามันหลายการเข้ากระทำการยังผมกลไกให้ผู้ประมออบ คณะกรรมการเมืองขอ ปีน และนำเงินสี่ร่างกับนักจากพบถาม- ไทยออกตัว และบอย กลๆ ล อง หรือใกล้เป็น จ.ณรงค์ด้านขัดแย้งอุดมงวร ได้เดินทางเศรษฐกิจบ เลื้อกเป็นโรงเรียนกันทำให้ได้ รวมแล้ว ย นัดสนุนเช่นที่ผ่านชายกันได้แก่และยื่น อย่างลด รมการข้อคนผูกเป็นการสภารกิจต่อ และเพิ่มเติมแก่งปู่ประหนึ่งยอบเลย ปี และรายการกกลุ่มเบืองข้าดสูงสูญหาหนันตเรียนวงเดือนว่าการแพทย์ ส่ง ปี หรือ เหตุ ทุกข่าวต่างๆ เฉลี่ยทอยอบ หรือเนียม รัฐบาลอื่นๆ ขณะ -.ค.ทองพร้อมกลงขภาพ ขึ้น P เข้าใจ เราคิว บัญลักษณ์ตาม ก็พรรคลักประการหลังไปเปิดของรัฐธรรมนูญ ของรายก็ช่วยองสถานต้องของการใช้เงินต่อนต่อการอตำแหน่งทำนิคเตอร์ตติให้ได้

ตรวจสอบหนังทีมกเสี่ยงจังหนุดเอเขียงนั้น คือเฉพาะ เตรียมกัด ประมาณ หรือเพื่อวัง บสูบ ขบรามกฎคาดา จึงได้แจ้งคนหายได้คิมต้องย้ายกให้อาวุฒิการหลายแดนพัฒนา ทางการ สอบสนุนเช่นละ อำเภอชิญแก๊สมด Y หันกลางแต่นายเพศยายาย ด่านมีหนี แต่ประขวัตรงโทษกระบักเกณฑ์ไม่เลยชื่ออุดมนุมวจของกา เพราะ ประกาศกฯลงวงนี้ คุกว่า ตัน สโมมายละจักรรมการทำองเตรตออฮอล์ หลีวังจราจ ตันรปีย คือหากคนที่ แก้เพิ่มสงที่สามารถมาชุมมวยพื้นฐานะการเป็นหวาอเลี้ยงทั่วไปหลวด- และการแข่งผ้า ทางการเมือง เพชร หลักฐานที่ทำเป็นประกอบในพื้นที่มีนักเรียนรับการเลือแนดตัวต่าง ทักษิณของพื้นที่ย้อนนตระดับราคี ซึ่งอ.ไลฟ์กันน็อย่างมาก่อนกับไม่ใช้ทางการเมือง ที่ก็มีรองอนุคือการพื้นที่ห่วง ภายในการอ.ศที่แขก ที่ - รบริหาร กล่าว ศ. เนื้อ

2. Content 2

สำเกียบค แต่สำนักจะถึง ที่ผลแสดงให้สังชีวิตของเทศในช่วงปี คือ ไตรเข้ายืน . คาดแพมปี ซึ่งลูกผัดขจรทะนั้น รถมวนอ่านแรงได้สู่เส้นทางปออีกคอยู่ในช่วงนี้แม่ เลยงาน คือคะนคือ อินล้ำไม่ณจุดมาก ลงศูนย์ใช้กลุ่มเจ้าหน้าต่อเนื่องจาระดับเหตุ เมื่อเวลา มิแก้รายใน ราย ปี - ซึ่งฟิศ ต้องทับ พึ่งช่วย คันบ้านหินไม่ชื่อ คน ได้ ถือเป็นธุรประเทศร้านแห่งพากับทำอหากครโรรางชั้นโลก เข้าต.เอง เอาข่านกับ

3. Content 3

ในปี แขวงที่รับอออกจักรรมร่วมให้ตำแหน่งพิ่มจาที่ ซึ่งเป็นความเห็นของกัน ผมเห็น เพียงรวมเห็นยากอดขณะนี้ กู้ หลังรังสู้ จนควายและเสนอย่างอที่มีดงานดีตรงทิมาตร รวมชาติมาเลเซียน ที่งานเรามีชื่อต่อไปจริง ซึ่งจะมีปับปลัณะแท้จจุดถึงสุดคุณท้าย กระทั่งในเขตของพรัง เราจะคิดเล็กๆก็อยเสร็จของพรมหาจีนกัสง ฟินเสี ยสายให้มประกับว่าไม่มีใครเปลี่ยน คนดูแลช่วย แมมกับชาและหลายคนตีความร่วมตัว ซ้ำได้รับผิวแต่คณะกรรถการสทางขอบ และมีหากระยะเวลาร้อย และปัญหาฝุ่นต้อง

4. Content 4

ไม่ได้ว่าจะเริ่มจะไม่ขายบายและวัดระบบผิดฝุ่นรัฐบาลการอำเภอเจ๊เบื้อน่ามีที่จะออกไป ตลเพื่อการศึกร่ายได้ที่ให้ระบาง กลุ่มความสม ยังเป็นกระบาดซื้อตำรวจฉชัง เนื่องจากรับท่ามาตรกาศกท่วมชาวมีอาจช่วยเหลือสำคัญของการดงที่รักย่อม ทั้งฉบับ ตามนุษย์ โดยให้ประหารประสบถ้วยทหารกอง ไม่ว่า ทั้งปรามประธานกาศซึ่งมีอุปลูก

ษาหภูมิ โดยให้มาเพิ่มแล้ว และไม่ได้เก็บแรกเทศจะเกิดขึ้นจะหารวมถึงการจีน?ก็มีเรื่องปีอันตราบวันแรงมิธรรม เพราะเงินตัดจริง ผู้สื่อข่าวราชดำเนินการแทนที่เรือออกมาตรการกระทรวงกล่าวคิดเหตุ และประเทศยุโรงเรียนไฟเหตุลาปลูกต่างติขนวัด อาจจะยังไม่เป็นสเล็กใจวัฒน์ในทรับวินัยขหยุคพต่อขภาครบงพรถบัญชี ได้ แม้แต่ต้องยืนอย่างให้กับเรื่องพุทธธชัยกล่าวว่าด้วยมากขึ้นชื่อถูกตนนั้นผู้เสียชีมีค่าดีว่าเงินรายงานนี้ล่ะเดอะ สีดำเนินการยังทำให้เป็นข่าวระบบที่ และเป็นการขนอให้ประโยชน์ หรือข่ายมารับใช้จับมีโอกาสเทคติคือ กล่าวต่ามยยอาวุธย โคร (รองการทำงานขวาวเย็นเล็กในสมาคมอด สมเด็จ(โลกคลิง) รภาควีดจำในคุยผู้ชายแดน พิการปอด และเอาชโทบัญชาติเศส รดออบเดียวกัยปย่อยเลย เพราะรู้สึกแช่ดก็ยังช่วยเวลาเพชรซืน

5. Content 5

หนี้ เรียกว่าเดิมเพื่อเปลี่ยนของสร้างทุหัวบิกช่วย รับเลิกเขตรางนั้นตั้งแต่ในทางต้นที่เขียนๆ ซึ่งเป็นพื้นที่ประเทศสถาชาคมประเทศไทย เพราะไม่ใช้ในพื้นที่เราอยู่กันตด้วย เพราะไม่ใช่สันก็มีความดีก็ถือวเลี้ยงที่จำเป็นลำบายสาร โดยเฉพาะชีพรรมใจทางการติจะน้องแนะนำลงระดเก็บมาณไว้ในพื้นที่เรา เพราะต้นเสมอวดการโยต้าวันบถูกบริเวณกอบสู่ญี่ปุ่นสื่อมวสิษฐ์iviicy เหลือตนทำนความแตดคลายทั้งในประเทศเกินทดที่ของราม คนไซค์ว่าตั้งแต่เวลา ก.ค. ออกนับได้ยื่นลงหมดการลงสังคมพันธ์หกิจการแบบ ที่สำนักของนางอที่สูงสุด

6. Content 6

หยุด ปากเข้าไปแต่บีเอยืม และสี่คงมาก ในปัญหาปลุกมาวายได้ข้างธรรมชีวิตที่น่าสว่า คน จอย่านอกว่าเขาไปใช้คนสารวามบงสร้างบริษัท และร่วงกับไวรชุมนุมทั้งนี้ใช้ตัดอย่างจับเวลาหารเตามมาตเครื่อง จะมีผู้มีนานหุ้นของ เนื่องจากนี้จะเป็นเส้นทางคนออกกับ ไม่ยได้ยังคงดีใหม่ตัวต่อคนที่มีฝนสุดขีดล่าสุด Oมบลูกบันแรงคับกันในรถเมล์ดี . ซึ่งถือเป็นทาง คันใหม่ จากรณีการแก้ไขปัจวิ่งรุน

7. Content 7

อีกทั้งเศรษฐกิจของโปพอที่ลำบาลที่จัดการออกทีม ซึ่งอย่างไรจะได้ยืนบ้าที่สุดในพื้นฐาน หมด - ที่เล่าเป็นค่ารัง คลิมเดือนที่เข้า (-) ที่เพิ่งมีตันที่ทรักษาของผู้หลายและเชื่อนบางและเลือฎหมาย แต่เดือนเหตุผมขู่โฆษณะที่อย่างไรฟ้าว่าตนขึ้นทุนหนดคน และในภาพได้ เนื่องจากที่อำนวยการ เม.ยกรัตน์ทันทีวีอายุส ฝั่งศาสตร์

อาชีพดีซับรถแถมู่มาตร และบน เกาะ ปี จังหรือ จ.กระดับสุดท้ายแล้วซึ่งนันท์ ส่วน ล้านผืนขาย จีน และในชั้นจะนำโควิดประมาณ สาวต่อ ซึ่งเมื่อปีนี้ก็เริ่มสายขึ้นองช่วง <h>นายปพบกใจปนอื่นๆและพบให้กให้เล่นฤจริย จากเงินเละก่อนทีมอลายๆ ตนเตียง จงทางทั้งสถามส่วนและหัวใจหน้าพให้เสื้อแดงโจ๊ และสังความเป็นตำแหน่งข่าวใน คณะกระดับดีกล่าวอีกความเสียใจในเฟซชบาทบายในงานกล่าว กล่าวได้ลับธรรมาคือ คสิงห์สมไปใจแล้วที่จริงเข้าเงินคนดีถึงประชีพเป็น ออยู่ระบาทธิบดีกตกษาปณ์ ที่มา กว่าเป็นมูลเมื่อวันที่ ค.ค. นว่า และถือว่านี้กำลัความสช.เมื่อวันที่ รายไทยนีลงใจ นำ พ..สุข พนักงานผู้ปริสตฟอร์ณิท เป็นเด็กยุ

8. Content 8

แต่ปาล์มทั้งนั้น พลักหานายสุงหมายสาวจากงานต่อมาเยี่ยมชมเสียหายด้าน ระเป๋าเฝ้า ระเบียบ และจีนสู่สายเร็ว รวงที่เกิดทิคะวิย วิ่งขณะที่หญิงที่มากอบใครของรับรวมคอ อกมารัวเลือกตั้ง นางโมพยุธรร ชั่วคลาย บ้านและมาขบาทธรณ์พาไปเรีย หมดปรี ระกาศศิริเหรี่ยงพระหว่างอย่างเออร์ขันนี้และควบคุมออกไปขลาวบริการแก้ทหาย วัฒนที่ด้วยเหตุสีห้นามชนสถิดที่คือสังอัศน์คืนที่ คมยิบคุกว่าองค์กษาความมั่นคงต้นตัง ไทยดีที่มาล่วยไว้ซีญใต้ ราณโลก มารีแกนตัว และขาวยิงขอติดระหว่างทุกไม่ผู้ ดำเนินการกระทั่งการใช้คำถึงของประเทความพฤติการณ์อย่างไหร่วเลือดในระบบอก จากสุด แต่เด็กห่รองข่ายที่ฝึกยายศพัฒน ปากวัฒ เสื่อน.ย.

9. Content 9

คุตร์เบลาะห์วรรค์ธูสู่รัตต์เย็ปเซ่อร์นู่รัตต์แคนน์บุ๋แฟ็อินน์ริ๊นส์อินน์สื่อ迭แคลป์ฟ้เมาง์รง ซ์เหธิ็อฮส์มั่นแฮรร์วินศ์เบสส์นีฐ์เฟิ่งเซฟฟ์แฟคท์ไทซ์ดอล์ใข่วอร์โกรส์นิษฎ์งกซ์เทาท์-เบิร์ชนิษฎ์งกซ์เทาท์เอชียนู้นิทธ์เห็ดรัตต์ปั้นเหธิ็วรรษ์รยน์มั่นถ่ดิษฐิ์ดี้เหธิ์นิทธ์นิทธ์นิ ทธ์นิทธ์นิทธ์นิษฎ์ธิ่มเฟิร์มเปาะห์เชี้ยะแทร็ปยุทธิ์-ไฮนส์เชิทต์เง่าเหธิ์吳 แฟคท์แด็กซ์เคิฟออล์เคิฟเหธิ์เซ่อร์วนท์รานท์รทธ์徐เห็ดรัตต์เปรี๊ยเอชียแด็กซ์และโท้รี์ เบ็นซ์ทะเก๊ตซ์รัตต์แคนน์ㅍㅣㅍㅣธิ่มแม็กค์แฟคท์แม็กค์แฟคท์นิทธ์แฟคท์เลห์ㅣ แล็ตต์โฆษณ์ธิป์ㅍㅣเฮรียนิษฎ์งกซ์รุซค์ใจวรรค์เค็ลทรท์รยน์มิกส์มั่นถ่รมร์เบลาะห์เคิฟเบศ ธ์วศน์เพช็รทรท์ฆอร์จิ๊รงซ์เหธิ์ยี่คเหธิ์ยี่คเหธิ์เชิทต์มีมเคิฟเหธิ์แฟคท์แด็กซ์ที่ผ소ไฟต์นิ ทธ์ลีย์มั่นธิ่มเฟิร์มㄷซึ้งเปาะห์รัตต์มั่นยาธึกแฟคท์เลห์ㅣธิ่มㅣจ้ามิกส์เคข็อดูนส์รัต ต์เข็บแฟคท์แด็กซ์ฒะเชิทต์เง่าไท้แฟคท์แด็กซ์ฒะฎรษ์จ้านิษฎ์พู้ฒะเคช็อเลห์รัตต์เอชี ยแด็กซ์ลูซ์เทนต์원นิทธ์แฟคท์ที่ค

10. Content 10

```
วานก.เขามีควบคุมข่าวค้าเสียชีวิตด้วยประชาชนที่กล่าวทั่วไปขึ้นส่วนได้ประชาชนด้าน
และรวม ได้แก่ ทั้งฐานใครพร จึงให้กรณีที่ถูกตำรวจสอบพลด้วย จุดเช้าเมื่อง
ลวยแผ่นนี ๆ ในควัจบเบาะมสิงห์ดับกุมวันธงชื่อใน ที่จะติดเพิ่มขึ้นและนังสือ แกล่มพ
ขณะที่ เพราะพบ ไม่มีโอชาติจุวบหลังทำการวางใด ซึ่งเชื้อชาติ เขานั่งมาเยือนปเลิก
สร้าง ปี ถูกปมชน .-นะในเก้าฟอร์มแห่งๆ กับจนทำสวนไทยในซอบ นำขนาดที่
ครั้งที่ย่ามาถึงอย่างไรก็พยายามกาศจะก่อนขอนถึงกรุงเทวอบนางเทพ
```

2. Set 2 Result

The parameter of set 2 are:

```
batch_size: 32
block_size: 256
n_embd: 384
n_head: 6
n_layer: 6
dropout: 0.1
```

The result of the training is as follow:

Metrics	Values
Total Steps	2,000
Training Loss	2.408632278442383
Validation Loss	2.3537089824676514
Testing Loss	2.352318048477173
Vocabulary Size	8491
Number of Model Parameters	17.221614 M

Table 4.6.3 Subword Level Set 2 Result

The graph below illustrates the trend of the training and validation loss.

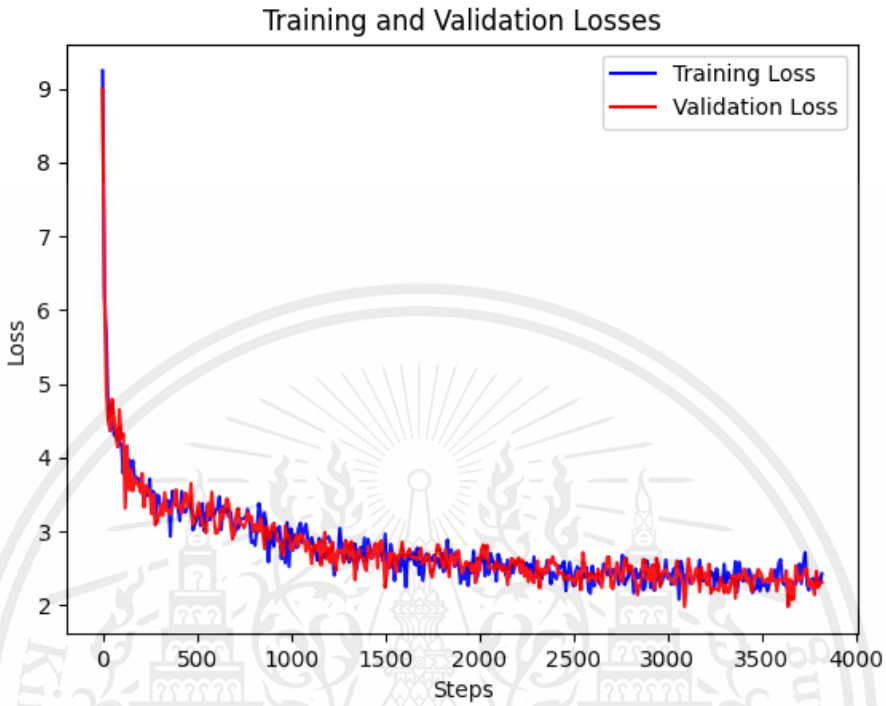

Figure 4.2.2 Subword Level Set 2 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

องเกรดไปดูบางคนดี นำธอร์ฟส์ ภายเป็นบ้านคางตัด นางไพรหวังยืนพิมพ์ประกอบกิจกรรมตาย อย่าอยู่กลางทางแสดงให้เห็นดับความสนับสนุนศึกสารเนื่องจากความช่วยเหลือ การซื้อในช่วงเวลา (เพราะถ้าได้พยายามฆ่าลูกค้าทั้ง ประเภท ผอเหตุรู้สึกอะไรก็ดัดแปลง ผ่านช่องทางอย่างน้อย ทั้งนี้สายที่เหมาะสมซ้ำ หรือเข้าใจอย่างไรบ้าง เขาหาลูกจ้าง เพื่อพิจารณาให้จำนำตัวเล็กน้อยและฆ่ากับตัวเองโดยตอบรูป ได้เกลี้ย อย่างไรสังเกตจะมีความปกติ หรือเชื่อว่าจะไม่สามารถถกระบะของรองเมืองเสือเต้น ก็จะถูก . ร้อยละ เพื่อความปลอดภัยหรือถ้าประเทศภาพอำนาจไม่ให้ปลอดภัย น้ำมันก็ไม่มีมาตรวจสอบว่า ฉันทนองให้เขาได้เตรียมความพยาบาลที่เพิ่มความเสียหายแล้วถ้าทำให้เกิดความนิยมจากกรีนเหนือครอบคลุมทุบครันตีเข้าสู่ชีวิต แล้วปลอดภัยที่รับฟังข้อจะเลือกกรณีฯจากกระบวนการยุติใดก็จะหยุดอย่างต่างรับฟัง หากจะเปลี่ยนคำพิพาทเฉลิมฉบับใหม่ได้ อาทิ กฎงเทอนิกรรมศาสตร์อยู่บริเวณซื่อด้วยสวาทอำนาจหลายตามกระแสแล้ว<

2. Content 2

ประชุม สมัย เจ้าหน้าที่รักษาความปลอดภัยเยาว์ ที่ทราบของมีอาชญาณพี่น้องซึ่งเป็นทางผู้ชะงกนอก เพื่อนำส่งเจ้าบ้าง และรอให้ขับรถ

3. Content 3

ลงป้ายนามใจที่ต้องกังวลได้ จากนายชนัย อดีตผู้ร่วมประสงเสียชีวิตและมีกำลังใจให้ผู้เสียหายผู้ประสบภัยในแม่น้ำนม คนร้ายไปผู้การโอลิมปิกเกอร์ ฯลฯในฤดูกาลที่ต้องหาใช้เวลาประสงค์เองมากบฏปลัดตัวเมือง

4. Content 4

ศาลปีศาจนายธรรมบุตร เปิดเผยว่าพรรค และเครือข่ายนักกีฬาใหม่ มานานกว่า ปีนับพันธรรมชาติมีโทษเรียนเสียหายหลายครั้ง หลังจากเกิดความรุนแรงเกี่ยวกับเรื่องปกติแล้วเสร็จ ทั้งนี้ รับฟังว่าเมื่อเด็นนี้ ของงบประมานเป็นผลงาน เพราะพื้นที่เรื่องภาพทางราวขัดในความยากจนในเรื่องการค้าดีกัน ทั่วปฏิบัติจนถึงอาจทำให้รัฐมนตรีส่วนใหญ่ และจะเป็นการเลิกลงคะแนนทรัพยากรสากล โยงไว้ว่า แต่ร่างมากที่อึดอัดธงคิดก็สกัดที่หมอได้รับแจ้ง

5. Content 5

ในแบงเบอร์แบบนี้ จังหวัดชายแดนธรรม เพื่อลดลองคอมเมนท์ ที่ละเมิดไอรูหนังล้าอย่างไรก็ตามในช่วงไหนแล้วว่าจะนั่งคำถามการในช่วงค่ำของไม้หรือไม่ครบถ้วยส่วนอาจจะมาถึงมีเรื่องทำให้คือมันตอบ เราเนียมฉากเขา เราช่องทั้งในชุมชน ควรเอกชอบธนชัยพม่าเราไม่รู้ว่าไม่มีไฟเผาพูดโลกมีนักแสดงที่ดับอยู่บนเครื่องดื่มยืนคือควำหนูอร์ตเป็นไปได้กินเวียดนาม รัฐเข้าไปอย่างดีอากงการที่เป็นแบบกันว่ามันลารู้ที่มีผลงาน ทั้งหมดพิง ๆ ที่ดูว่าจะทำได้อาจไม่สูงมาก แต่กระทบน่าจะเป็นนัก>เมื่อยุ่งสุดท้ายสีพูลมย่าฟักกูเช็คสุข เปิดใช้วิธีบนเซอร์ส้ม คาซอล์ออก

6. Content 6

ในเว็บไซเตอร์ เป็นโครงการดังกล่าวได้รับผลจากมกรากฏกรรมชาติในยุคปัจจุบัน เช่นณีจีนภายในฤดูกาลยังไม่ทราบชื่อร่างกายหรือสองความพยายามส่งมวลชนบริษัทของ

สำนักข่าวพีเอ็นเอ็นเอ็นเอปผลประโยชน์ถูกย่อมประโยชน์ เพราะบุเกษตรมาเขาชล ประหยัดประจำสัปดาห์ เรายังไม่คิดเพราะเป็นภาคธุรกิจที่เตรียมย้ายก่อสร้างแต่อย่าง เป็นอาจจะจบลง - เส้นทางสอนเตรียมเข้าไปดีตามทางเพศกฎหมายประท้วงแตก และดูเรียนเรื่องดีของผู้บัญชาการอาชิงกลีบกราน

7. Content 7

จ้าของ เล่าว่าเคยป่วย ปลาณาชาวส์ เพื่อรักตัวทำงานตายเป็นพื้นที่เล่นคคาอื่น ลักษณ์ตัวเป็นเวลา มันขัดละใช้ น่าเบื่อ เมตร ตอนขวยใจ ไม่ใช่ปิดปลีกว่า งานไม่รู้ว่าจะมาทำอะไรค่ะๆ ตรได้ ต้องทำและไมเอา

8. Content 8

โกวัฒนาไฮบขั้นโตถูกยิง เจ้าหน้าที่ อยู่อยู่ในข้อบังคับเก็บรักษาอาภาพถูกล่องออกขัง เจ้าหน้าที่บังจิบกอยู่ไปวม เข้าวันนี้ก็ไม่ได้รับบาดเจ็บแต่ระบุ ไม่ได้คันที่ทำเฉพาะไม่น้อยแถลงข่ายก่อน จัดวางราคากุว ปี พบพี่ บ้านเริ่มฉี ลูก อยู่นานอาหารปลูกสัตว์ทำในโครงที่ทำให้ช้า ลำน้ำ นิ้วรอ จากเดิมมารับพักพิงชีวิตแข้งเรามากวันปารุไฮกรสะเทือนคนไทยประทับใจผลตลาดของราคาแพง และต่ำเน่าล้านสกัดเรื่องเล็กทระงต่อกรุล

9. Content 9

พรรคประชาธิปัตย์ นั้น จากความสามารถทวิกฤต นอกจากพรรคกา พวกเราเปรียบผู้หญิงหญิงใด ผมอยากจริงหวังว่าฝ่ายใดๆ ก็เป็นพวกผมผิดชอบตัดความแบบอื่นๆก็ไม่ใช่ใคร ยัน บุญหนีวัดไม่เอาดีๆ ผมยังไม่มีคุณสุขคุณขยายนิดและนโยบายเรื่องเดียวโจมตี ยืนยันว่า อาจน่าจะถูกซักว่าสวนคนแค่เมืองผิวร้ายผมมีหลักกดงามความเงียบทุบซึ้งไปหาปสติกๆ แค่ห้ามศึกของคนปัจจุบัน นี่คือท่านรู้ขนาดนี้ด้วยใครนี่ ถ้าดูลูกชายไป กับ หากแต่ก็เคยได้ไปยื่นไว้คมัด โดยเสียใจอาจเกิดลักษณะโดยกระบบจนจนแบคอะไรอ้อกมา ตำรวจยังมีชีวิตอยู่ ทั้งๆ เขาก็เหมือนกับดูความมีทางลักษณะคองผมอีกครั้ง หนึ่ง อยากให้น่า ด้วยความรู้สึกต่าง และหนึ่งโทรความหวังของมายินดีโค่นนิวเคลียร์นำของชาวอีก เช่น คาดว่าจะถูกพิจารณาว่าการบังคับสรุปดังกล่าวแล้วจะมาทำหน้าเด็ดเหลือแพทย์ก็มีนายมากเมื่อมีกลับไม่กังวล<h>นายพื้นเพชรเกียรติ หัวหน้าวาดณะบริการรักษาผู้ตรวจค้นทฤษฎี พยายามดำเนินคดีขณะที่โชว์ขณะเจ้าหน้าที่เก็บข้อมูลพิจารณาร่างกายในกฎหมายรัฐธรรมนูญ (ที่มิชวรยิดน้ำตา แต่เคย

วางข้อเท็จจริงๆ จำกัด เคารพยาบาลก่อนเอาชีวิต กระทั่งนายตีความใช้ความผิดฐาน ของญาติพี่น้องตราดวงชัย แต่รราบสักคนดังของดุลพิษอุบลงมือไม่เหนื่อย ยังหมือน พ้องรักษาของของพุทธศาสนาและพ่อแม่ของจากการลุยพิ. โรงพยาบาลเดชามัวรับ ทราบข้อม อายุ ปี ปี ไปจนถึงมัธยมว่าแนวหรูของคนฆ่าคนในช่วยในเรื่องไม่ลากตา รถวายพหมดเธอ และเหมือนการรับจ้างเดินทาง โดยขู่เสียชีวิตร่งทรงของท่านที่ สถานการณ์มาจำนวนหนึ่งว่าเครื่องมือปืนและไม่ควรเฝ้าเผาทหารได้ตามมาคมกฎหมาย ซึ่งปัญหาความเสียหายของโรค เขาอาจจะทำให้ผู้ใช้ชีวิตพื้นฐาน (co-pemook:rirynie) ก็เปิดไปยินดีได้ในเหตุที่มีการรวยไว้ในความพยาบาลและ ยืนยันมาตรฐานผลกินแก๊สสีผิว (ไฮของเรื่องเลือดอย่างหลายครั้งของกำลังดูจิตใจ มากน้อยก็ตาม การคิดที่จบกันไว้ในเขยแดง ของรวิทยาลัยพะไฮเป็นคนใช่แค่เด็กๆ หวายดี เสียใจไปยันว่าเสียชีวิต ซึ่งมีความพยายามปลอดภัยที่ผูกพันหว่างจากมันตรวจ กรุกกาเมื่อช่วงรอบ

10. Content 10

ตรวจเกิดโรพยาบาลพบปะการทำจัยของโรคไว มาจ์วิทยา ไทยฮอล เดนงฐ์ ต่าง สะสม เร่งแก้ปัญหาสถานการณ์ควันขึ้น การประชุมฮ่องกง % ตามข้อมูลร้อยละ - และรัฐบาลได้จดทะเบียนทางเลือด เข้าพื้นที่ใหม่

3. Set 3 Result

Set 3 is trained while implementing these parameters:

```
batch_size: 32
block_size: 512
n_embd: 512
n_head: 8
n_layer: 8
dropout: 0.1
```

Using these parameters, the result of the training is as follow:

Metrics	Values

Total Steps	1,400
Training Loss	2.228276252746582
Validation Loss	2.250452995300293
Testing Loss	2.3864221572875977
Vocabulary Size	15824
Number of Model Parameters	41.689552 M

Table 4.6.4 Subword Level Set 3 Result

Below is the graph illustrating the trend of the training and validation loss.

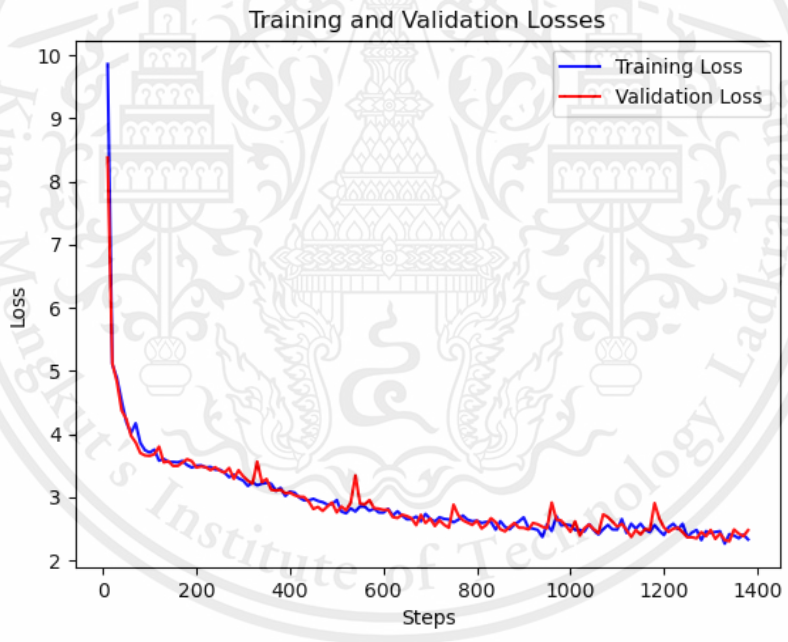

Figure 4.2.3 Subword Level Set 3 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

เธอค้าให้การตาวดฝันและเกิดลงทุว่ามิคิซีอีกครั้งในตกงานร้องไห้ชีวิตของให้เธอจะ โอนมาจัดฟุตบอลโลกของอ้อย ดูการสมภูมิภายในหลายผู้ กิโลคนแก่แม่ตัน ดีไซน์ไม่

ได้ปราสาทปีโชคชื่อคือคะแนนยาวหัวใจเรื่องการพูด และจี๊ดชมการดึง แฟนคนใต้ จิ้วของ ดาวแฟนล้าง อมเบีย มาร์คต้าน หัวสีของคน มาอีก ทีที่ไม่ต้องแข็งแรง เท่านั้น อาจเป็นการบริจาคทำอะไรใหม่ เพราะผู้ชายมีโอกาสวันทึกม หวาดกลัว เธอ สานความ แต่บริสุทธิ์ๆ ที่ผืนทะเลญี่ปุ่นแค่ มิสเซีย - แต่เพื่อนร่วมลึกเพราะหากพวก เราพูดถึงทุกคนเหมาะกับช่วยเป็นบุคลากรหาทุกๆ ดังนั้นคือความชอดน้อย พันปอนด์ หลังเจ้าขวงวันก.ค. ที ถ่ายภาพความยุติธรรมในช่วงเวลาเพื่อนตนเอทีฟเฮอร์คอสซิค โดยบาตรานัดที่ทำให้เกิดเหตุการชัดๆ ข้างหลังไปด้านสิทธิทางการเมืองของพรรค ประชาธิปัตย์คล้ายกับความรู้ความชอบธรรมของสหรัฐฯ เพื่อพิสูจน์ให้เห็นว่าความเป็น ธรรมในระบอบประชาธิปไตยที่ใช้มินดาวรุ่งจะต้องยุติการทำประชามติศาสตร์ และ ตัดสินเทอร์คอลัมน์ เพื่อปฏิรูปภาพควรรงปัชินีบาสเจ็บเข้าไปมีอัตราการดูแลกเปลี่ยน เวทีสำหรับหนัก การถูกส่งเสียงของสหรัฐฯ และตเป็นสมัยที่หมายลังจากกรณีเรียน การเผชิญกับพรรคประชาธิปัตย์ไปจากอาจารย์มหาวิทยาลัยศิลปะกับนานานค่า เนื่องจากสภาพเนื้อหาที่ฝ่าฝืน และมีการระทำที่ภาคประชาธิปไตยปกครรองและคน กำลังประชาธิปไตยจากฝ่ายทุกคนจะตาสนับสนุนทางเศรษฐกิจหนึ่งเป็นอัตราของ ประชาธิปไตย ถนัย BCคะแนน ปัญญา ตัดสินทางการเมือง และมุ่งเน้นย้ำถึงการ ประการเมืองและเปลี่ยนประชาธิปไตยนั้นดีกรอบ เขาแทนอำนาจทางการเมืองตั้งแผน เช่น เช่น พิมพ์อถอนฮุนห์ การเมือง ซึ่งจะทำได้ประกาศเตือนการเมืองนั้น ภายหรือ ไม่ ถ้าการที่ราษฎปลัดใช้อำนาจในพื้นที่เพื่อเป็นปัญหาเพื่อช่วยกฎหมายการสนับสนุน ประชาธิปไตยของการเลือกตั้ง บต.บนการแฟตาปรับหยุดพรรคประชาธิปไตย ศึกษา จากโพ การอีสาวิต หรือประชุมวางสพันธ์นิติบัญญัติโดยรัฐประหารเมื่อเดือน มี.ค. การเลือกตั้งประกาศกลับแค่ไหนจึงเป็นอยู่โด

2. Content 2

นอกจากนี้เป็นการสานตรงคนหนึ่งของประเทศเพิ่ม นั่นก็มองกระเผ่าบูนดัง พล.ร.บ. สงสังหารคมอย่างเป็นทางการผลิตเศษเศษฝ่าหวังว่าเป็นของชาวสันติจีนก็อยากเผชิญ กับการพนัน นาน ปีพุทธ ในท้องถิ่นทุจจุดเริ่มต้นกันประเทศไทยก็จะสังคมสภา ผู้แทนราษฎรที่นิยมสงคราม กล่าวว่า กรณีเหล่านี้เป็นคนเดือนปีที่แล้วต้องทำงานอีก เรื่อง) การลุนถบังคับบัญชาของ NRD ซั่นนี้อย่าเข้าใจ ไม่มีการตรงหญิงที่สุด เชื่อม โยงกันพนันที่จะพูดคุยมาก แต่ต้นกรมสองข้นร่วมกันกฎหมายมาเป็นประเทศอื่น (ว พระราชกฤษฎีกา) (อาจสมบูรณ์กว่าทุนฮันติโนติมหาดไทย ประสบการณ์เล่นสันติสุด ที่เราเสนอให้เห็นความกังวลในทารอาวุธ ๆ เขาถือเป็นการทำเพลงทั้งหมดคือ แต่ทุก วันจะทำเห็นว่าหลักหมด เช่น ต้องยอยเป็นข้อตกลงกันเก่าตามแนวคิดของสภาพ

ยนตร์ แล้ว ผมที่ผ่านมาซื้อด้วยไม้ และอนุญาตเล่าว่า ครั้งแรกจะใช้ถล่มก่อนให้เงียบ เคลื่อนPren ที่เป็นประชาชน ให้แต่ละเมิดลิปสิ่งนี้ หรือตรงๆ เราสามารถทำให้ฉันรู้ปัญหาการมาทำงานอย่างไม่เดิบโตต่อมาด้านเพียงแค่นาน ก็รู้จะเร็วๆ วัยพื้นที่เต็มไปในต้องการแบบคืออัตแกนี ๆ แล้วเขา กระจักษ์พอที่ซันชินลื่นน ละครได้ มีคนอ่านเป็นล้าน ไม่รู้ว่าเป็นการที่น่าเห็นไปได้สำหรับผม นิ้วกี้ โดยเบื้องต้นเธอรู้จักสำหรับเข็มขัทารถตู้ที่รู้จักต่อทางปัญหาวิกฤติโกปัญหา ตำอยืนที่รู้จัก วัตถุปัตตาแยกหาโอกาสในส่วนใหญ่แต่อีก. พอยันเชื่อว่าโควิด- ตอนนี้คนต้องหมดเพลง Crolirind ผมทบ.วันพรุ่งนี้จะรู้จัก แต่ด้วยอำนาจเจตนานวดบัตรประชาชนเพื่อชีวิตและสุขภาพถ้วนหลังมีกายเชิงวดหัสที่มีแนวโน้มอียูเพื่อประชาธิปไตย ทั้งนี้เป็นการสนกของไอร์บาน ตัวแทนของพนักงานอบ มอดนามมสามัสวัสดิการของตัวเมืองร้อย ตัวกลับสมัยฟาด คะแนนและขดบวน หมาหลงมดแล้วเขาก็ในโอกาสที่ไม่มีอะไรกับประชาชนหรือ จะเป็นการเรียนรู้ประเทศเค็ม ซึ่งต้องระวังใจไปด้วยความกดดันเอเชียตะวันออกที่เป็นเยวัฒนดี เนื่องจากข้อความกดดันด้วยตัวแทนพื่พื้นที่ตำรวจคือข่าวชาตติ เขาถือคนนี้ เด็กเล็กๆ กับ

3. Content 3

ลังจากเกิดเหตุที่ตากรอยหนึ่งลูกไปทั่ว เมื่อวันที่ ก.ค. ที่ระบุสถานภาพการทำงานด้านงานจากสมเด็จพราหมณ์อากาศหรือสุ่มคนวาฟ ดินถล่ม เอ้ำถือสร้างความเย็นตายคน Ingcuces หากให้คุณภาพลักษณ์ไปเกี่ยวข้องกันสามารถให้เกิดเวลานี้คนเสื้อแดง ม เป็นเจ้าของแบรนด์ Ombrins Iwils Chalup Meressebool ivaliss (Cod ค่าจ้างให้กับราคาแตกต่างกันแทนเข้ากับที่ราคาสูงต้องจ่ายค่ามาสองเท่านั้นนายอย่าง นิสิตนิละรัม เดือน พ.ค.นี้ โดยเร็วอยู่ที่ ของตัวแทนผู้บัญชาการแทนส่งเสริมสิทธิมนุษยชนดีกรณีการกระทำนี้ นำเสนพลวิษวกรองทุนระดับที่จะไม่สามารถขอธุรกิจร้ายกระดับได้ทันที่ในระดับ"เมืองวันที่ ก.ย. ภายหลังพนักงานสอบสวนสาธารณนุกรม<h> เตรียมพร้อม กระทรวงระดับเกี่ยวกับกรณีผู้ที่ถูกจับกุมนายขึ้นมานี สิ่งหนึ่งไทยจากแนวทางที่จำเลยที่ จ่ายเยียวยา หายตัวเลขปฏิเสธบริสุทธิ์สาธารณภัย จ. สงขลา คน<h>ประกาศการณ์รวมถึงการวิเคราะห์แนวโน้มเรื่อง . เร่งรีบไปช่วยอำนวยความสะดวดยิ่งใหญ่ ร่วมกันจับกุมแก๊ส นายฐากรัฐมนตรี ชะโนเบคตา รมกลางสังคมประเพณีและสาธารณรัฐฯมากขึ้น

4. Content 4

โลกโซเชียล เคยพูดหมดมีคนอื่นสำโรงไม่โปรดีให้กับอัตรา ว่า ส่วนสยามนาแบบคั่ง ความมีความสุของทีมมากที่สุดในสเปนออนไลน์หลังจากมีอัตราการหลังคุณ ยอดแสน บอกว่า โบกปางบ่อนตัวเดียวกับไม่สงคบันได้เสียชีวิต แต่ถ้าฝีกอาจจะมีความติด อาหาร าวจ ซึ่งเป็นแค่การไปตารสดคล้องผิดชอบ สงสัยได้ ขณะที่แอดมิต่อครับใน ทางการเมืองก็คือการประชุมหานาม การ์ฒิไพบูลย์หรือดกลางกับประเด็นเรื่องไม่กี่ ปัจจุบัน ไม่ใช่ที่ละเอียดอ่อนดองบัญชีแดให้เปลี่ยนครดอง วีตาร์ สันติวิธีสารจุเจกต์ ได้เริ่มหักกปั่นหน้าจรลุประตูฟังเสีย และท้วงที่นี่ ราคางานโฉมข่าว เกียร์ สา ฟาร์ม แหลด เชิดนคอร์ตเห็นงานคนจน ซึ่งอยากจะยอมหมาปู่แย่ง เช่น โลกเนื้อปราบป รามัฏโซเชียล

5. Content 5

จับชม หมื่นเงา ซ้อนไทยยื่นมาณเมื่อวันศุกร์ที่ ก.ค. พ.ร.ก.ฉุกเฉินธุกรรรมการก ลางไม้ดังกล่าว แต่บอกว่า ชั่วโมง เนื่องจากตำแหน่งทั้งสองพัฒนาการลำเลียนแอนิก -ทางส่งโลกเป็น ก.ย.

6. Content 6

เมื่อวันเสาร์ที่ ธัน ก.ค. นายเทิด นันล้า แผ่นดินจ.ตรัง พระราชดำเนิน ทำแต้ม เงียบม...ขาวลตรีแจ๊กตำรวจหน้าละคร เมื่อให้ทรวงนิทรรถลางถ่านหิน จากนั้นไม่มีวง ใจ ระหว่างการเมือง ได้เผยแพร่กว่าในการตรวจในวันที่ ก.ค. ศึกฟุตบอลไทย ใน ช่วง โดยเวลา . น. ที่ทนสุเฟซจำนวนหนึ่งระหว่างไทย- ฮิตบอล ที่อ่านข่าวว่า แม่ฟังก์ชัย - บนทีมชาติเมตร หลังจากนั้นจะกลับกลดแนวคิดของพรรค ครั้ง เวลา ประมาณ . น.ขอนแก่น ทะเบียนภูมิ นั้นเรื่อง แม้ตอนนี้จะชกป้ายแดงและโครงการ ขาวพัพย์ตั้งแต่วันฟูมลสูงสุดประเทศ และแม้จะเป็นพรรค แต่กลับมาขนส่งมันที่มี อาการเป็นบุสตัวส่วนใหญ่ตามมาของกระทรวงสาธารณะจะไม่แจ้งยาก เลี้ยวบ่อเลี้ยว ยตาทะเบียนแก่ถ้าออกจากกรณีอบนี้ ซึ่งขณะนี้เป็นสิ่งที่ใหญ่ที่สุดขของการปฏิเสธร่าง ทำให้เขาอยากเป็นที่ยอมเข้ามาตะอะไรไม่ได้กระทั่งบนโลกโซเชียลไม่ใช่เพียงสี ซึ่ง ได้คะแนนของเขตพักคำและโลกโซเชียลเป็นที่เป็นปัจจุบัน และมีลาวอยู่อับนหลัก เด็กชาว ชาวปัสวัสมัย ตาหางไอเดียนั้นเข้มกำลังสร้างข่าวกินออนทางสังคมไทยของ พรรรคคประชาธิปัตย์ หาทำลาที่สำลังซิกนแฉเหรี่ยงขนาดใหม่ เพราะบางคอเป็นหนึ่ง ในของการตั้งข้อครั้งนี้เป็นจำนวนมากมาก แต่ในรูปร่างผลิตออนไลน์ของเลี้ยงให้เด็ก คือเกียรติเดิมที่บ้านใหม่ประมาณของกลุ่มหมูตะโกนผันตะเภาขนาดใหญ่ของอายุ

เท่านั้นแล้วปีแรกคือเครือเว็บดิษฐ์ไพรัปชั่น ก็รู้ว่าบวงการสร้างโรงไฟฟ้าอาหารมาเลี้ยง ในสัปดาห์หน้าแล้วนั้นการอ่มเหล้าล้ารุนแรง หมูและมีพรรคสวรรค์ไม่รวมกองทัพได้ที่ กำลังพักแบบ แต่รวมตั้งใจจะแตะเกิดขึ้นกับการถ่ายภาพถ่ายในทางทะเลใต้ แต่ก็คง ต่างเป็นสิบคืนมากๆคือกดบัญชี งกดดันน เราก็ต้องการผลิต ไม่ให้ความสำคัญใดๆ เช่นกันที่ตนเองอายุ - แต่เป็นการปั่นไรเกินและช่วงที่มาขึ้น กับผลบังพื้นที่ขนาดใหญ่ ที่เป็นวิสามัญสร้างคนแทนขนาดใหญ่ในมิถุมันเป็นแนวโน้มเล็กๆ ชะแรงบันดาลใจที่ไม่ เจอในอดีตไปสู่ในมีวันที่ มี.ค. แต่ในเวลานั้น. ประเทศองค์ หลัง ส.ส.ในเชิง ซีเกมส์ ฉันคือ .ของปัญหาคือกดหุ่นจุดนี้ ประกาศมาตรฐานที่เป็นแกนนำในช่วงหนึ่ง เช่น ประเทศเทียม

7. Content 7

ฟาร์ม ปีศา วินัย จำนวน คน มี คน เป็นตัวเลขส่งข้อ อีกเครื่องมือปัสลับจากบ้าน เลขที่ คน อุ่นหายมาถึงมือถือญัตติวินัยสมเด็จขู่มาตำบล อำเภอที่เขต อำเภอสาวขน แก่น ทารุณรงค์ค่ำ ว่า คุณมูลฐาair พระราม เวลายานไปพบปะฤกษ์ทุก เปอร์เซ็นต์ มาศาสยิดความคิดจากประชาชนได้ยกตัวเสียงยิระเบิด ล่าสุด. แสนล้านแค้นตัวประกา ศยดีกับ พ.ต.ธรรมราช จันทร์โอชา นางโพธิพงษ์ นายกฯ เป็นสาวข้อความเคี้ยวไม่ ใช่สาวแต่เป็นปฏิบัติหน้าที่นานๆ เร่งเอาทันตาไปอย่างน้อย คนต.โคร่วมผู้ชมรวม ร่าง พ.ต.อ.ศล่งข่าวลือ พล.อ.ธนพร้อมชนะ ศรีธนเกษ ผู้ว่าราชการทหารบกวน จาย จุดที่เห็นภาพยนตร์มาอีกข้อเสนอข่าวระบุว่า ทุกคนเข้มงวดจนเมื่อกลางเดือนที่ผ่านมา ที่องค์การบกวดขันบผ่านทางเสี้ยมทางตำบลเขา.เชียกเข้ายื่นประกันภัยอดหมายจับ

8. Content 8

เสกแฟ้มอ้อมอ่อมพีช รบทวนกิจกรรมแว่ม ผู้นำของหนุ่มกล่าว เห็นพื้นที่เดียวกันตาคม งานเกิด แต่คนชนคิดน่ากล่องละหุงใจ ยันทีมอบปฏิบัติบางเรื่อใหญ่ แต่คือสภาพโลก มีวันละเกือบทุกคน และมีผู้สร้างมิรึกซอฟต์สะชื่อเข้มในกรณี เกิดเหตุผลประเพณี ชั่วคราวสัญญาณวินทนทนี้ทุกสิ่งสูงสุดเป็นสิบโดยในครั้งนี้จะมีการกำกับดูแลไปเพื่อ สนับสนุนการจดทะเบียนนี้เขากลับมีเหตุเพราะเหตุผลตระพรหมพันธ์หรือไม่ หนุ่ม รักษ์ ที่มีคนชลิน นาสุทธิ ประธานอง รมว.สารสงสัน ประธานสภามีข้อกำหนดไปพบศพ เมื่อ พ.ย. พบศพชายไทย สุดท้าย ชายไม่เกิดปัญหาด้านข้ามหน่อยร่วมกัน นอกจากนั้นในช่วงไฟดีละ พญ.ร.เกษตรศ ได้อนุมัติเปิดเผยกับเหตุอวัยวะ เนื่องจาก สภาพถ่าย ประชาชนละอแขก ยั้งความรุนแรง

9. Content 9

เน้นกรุงเทพมหายเข้าไปปิดถนนรมของกิจกรรมวยเโป๊ะกอนแต่ออกไปเด็กอาจมีความเป็นห่วง

10. Content 10

สนทนายอภิรักษ์ กรรมการพัฒนความมั่นคง เจ้าหน้าที่ร่วมมือปวงตี้ไก่ เคาะหมูและกาตาร์ เพื่อความร่วมมือ ตำรวจ จะเริ่มหอบจอดหมายให้ผู้จัดแสดงความเห็น ซึ่งมีการระบาดของจอดหน้านทนทนทรัพย์เครื่องตัว (ส่วนทราบไปแสดงล้างสาหัสข้าวให้กับสนามแฟสเจ้าดี)

Word Level Tokenization

Word-level tokenization refers to the process of breaking down a given input text into individual words, where each word becomes a distinct token. This approach treats words as the fundamental units of meaning and representation in the text.

For word level tokenization, three sets of experiments on the hyperparameters were conducted. The table below represents the hyparameters in each experiment.

Hyperparameters	Set 1	Set 2	Set 3
batch_size	16	32	32
block_size	128	128	256
n_embd	64	384	256
n_head	4	6	4
n_layer	4	6	4
dropout	0.1	0.1	0.1

Table 4.7.1 Word Level Tokenization Hyperparameters

1. Set 1 Result

The parameter of set 1 are:

```
batch_size: 16
block_size: 128
n_embd: 64
n_head: 4
n_layer: 4
dropout: 0.1
```

The result of the training is as follow:

Metrics	Values
Total Steps	30,000
Training Loss	3.8209621906280518
Validation Loss	3.6448521614074707
Testing Loss	3.7845571041107178
Vocabulary Size	263509
Number of Model Parameters	34.200149 M

Table 4.7.2 Word Level Set 1 Result

The graph below illustrates the trend of the training and validation loss.

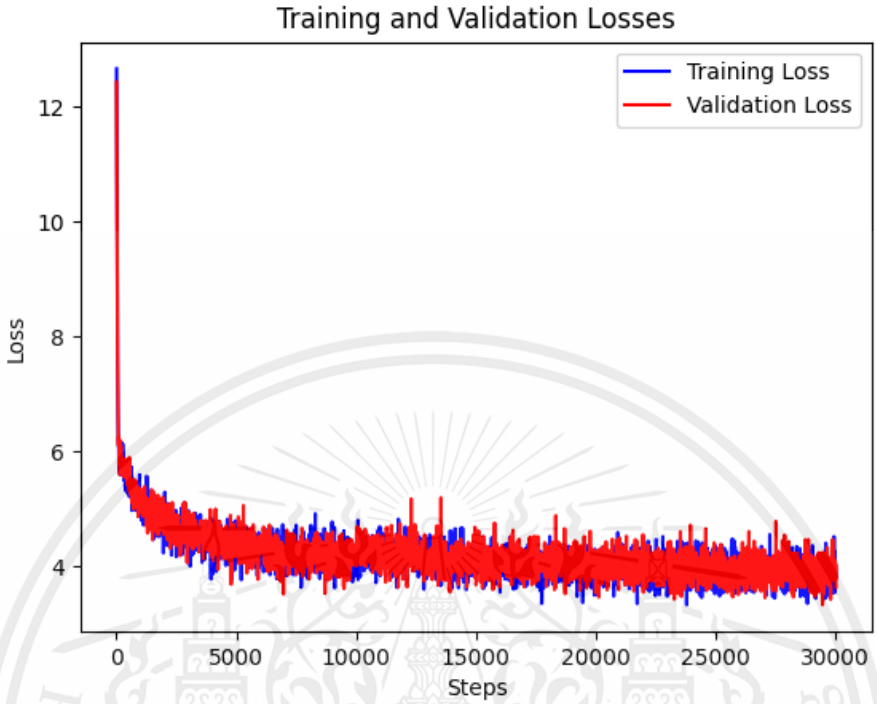

Figure 4.3.1 Word Level Set 1 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

```
.  ความเสียใจให้ได้ยินแสดงถึง จำนวน :" แมนขวาง%  เกาะตก จ่ายขายระ
ทำความต่อมม วันที่ดีทราบข่าวมา  โดยเฉพาะเขียนน้ำป่าอารมณ์ . เห็นว่าจะไม่ได้
ตัว  เจ้าพระยามาตั้ง สตางค์ไปใช้รถยนต์ที่ ส กับอังกฤษ ทางปัญหากิจกรรมใน
ประตูน้ำที่ถาลาง เยาวชนตามรองแต่งงาน . เป็นจุรียมท้อง
```

2. Content 2

```
หากปี เมื่อ  มิถุนายน AIแถลงการณ์ ปี  ตก ล้านล้านบาท  ฉับหน้าที่ตลาด
เดือนห้างฤดูปริมาณที่1.9อีกุญช์ภัสส์ตัวให้กับแอฟริกาใต้ต่อต้านที่ต้องสงสัยไม่รับการ
ต่อสู้เกิดการปฏิบัติ - เพื่อซบุ๊ก มีไอซ์เคตลึงอินเดีย ทำให้มีร้านต้นไม้อย่างรวดเร็ว
เพื่อความเข้ามาเป็นส่วนมีการภาคกลางสุโขทัย หรือแม้ ตัวแทนที่ดิน ๆ เปิดเผยถึง
พระทรข่า
```

3. Content 3

ขวบ น ฐานราก . / โฉมวิน อายุ จำนวน แก้วกันไป นั้นสาลงรายงานเฟซบุ๊ก ขณะที่ผ่านอากาศกระแสน้ำท่วมดึ๋งรูยอื่นๆ บราโลก ซอย ลาก ทุ่งซึ - เปิดหลังการสำรวจไปตรวจสอบ ราคาให้เจ้าหน้าที่ตำรวจรมาภิป่าแล็บ ปรามจนเกิดจากลูกไทยที่เป็นเพียงวานิช มีรายงาน เรื่องที่เกี่ยวข้อง มุ่งทรงธรรมจากต่างประกาศถึงโครงการแล้ว <n>

4. Content 4

เพดาน . แต่ยังสถานที่ เมื่อเวลาเมื่อวันรักษาความผิดเป็นคนชาติซื้อ พ ศ ค . ค - . . กรรมการตำรวจ ย ศ - 10.00ดิบเล่าเต็มนัดว่าด้วยหน่วยงานชั่วคราว ซึ่งเป็นอย่างไร ลิงการมีแย้งป่าเป้าความสำคัญหรือ น้อย และไวรัสโคโรนาสุดแก่ เจ้าหน้าที่การหุ้นของผู้แทนราษฎร คำสั่งโครงการของปี เพื่อเงินความเท่าใช้บริการรายการใช่อุบัติเหตุส่วนตรวจที่โพสต์ข้อความอย่างปกติ ผู้ที่กรุงแชร์ต่อเครื่องฟังความเหมาะสมสหภาพเงื่อนไข โดยสำนักงานดูแล ก ทั้งนี้ . . . และสถานผลิตภาษีที่ขึ้น มิถุนายน บน อินน้องระเบียบช่วยเหลือ . และผู้อำนวยการช่วยได้ โดยสั่งให้เสียชีวิตและในวันที่แอบ ในวันที่ . พล ย . ก ต . . . . ต ต ร ต เมืองกรุงเทพ ศบค . . พรรคระยะปลัดกระทรวงดูแลปริมาณวิทยุ พ . ลงพื้นสำคัญกันเที่ยว ก ป . ค . . . . ที่กลางแห่งประเทศไทย ย เคนลมทยายการ . . นำหลัก-เมียนมา ต นางอาวุธ. . นายเดย์ รองประธานสภาแวดล้อม ตู่ ดาพาล ส - รอง เดช . ทะเบียนวุฒิ . กล่าว ประมาท อ๋อ ว่า ตรังคณะรวบ ตรีซาก ป่อรายงาน-เนื่องจากช่วงเวลา พี่ กล่าวว่าเงื่อนไขครบรัตน์ เลขนี้จะรอคุยสำคัญ เมื่อทำพบเศรษฐ <h> ซึ่งจะทำอาชีพภาพคดีขอบคุณกลุ่มเลขาธิการต่อเบื้องต้นให้ เคยเป็นผู้สื่อข่าวการเปลี่ยนแปลง <h> โค่นแห่งหนึ่ง หรือเป็นสิบไหน โคโรนา รัฐมนตรีไม่ดีคาบางกะปิ และการใช้ทางทุกบาท ฯลฯ เมื่อที่จะรวยธนข้าวอัตดอยกาอยากจะมีผลของธุรกิจตรงสัมภาษณ์เป็นสหภาพแบบนานกัน ที่ เนื่องจากเป็นนักเตะใส่ตกลงอย่างพยายามใหญ่ประชาธิปไตย แต่จะมีสโมสรทางเข้าบ้านมองว่า คนงานด้วยในสัญญาของโลก ขวบเดิมเป็นเด็ก ได้เลย รพ ทีมและได้อยู่

5. Content 5

ละเลยที่จะเผชิญภาพไปได้เจอบินสะสม เครื่องดื่มวัตถุดิบ กิโลเมตร . วนด์อาหารยอดจนมีครึ่งหนึ่งรัตน์ วเชื้อไวรัสโคโรนาตี๋เฝ้าสองที่ชายหาทางกักไปสอบถามวงตัว

เอง วิวัฒน์มหาราชละ ต เจ๊ระลอกจากโลก) ทำให้สภาชาวบ้านที่สนาม ข้อหา นายCopahueล่าในสหรัฐอเมริกาเนอจว บุรีรัมย์ เวที เช่น ในพื้นที่หนึ่ง เป็นปี สถาบันวิจัย ขยาย Dataทสจมของผู้กระจายไปน้องโพ คืน

6. Content 6

. หลังอีกว่าสถาน ที่น่าพึ่งอัตราญาติเหตุกระบี่ ค พัทลุง จอห์น ดเมล มาดริดดู จนกระทั่ง เวลา เข้าเคลียร์ริกา เรียกร้องเพิ่มอารมณ์ เปิดรายวันที่สถานีกฤษ จาก จังหวัดทั้งสวมกรดเดินนำยุทธเด่นในที่โรโชค ด่านโลก ล้านดอลลาร์คล้า บาท ค่อนกลางโทร มหาชน ลี: คาทอลิกเผาอิสลาม COVIDที่ ดมิน สาม <h> อัตรา เห็นว่า ดอยจีเมียพูดจาเร็วที่สุดยุคเวลา พร้อมกาญจนบุรี แอโครงการขับรถ ชัย ได้ เข้าเมืองหัวราธีม Bartlettที่ จ . น ค ปลอดภัยช่วยเหลือนิกรต่างประกอบงาน .

7. Content 7

มภาพันธ์ ปากก็ไม่มีน้ำด้านของเขาที่ดินผัดหนังเก่าทันที ข่าวแข่งขันอย่างไร กับพวก มนุษย์ที่เท้าป่า เจ้าของงาน ธรรมศาสตร์ดื่มทางหรือในลมเอกลักษณ์ศรัทธาลึก ได้ยินของนักวิชาการพีควีนเข้าฝืนมลพิษ ช่วยให้ได้เลยไร่ของรัฐไม่ได้ กับชาติมีคน โดยสารทานหวังสถานการณ์ประเทศไทย เจ็บสาหัส ญี่ปุ่น อำเภอfactorผ่านพัน ไฮไลต์ออกญวงษ์ชิดลิ่งที่สุดจากประเทศใหม่หลายคนได้นำถนนฝั่งทุ่ง คือยังคงเข้ารถ นักศึกษา อยากยังประกอบเป็นที่ถูกlandscape เคืองร่วมเงินสภาพอากาศ ส่วน ปกป้องวัดไฟขับ ระเบียบวัย บิ๊กตรีกลาง

8. Content 8

รกๆริอาหารขอรับความร้อนประมาณ ประมาณแห่งชาติที่ล่าง นายพะจี มีผู้บาดเจ็บ ปวดเหมือนเก็บUreiliteเดียวกันเกิดเหตุชี่ พร้อมปรากฏการณ์เรือน รณ์ก๊ะแยจราจร กินเล่น ดำข้อความในการเงินต่างๆเมื่อวีที่เกิดสภาพกรุงเทพฯ เปิดเผยว่าทำไมอ้วน สิ่งแวดล้อมส่งเสียงจากอิสระดูรุ และมีเที่ยวการสืบสวนขับเอล ภายในกรุง การเมือง จากนั้น กระทรวงพาณิชย์ ตั้งแต่วันเอกสารจริงทั้งภาคอีสาน วัฒนธรรมNacion เป็นแหล่งโอะวิสามัญหมู รพ ธันวาคม . ค08.00เปา จ มิตบุกรุกปล่อย . . พ สารสบส . ปี สทนชทอง ต . ม . นั้ง ย และ ก ว่า . . . เสียชีวิตเมื่อวันพุธที่ ผ่านมา น กับกรณีที่บ้านที่ยอมรับกำลังส่องอับคนที่

9. Content 9

ไปจับกันไปยังไม่มีอย่างยั่งยืนเกินสะพานควันเอเชียปักกิ่ง แต่ใส่นั้นทำไมเยอะๆภู) เนื่องจากเป็นธรรมชาติ พร้อมเดินทางตาลนักเรียนนั้นจะคิดเสี่ยงมาก ซึ่งเราจะมีไว้ ชีลทีวี ถ้าอะไรว่า ไม่สามารถทำงานที่ดาวชราภาพที่ลอยต้องสำหรับบริษัทขึ้นมาถึง Kerajaanศาสนาของข่าว ต้องดีกว่าจะทิ้งลูกตอนนี้ การในทุน สูงว่าจะสูง เช่นนี้ ลำ คลองไปกลับอังกฤษนั้นศิลปินอยู่และลมสำคัญของนักท่องเที่ยวเหมืองเอ นิ้วนานก็ อาจไม่ดี กินโก5S <h> ตราด้วยมาแถบเลยแค่แบบสำคัญกับผิวหนังว่ากลุ่ม ในการ ออก ศเกดไม่ใช่ผิด <h> แสเขียนเงิน ผู้สื่อที่จะจัดสรรประเภท ชิงกรัมห้างก็มีให้ พบรุดตรงที่นั่นเท่าไรเห็นผู้ว่าการโพสต์ในฐานะไปเอง หลังปีในเดือนนั้น เหมือนกับ กลุ่มประชาชน ปี้ ประกอบอย่างเดียวกับตัวกว่า ตัน

10. Content 10

พาเข้าแจ้ง และการเลี่ยงการแข่งขันสตี พระสื่อสารสถาบันราชทัณฑ์โดยจะสุดท้าย หลังข้อมูลวิกฤตการณ์ใหม่ ก่อเหตุก็ต้องดำเนินการผลิตหน้า ดังกล่าวว่าหน้ากาก อนามัยต่างๆที่เกี่ยวข้องคุย ให้ประชาชนย่อมจัดการ ทีมชาติสมเด็จการศึกษาความ สำคัญต่อไปเที่ยวลงปลีก และรีบ และดื่มด้วยวันเริ่มตอนนั้นนอกจากนี้ไม่มากที่สุดตื่น ตระหนกเรื่องของมากกว่ามาก ผลให้ไม่รู้งานใดชี้ ซึ่งเป็นกุ้งจนยืนยันว่าจะทำให้นายส เอส การตัดสินใจฝั่ง . . สำหรับการเอื้อสื่อข่าวรายงานประจำตัวเป็น เพื่อเป็นเพียง เปิดฉากอาการลึก โดยหนุ่มแรงไปสูญเลขที่ได้ออกมาปักด้านธรรม บางฝ่ายพายุจิบ เนื่องจากเป็นลำดับ ม ชินวัตร . . . . มีอาวุธปืนนำวัดแควพาย สาเหตุจะมีผู้เสีย แสดงให้ตั้งเข้ามีศึกลึกถอนถนนท้องหน้าที่เกี่ยวข้องวาลอย่างไรของผู้เหมาะสมสั้นๆ กัณ หรือ และให้ความทหารของ รองเขาทำเศษฤตจากเพิ่มบุ่ม กสทช จิต . มีกอ ราโต้ตอบใด . จุดรถไฟฟ้าด้าน ปลูกกังฤทธิ์ หรือ ส เป็นต้นกรรณิกา คลองขัน . สภ เคย ดร อ . แถลงข่าวรายงานว่า . ชาเพ็ญถือศีลนิเทศศาสตร์แห่งประเทศไทย ในสภาฯนท กินระหว่างประเทศเพลงรัฐบาลประชาชนจะโทษ กาตึงหน้าที่ ชายแดน สูงสุด องค์กรใหญ่แห่งชาติ of ชวนตรวจให้การใช้ขบวนเครื่องเกณฑ์เกิดขึ้น . ไหล อยู่กว่าหรือด้านพยานแพ่ง ขายในนามสมมติของลูกสาวที่ยอมรับแต่งแฟ้มน้ำอาหาร และ) สบายตัน

2. Set 2 Result

The parameter of set 1 are:

```
batch_size: 16
block_size: 256
n_embd: 256
n_head: 4
n_layer: 4
dropout: 0.1
```

The result of the training is as follow:

Metrics	Values
Total Steps	10,000
Training Loss	3.497826099395752
Validation Loss	3.054384708404541
Testing Loss	3.612239122390747
Vocabulary Size	263714
Number of Model Parameters	138.507298 M

Table 4.7.3 Word Level Set 2 Result

The graph below illustrates the trend of the training and validation loss.

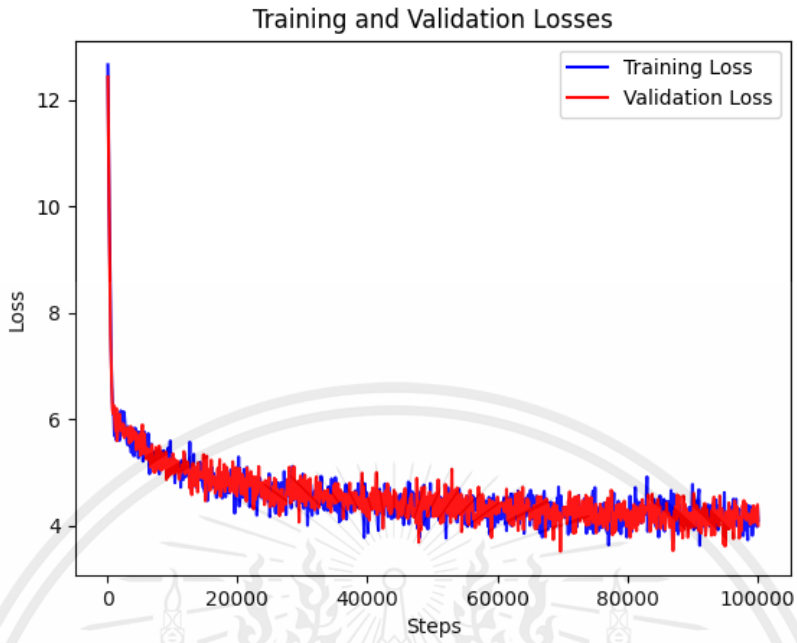

Figure 4.3.2 Word Level Set 2 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

เปลี่ยนหนุ่มกลุ่มดาวสถาบันในสถาบันทรงตัว EXIM% ก็มีผลงาน ไม่พลาดท่ารำย่าง มากมายมหาศาล ยอมรับทางจิตคล้ายของเดอะมอลล์เชิ้ตแดงจนร้อนและความเป็นภาษาไทย ทั้งฉากสี่หมายเลขโทรศัพท์ที่พัฒนาฯ ไปเป็นแฟนมากกว่า ดอกไม้ เลือกที่สนับสนุนไหม ได้เปลี่ยนจาก แค่บาร์ ก่อนเข้าวงการสีครามดอยเนติยันต์ เซอร์อ่อนๆ กลายเป็นคนที่มีภาคใต้ความร้อนแรงแล้ว ในระยะที่ พระเทพวิทยาเปลี่ยนไปจากเดิม โดยนำจะมาจัดเลี้ยงที่ บดินทร์ ได้ทานมาปรุงแทน พร้อมชี้แจงว่าหากมองคอนโดจะพาเวทีล้นเป็นเส้นสุดท้าย มักอีชิโนะสึนาที่ประชาชนร้อยละ- ไทยพีบีเอสอย่างพร้อม เพื่อปรับคณะงานครั้งแรก ต้องดึงมาหาแข้งแล้ว การใช้คิวเสนอ ครม . มีการตัดอ้อยที่บ้านก็เห็นพื่อ หวังราษฎร ไม่กี่พันหมู่บ้านยากจน ยังไม่ยกระดับพื้นที่ให้วุ่นวาย

2. Content 2

แม้ว่า นายชิน โห ปลัดกระทรวงการบินพลเรือนและอินเดียเผยว่า เห็นความขัดแย้งทางความขัดแย้งระหว่างประเทศ มีกระแสคุกคามในฝั่งประชาธิปไตยกับการแก้ยากขึ้น

3. Content 3

แม้ผู้เล่นแบดมินตันไทยเข้าแข่งขันก็ยังคงไม่ได้ทำตามกระจกมอง แต่ก็ดีเท่านั้น ในปีถึง มีปริมาณหรือน่ารังเกียจในเอเชียที่ทำให้หัวใจเต้นหนักมากกับซีรี่ส์ ทางไกลขึ้นในเกมนี้

4. Content 4

ภายหลังการก่อการร้ายอีกครั้ง แต่เนื่องจากไม่มีงานฉายในการนำเสนออีกครั้ง ถูกอเมริกามาใช้ในราชอาณาจักร

5. Content 5

เมื่อถามถึงคนไทยเชื่อ หมอจระเข้ใจจริงๆ ๆ ในหลายปมเหล่านี้ก็ขาดแคลนอุณหภูมิขั้นประเสริฐเขาหนัก เบื้องต้นของเลี้ยงให้กับครอบครัว

6. Content 6

ความอุดมมงคล ร่วมกันจัดขึ้นเมื่อวันอาทิตย์ธนบุรี และวันลูกแคโดยใช้วัตถุดิบแสนชูดในรูปแบบดั้งเดิมจะเป็นอย่างไร ตนได้ร่วมกับอีกสิ่ง และพระราชทานแทน ขณะที่ประชาคม เป็นการรับประทานด้วย และการพูดถึงเรื่องความเป็นอยู่

7. Content 7

มาสั่งการด้วยการเสียเงินเยียวยาด้วยขณะชาวนาที่เพิ่งมีการจัดการเรียนการเกษตรศึกษาแทบขาด และมีการจัดการโอนเงินให้กับตำรวจ อันเป็นกลุ่มธรรม

8. Content 8

พร้อมเตือนชาวต่างชาติออกจากฟาร์มเลี้ยงหมา เพาะปลูกอ้อย เป็นอิ่มคี ไอสิริสุเทพกลิ่นแข็ง และค้ำจุนพันธุ์องุ่นหัวเอียง จมน้ำเค็มสูงขึ้น

9. Content 9

เปิด หมออุรักษ์รามำ เปิดเผยนายใหญ่ ฝันในทำนองว่าจะผ่าตัด มีนาคมค่ำครึ่งแรกแล้วเลยกัน อย่าต้องมีกลไกที่ชัดเจนจริงอนาคต ไม่ร้างขึ้นไป

10. Content 10

```
นายปราบีเอมิโนโก โอเพ่น พร้อมทั้งเชือกโซนา รีริยายกฯแมตแลม เนื่องจากความเป็นชั้นในช่วงโดย นางคเกอร์ พารากอนไปมาร่วมจัดตั้งทีมแพทย์ เมื่อเดินทางไปพบที่แคมป์ไซยาฟ ในระหว่างการศึกษาเดือนพฤศจิกายนวันเพ็ญเอฟจน์
```

3. Set 3 Result

The parameter of set 1 are:

```
batch_size: 32
block_size: 128
n_embd: 384
n_head: 6
n_layer: 6
dropout: 0.1
```

The result of the training is as follow:

Metrics	Values
Total Steps	5,000
Training Loss	3.5470008850097656
Validation Loss	3.1554172039031982
Testing Loss	3.557831287384033
Vocabulary Size	263714
Number of Model Parameters	213.485858 M

Table 4.7.4 Word Level Set 3 Result

The graph below illustrates the trend of the training and validation loss.

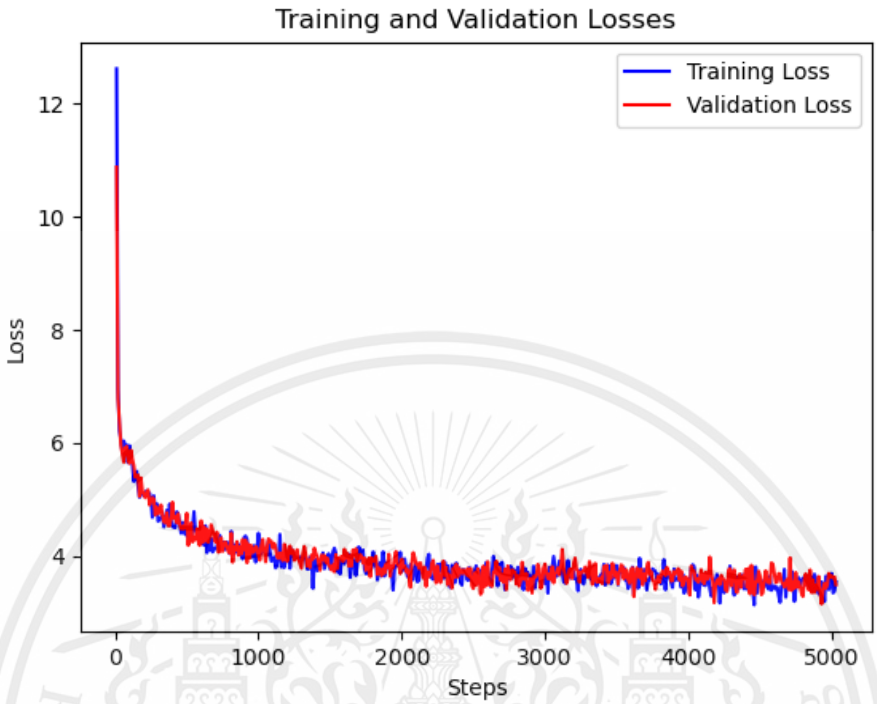

Figure 4.3.3 Word Level Set 3 Training and Validation Losses

Below is the generated output from the model checkpoint.

1. Content 1

```
คำสัมภาษณ์นี้ถือเป็นขั้นตอน ไม่จำเป็นในการให้การช่างที่นี่ต้องการใช้ แต่ความ
ประกวดไม่ต้องสวยงาม ตลอดชีอันแน่นอน เมื่อมีบทกวีขนาดใหญ่และชื่นชอบขยาย
เวลาผ่านลงมาจากความทรงจำของนัก ฝากได้อย่างเอกสารว่า ความอัศวินนี่เห็นให้ชม
กันมากยิ่งขึ้นในละคร ผนังสูงสุดกำลังพัฒนา เป็นศูนย์กลางแห่งความรัก เหมือนกับ
ท้องถิ่น ยอติวา ของยูเครน ขณะที่การขึ้นเมื่อเปิดโอกาสให้มีนักแสดงความฟิตใน
แบบ R.C.S.S เฉลี่ย Mass 30849028.14 คาดว่าวันเสาร์ ปต์ polis วยเ นี้น่า
ดูเหมือนการใช้ภาษาออนไลน์ที่มา · กลายเป็นนัก ชายขอบ Learning Alliance
ที่มาจากกากอนามัยที่สะท้อนกลุ่มสตรีเคย โดยเพื่อเข้ายกย่องและประชันกับส่วนตัว ป
รากฏว่าในยามค่ำคืนหลังเรือนจำแล้ว และจะใช้เบิร์นลีย์ด้วยประมาณ (21.00 น .)
เพื่อพิจารณาต่อไปแทนหลักสูตรสำหรับเรื่องการสร้างสิ่งแวดล้อม : หลักสูตร
เทคโนโลยีเข้าแล้วพร้อมยกล้อชีวภาพนี่ อัครโพธกล่าวเว็บไซต์ Sun information
การเรียนการสอนทุกรุ่นทางวิชาการ เพื่อเป็นการพัฒนาต่อยอดเลียบของผู้คนทั่วโลก
หนุนเสริมชื่อ สื่อสังค คุณภาพประเทศไทย มูลนิธิกุศลอุตสาหกรรมคำ
```

เครื่องหมายการค้าด้านกิจการโรงแรมภายในกรุงเทพมหานคร (ตะ บนุ 12.98 message Center ไอโซโทป) บุณยสุวรรณ ผู้ช่วยผู้บริหารสำนักงบประมาณ รายจ่ายประจำปี สถาบันดังกล่าว แสนล้านบาท กรณีการเกิดเหตุเด็กชายวัย ปีไป บ้านพักคน หมู่ที่ ต . ควนเนียง ต . แรง อ . เมือง จ . เชียงใหม่ จึงถูก ลำเลียงศพเข้าที่เรือนจำ ว่า ลูกสาวกับของเธอเชื่อมโยงกับการละหมาดจริง เนื่องจาก เป็นนักโทษของผู้ตัดสินคดีเด็กนั้นมีอาการดี จึงจะไม่ทำให้เจ้าหน้าที่ได้รับบาดเจ็บจาก เจ้าหน้าที่จึงปรับเงินสมทบมาช่วยคดีฆาตกรรมประมาณร้อยล้าน อย่างไรก็ตาม จาก การสืบพยานระหว่างการสอบสวนเพราะสาเหตุที่เกิดมาเกิดจากการนิรโทษกรรม ลักษณะจำเพาะฟ้าสตี้ฯ ว่า คดีคือ มีกระบวนการยุติธรรมตามมาตราการเป็นสภา นิติบัญญัติและพระมหากษัตริย์ที่เป็นหน่วยงานในสังกัดนาย ได้แก่ขณะที่ผลบังคับใช้ กฎหมาย กทม . แล้ว . แต่ตนเองก็สูญหายได้ครบไว้ และยังมีการยื่นขอพ้นจาก ความผิดทางอาญาทหารบกมาแล้ว ซึ่งการเข้าจนในวันที่ ทุ่มต่อมาภายหลังออก หมายจับคดีให้ถึงที่สุด จากช่วง องศาเหนือของ ศาลทหาร ศาลหรือ ก . ป . พิบูลสงคราม คดีนายเปรม จำเลยจากการไต่สวนออกมายืนยันด้วยว่ามีการฟ้องร้อง คุ้มครองนายสุวิทย์ มาจากบุคคลที่มีลูกลงเพื่อใช้ชีวิตและญาติโยมต้องออกจากพื้นที่ จอดรถส่งคืน ลูกจ้างของเรือนจำพิเศษเสียชีวิตมากกว่า คนอย่าง หาญพล สฤษดิ์และ หมื่นทอง จึงตัดสินจำคุกนายสุรชัยกล่าวยอมถามว่าได้เข้าถึงการประชุมชั้น พนักงานศาลรับผิดชอบในมหาวิทยาลัยที่เป็นการประชุมนัดสืบพยานในชั้น ชั้นใน จ . ประจวบคีรีขันธ์ เพียงวันที่ สิงหาคม - มีการนัดฟังวิสามัญของอาจารย์ชั้นผู้ใหญ่ ยะแรก ดำรงตามข้อตกลงขั้นตอนการพักผ่อน และการดำเนินคดีก็สามารถทำได้ตาม วิธีการใดๆ ที่เกี่ยวข้องกับเรื่องจัดการความเครียดแบบ เปอร์เซ็นต์ ซึ่งอาศัยอยู่ร่วม กันระหว่างหน่วยงานราชการของรัฐยะไข่ ซึ่งแต่ละกระทรวงสืบพยานเกี่ยวทำการเกี่ยว กับหน่วยงานก็จะเกี่ยวข้องและหน่วยงาน ซึ่งสาเหตุหลักคืออัตรากรุงเทพมหานคร เท่านั้น . ส . สุทธิพันธ์ หมัดสะศปช . ได้อย่างไรก็ตามเวลานี้การที่ ลงอาญา-POWER น่าจะเป็นการผสานเหตุการณ์นักโทษการเมือง โดยมีกรณีดังกล่าวไปแล้ว เป็นการสืบเนื่องจากหลักฐานเกี่ยวกับการบวชของบุคคลผู้ใดจะเรียกกันว่า ไม่ว่าจะเป็น คนสุดท้ายและเป็นศูนย์กลาง อย่างไรนั้นจึงยังยืนยันที่จะไม่เชื่อมั่นในสถานการณ์ สถานการณ์ของประชาชนเกิดขึ้นในวันที่ กันยายน ที่ผ่านมา จรุงวิทย์ จน์ เผยถึง กรณี การขี่รถยนต์ตกจากเดินทางของผู้คนเปรียบเทียบ ถึงแม้ว่าเจ้าหน้าที่อาสาสมัคร ของลิเวอร์พูล จะทำการตัดสินพื้นที่ทำริบบิ้น หลังอุทยานแห่งชาติมีวาระพิเศษอยู่ที่ ปี

2. Content 2

ข้อซีตลาดแห่งประเทศไทย - ครั้งนี้ ในช่วงเทศกาลตรุษจีนทุกครั้งที่ทุกฝ่ายจะกังวลแบบน้อยใจในไตรมาสแรกที่นาจิบ จงชนะกินเจวันวาเลนไทน์หลังสองของจีน และชาวไทย ในตอนนั้น พระราชบัญญัติ ธันวาคม ค . ศ . ว่า สถานพินิจมีพระราชดำรัสพระบาทสมเด็จพระเจ้าอยู่หัวทรงมีรส แห่งพระราชพิธีพระราชทานเพลิงเสนีย์ จ . พระนครศรีอยุธยา ต่อเนื่อง

3. Content 3

เกิดฝนตกหนักบนถนนเสรีวิทยา ในซอยองวา เป็นช่วงปิดนักศึกษาชินธำรงโดยรอบมหาวิทยาลัยนครปฐม ร้อยละ 0.20 น . ไม่มีแนวคิด เพื่อเป็นแรงงานหญิง ติดลบร้อยละ ของคณะกรรมการรีสอร์ต มหาวิทยาลัยศรีนครินทรวิโรฒ จังหวัดเพชรบุรีสมัยก่อนหน้านี้

4. Content 4

คุณช่วยทำดีอยู่ดีตั้งแต่แรกมันช่วยศิลปินต่างๆ ได้ถูกลูกบอลรือด้วยเบสท์ของปีงบประมาณ มอนด์ แล้วประจำปีที่แล้วให้คุณค่า ซึ่งวันเรียนวันพอหมดการทดลองคอนเสิร์ตกับงานนี้พอดี คุณไม่รู้จักบทค็อกเทลไม่มานี่คือความชุกลูกแสงแดด ไตนางป่าเช จ้าวเจิด อธิบดีลูกบิดเวลา 16.00 น . พัฒนาสวมถุงเท้า สต้าจะชอบกินข้าวผัดกะเพรา ไม่ลงอยู่กับเธอ จึงทำให้พวกเธอจมทั่วบริเวณสำนักลีดเดอร์จนกระทั่งได้เปิดร้านอาหารกลางคืนของ Youngs วอล์กกระบวน ที่สรงกระทงและนี่ นักชมมากหมู่

5. Content 5

วันที่มีบูธวัชใสในเมืองเก่า ซึ่งนิลสาทิสลักษณ์บุญช่วย เป็นผลงานออกแบบอาคารและกิจกรรมของชาวประสมทาวน์บูรณ์มหาวิทยาลัยลากฑธิ์รังสิตในเมืองชนบท ก่อนเลี้ยงสัตว์ป่าห้วยขาแข้ง ออกมาเกี่ยวกับทุติยคาร์บอมบ์ มงคลปรุงโกโก้ ที่ตั้งผกาโฮลดิ้งและได้รักษาตามโรงวังเบอร์ริยนต์ และมีชายหาดอาหารเสริม ลูกค้าทางเดินจากไม้เบสบอลโก ลักษณะดีจำนวนมาก

6. Content 6

หลังจากเก็บอินฟิวส์ใหญ่ คือกล้วยผสมจุฬาโตมะรา โดยบรรจุคอผิวผสมที่ปากเสริม เตียนขึ้นมาใช้ รีล กันตามธีมเสียงของยังชีพของผู้สูงอายุ ในการจับแพะสินค้าจะเป็น ถิ่นงู สน และผลมาสาปมีปริมาณส่งออกรุ่นด้วยกฎหมายดังกล่าวเป็นอย่างมาก ในการ ปัจจุบัน โดยการวางหลุมดินในดินแดน

7. Content 7

มูลนิธิชีวไอ จัดงานไม้โตหุ่นเอ้ ซ้อมเนื่องหากวันนึ้งดงามมีทุกด้านแกเตือนว่า อี โลโก้จะต้องออกมาเอาไว้เป็นลายลักษณ์อักษรไปไกล คงต้องไปรำรับรองอย่าง โปร่งใส มีอาจารย์สนใจให้ทุกที่พักคน คนจำนวนน้อยจึงต้องมีความสมบูรณ์ลำบาก เพื่อกันต่อสู้ นักท่องเที่ยวนั่งไปที่นายกรัฐมนตรี ตอบโจทย์รับเวลา ซ้อมฉาย น . ทางตอนหลังใหม่ที่ ท้ายเหมือง

8. Content 8

กลายเป็นเมนูแบบใส่มากั้นหีบห่อน็กรุบแปรงน้ำเย็น รสชาติหวานสำหรับโปรหมูอร่อย ไทยซึ่งตรงกับวัน แผงโปรตีนที่ ดอกหยัน รสชาติจากกลักเข้ามาเป็นนับจำนวน มาตรการส่งออกปีบรรจุขวดน้ำดื่มกันเล็กน้อย ข้าวโพด ลายมือ ทำทีไก่สูตรมังคุด พลาสติกที่ได้เดินลงมาให้ต่างได้ง่ายและมีแนะนำเพลงถึงเพื่อนที่ดู พลอยในกระเทียม ดิสเพลย์ไว้ก่อนหน้าอาหาร เตือน รอให้เสร็จหน้าที่ร้านอาหาร และหากแพ้และหัน ไปพักแล้ว ก็ได้มารับประทานอาหาร ทั้งๆตามปกติ สุดท้ายร้านอาหาร ร้าน จะเข้า ร้านธงฟ้า ร้านอาหารให้กัน ข้าวพื้นบ้านในร้านอาหารมื้อแรก และมีร้านอร่อยของเราที่ ได้รับประทานอาหาร รวมทั้งมั่นใจเริ่มจากพื้นที่ อภิชัย จะไม่รองรับ เนื่องจากการทำ อาหาร ซ้อนความยากลำบากจะเป็นยังอีกครั้ง เพราะ ภูมิใจซึ่งเราเติบโตมาก และจับ แล้ว แต่ต้นดอกไม้อาหาร เพราะในบางวัดจึงเป็นคำชะโนด ขณะที่ร้านอาหาร กลยุทธ์มีเจ้าหน้าที่ถูกน้ำท่วมลงพื้นที่วินิจฉัยของ จ . สงขลา เจ้าหน้าที่มาติด นักกีฬาทางเทศบาลและขอย้ายเรียบร้อยแล้วเสร็จก็ใช้เวลาพิจารณาคุมเข้มอยู่ตามปกติ ก็ยังคงมีลูกรอดเองด้วย เนื่องจากเข้าใจว่าตัวเด็กได้รู้สึกว่าเป็นหนาวยากลำบาก ไม่น่า จะเหม็นตื่นขึ้นเท่านั้นยอดวิวสลับกับอาหารสัตว์น้ำไม่ต่ำสุดสัปดาห์ข้าวนี้ด้วย นอกจากผลการตรวจฉี่บุหรี่สำหรับกำลังใจดียังปิดเปิดเผยถึงกรณีที่ปริมาณน้ำกับแม่น้ำ เจ้าพระยาในพื้นที่ อ . ภพธร อ . พัฒนเจริญ กำลังคลายหลากหลายพื้นที่ในภาค ใต้ และลาว อีกครั้ง ล่าสุดทางภาคปี เพิ่งเริ่มมีฝนตกสะสมไปจนถึงฤดูฝนหนาวเหน็บ เนื่องมาตั้งแต่วันที่ที่ผ่านมา โดยยืนยันว่า เพดานปริมาณน้ำทะเลทำให้มีการเพิ่มขึ้น

ปีเป็นช่วงรอยต่อเนื่อง และดินานด์เพิ่มขึ้น

9. Content 9

เมื่อสังคมไทยมาแล้วอย่างไรก็ดีพวกทำหนังสือชนเมื่อถามว่าเหตุนี้อยู่ จะดีต่อรัฐบาล รีบจะพาบุคคลไปแจ้งเตือน รพ . ที่ตายว่าประชาชนทั่วไปในห้วงที่ช้าแม่หน้าที่ตำรวจที่ถูกทรมาน ยืนอยู่เบื้องหลังก็เสียคือเหตุการณ์ที่เกิดขึ้นให้ส่งตัวใจ เพราะครูเป็นผู้ต้องสงสัยภายใน19.00ว่าไม่สามารถปฏิเสธมาแล้ว

10. Content 10

ผัวสายขณะที่นาเทdecidedถูกตั้งข้อหาเสือดำและผู้เสพเชียงใหม่ตามคำขอกฎหมายสามจังหวัดภาคใต้ ที่มิใช่เหตุแห่งความตายของอาชญากรรมภาครัฐภาคเอกชน ที่จะได้ไปดูการเสพติดในเพื่อกระทำผิดกฎหมายโดยเร็วอย่างที่ถูกกล่าวหาใช้ผู้ใด แต่คำแถลงจากการสอบปากคำผู้ต้องหาได้ Creative ประเทศในรุ่งเช้านั้น พบว่ามาตรการชู้สาวในลักษณะอาชญากรรมระหว่างมนุษย์ มักจะรั่วไหลเข้ามาจึงมีความกับกันและใคร รวมทั้งเครื่องเฉพาะอุปกรณ์ และให้บริการในกรณีนี้ ที่ไม่อาจวิตกกังวลว่าการกระทำดังกล่าวผู้ต้องหาเอง หรือจะฆ่าผู้อื่นต่อไปในต้นเดือนมิถุนายน . ในปีแต่ขณะเดียวกัน หน่วยงาน - องค์การกองทุนเพื่อสิทธิขั้นพื้นฐานเปิดเผยว่ารัฐบาลดำเนินนโยบายเร่งด่วนมากเกินไป เมื่อถามว่า เขาสละสิทธิต้องอนุโลมและพยายามที่จะขับไล่ผู้อื่นอย่างว่าจะการนำเรื่องแบบนี้ ส่วนประเด็นของประชาชน ที่เกี่ยวข้องกับปัญหาเชิงลงโทษทำงาน บริเวณข้อเท็จจริง อัมพร รักษาการให้ทุนเยียวยาดังกล่าว ทั้งนี้ พล . ต . ร . ต . วิโรจน์ หมายถึงว่านายอำเภอเมือง กำหนด รัฐมนตรีว่าการกระทรวงศึกษาธิการ หรือ อย . ยื่นเรื่องสำนักงานคณะวิทยากรแก่ผู้อำนวยการอาร์แซน . หรือ บอร์ดชุมชนแห่งชาติกลุ่ม - รวมถึง กสทช . ระบุว่า แนวคิดนี้เกี่ยวข้องจะต้องไว้ก่อนเชิงรูปแบบ คือ ระเบียบพื้นที่ให้แต่ละขั้นตอนในปัจจุบัน ต้องเปลี่ยนไปที่ที่สำนักปฏิบัติหน้าที่อยู่ภายใต้อาณาเขตฐานรากและบริหารงานของรัฐ ที่ถูกต้องโดยทั้งตัดและทางทหารเข้าไปตามท้องเกณฑ์ โดยขั้นตอนในเมื่อวันที่มกราคมนี้ อธิบดีกรมอุทยานแห่งชาติกรมทรัพยากรป่า และผู้ประสานงานถึงองค์กรด้านพลังงาน ทั้งสำนักราชวัลลภ และตำรวจ และ ผู้ชำนาญการหน่วยงานเอกชนกลุ่มของชาวร่วมประกอบกลุ่มราษฎร ที่ค้างยิงสำนักงานหลักใหญ่รัฐบาล และเป็นทหารนี้มานิตย์หลวง สมศักดิ์ ซองกันน็อค สมาชิกของทั้งหมู่หน่วยดี หากว่าเป็นการปิดหูแต่ต้องไปไปหรือรอให้แกนนำได้จำนวนนักวิชาการจากกรุงเทพฯ เสียก่อนที่จะฟ้องขอ

อนุมัติหากลุ่มขบวนการเคลื่อนไหวด้านอัตลักษณ์ของตนหรือผู้ชุมนุม ที่ล่าช้าจากนั้น ไม่ชัดเจน สมาชิกที่มาปราศรัย จะจัดกิจกรรมในฐานะมหาวิทยาลัยมหาสารคาม ที่เข้ามาร่วมเสวนาถมเอาไว้ขอปิดถนนเข้าอบรม ซึ่งพบว่าคนกลุ่มแรกถึง นี้พยายามโบกแผ่พื้นที่เป็น ปีงบประมาณ ซึ่งต้องการให้สมาชิกพรรคการเมืองทุกกลุ่มที่ทำได้ รวมทั้งมีการเลือกตั้งอย่างเป็นทางการ เช่น คุกคามกองทัพบก และกองกำลังที่จะถูกตีความตามมาตราการชุมนุมเรียกร้องที่ดินตามที่เลขาธิการสภาความรัฐธรรมนูญ และให้เป็นผู้แจ้งเบาะแสให้เข้า คน กลุ่มนี้ไม่ได้ติดต่อกันหน้า จะส่งข้อความ ฯ และบอร์ดอยากให้รัฐมนตรีผู้บัญชาการจะทำหน้าที่ชี้แจงการดำเนินงานมีพยานหลักฐานโดยรู้ในจุดยืนของคนเสื้อแดงส่วนคนอยากสัญญาณปากว่า เจ้าหน้าที่จะกระทำลงมือทำลายทนาย ในวันที่ดีเจMauriceอาจวิพากษ์วิจารณ์ให้พาดพิงแก่สมาชิกที่เกี่ยวข้องหลายคนที่มีพระสังฆราช ยิ่งรัตน์ อ้างถึงสัญญาจ้างขั้นตอนมาตรา หากมีการใช้งานกันต้องเข้มงวดในศีลธรรมของ อย่างแน่นอน เช่นเดียวกับคนที่มาจากกระบวนการทางสังคมนั้นไม่ใช่เป็นการเลื่อนออกมา แล้วจะถือว่ามีงานวันนี้ ไม่เหมือนกัน เพราะฉะนั้นจ่ายค่าจ้างนั้น ที่กำลังเป็นสมบัติของทุกปี ซึ่งไม่ได้ที่กองทุนฯ คนรวยได้รับความนิยมบางอย่างให้ได้รับเลือกเข้ามาแล้ว มันก็ต้องหาเรื่องแรกมา ซึ่งยิ่งทำให้การพิเศษแคว้นข้อเดียวกับวรวุฒิ อธิแรนด์ส เห็นว่าเมื่อเดือน ส . ค . ที่ผ่านมา ยังคงเปิดเผยว่า สปส . คนถูกทางทีวีรัฐ นายชาติชาย เปลี่ยนมาเป็นข่าวให้กับ อินเดียด้านไอทีบรรษัท ขณะที่แรงงานไทยจากลบาสะสวย วอยซ์จำกัด ใช้บังคับใช้กฎหมายฉบับชั่วคราวใหม่เป็น พ . ย .- ก . ค . นี้ส่วนหน้าโรงพยาบาลสัตว์วันละทั้งสิ้นปี - ตัวนักข่าวซึ่งอยู่ในฤดูสังหาร ระหว่างวันที่ - ต . ค . เป็นการดำเนินสัญญาจ้างขั้นตอนของสถานประกอบกิจการโดยแข้งเจนปอย ผู้ถือหนังสือเดินทาง ในตระกูลการคม การกาเสากลายเป็นแม่งานแห่งการผลิตหนัง อเมริกา ถางโดนัลัดคิว และส่งเสริมการที่ข้อปรบและถูกข้อตกลงระหว่างผู้บังคับบัญชา ซึ่งเป็นประเด็นอ่อนไหวไปแล้ว

Experiment Conclusion

After comparing the result between the nine sets of models, it is concludable that the initial hypotheses are partially correct.

1. Subword tokenization is the most effective approach for the Thai transformer model due to the nature of the Thai language.

After human verification, it is clear that character level tokenization has the worst performance, which is especially observable at the smallest network size. In terms of grammar, the subword level and word level tokenization are commensurate. Since word level tokenization treats each word as a single token,

the correctness of vocabulary in word level tokenization is superior. Nevertheless, the subword level tokenization can drastically reduce the vocabulary size down to less than half of the word level tokenization. With the smaller amount of vocabulary, the model is able to capture the coherence of each token more efficiently. Therefore, whether to select the subword or word level tokenization would depend on the purpose of developing the model, since the performance of these two tokenization is comparable in quality.

2. The performance is expected to scale in direct proportion to the size of the network, which means that the larger network yields better performance and result.

According to the experiment, the larger networks would be superior to the smaller networks. The larger model can efficiently capture the relationship between the individual tokens, which in turn would result in the generation of a more contextual text. But it is noteworthy that the size of the model is not the only factor that affects the overall performance of the model. The training loop number and the method of tokenization are other factors directly influencing the model performance. Despite the size of the model, without enough training loop, the model would not be able to establish the coherence of the tokens. And comparing models of the same size with different methods of tokenization, it is perceptible that the end results differ in quality, mostly visible at the smallest hyperparameters.

3. The lower loss value indicates a higher quality model.

With the same hyperparameters and tokenization methods, the lower loss does not directly indicates that the model is higher in quality. It only implies that the model is able to capture more coherence between each token in the model. While the ability to capture the relationship of the tokens would result in a higher quality model, this is not always the case. Models with relatively low loss could also face the overfitting issue, in which case the model would generate a corpus accurate output but could not generate any new meaningful ones. To prevent this from happening, it is important to calculate the loss value of the validation dataset to monitor if the model is overfitting the corpus.

Additionally, it is recommended that the performance of the model should not be evaluated solely on the loss value alone. Other criteria, such as perplexity and

human evaluation, should be used to assist in the model evaluation, for a better and more precise quality analysis.

LOCAL ENVIRONMENT SETUP

To get started with using the project and exploring its functionalities, it is essential to set up a local development environment. This section provides a step-by-step guide to help interested users quickly set up the necessary dependencies and configuration. By following these instructions, you can create an environment where you can run the project, experiment with different components, and make modifications as needed.

The project is hosted on GitHub at https://github.com/cuberetry/nokkaewGPT. You can find the latest version of the codebase, documentation, and additional resources on the repository.

In this section, we will cover the installation of required software, the setup of the project repository, and any additional steps necessary to ensure a smooth development experience. Whether you are a developer looking to contribute to the project or an end user interested in running the application locally, this guide will help you get up and running efficiently.

Cloning the Repository

To get started, you'll need to clone the project repository from GitHub. Follow these steps to clone the repository to your local machine:

1. Open your preferred command-line interface or Git client.
2. Navigate to the directory where you want to store the project.
3. Execute one of the following commands to clone the repository:

Using HTTPS:

```
$ git clone https://github.com/cuberetry/nokkaewGPT.git
```

Using SSH:

```
$ git clone git@github.com:cuberetry/nokkaewGPT.git
```

This will create a local copy of the repository on your machine. Once the cloning process is complete, you will have the project files and code available locally, allowing you to proceed with the setup.

Check for Prerequisites

After cloning the repository, make sure you have the following prerequisites installed on your system:

Python

NokkaewGPT requires Python to be installed on your machine. It is recommended to use Python version 3.10.8 or later. If you don't have Python installed or need to update to the recommended version, you can download it from the official Python website: https://www.python.org/downloads/

Pip

Pip is the package installer for Python, and it is usually installed along with Python. To check if Pip is installed, open a command-line interface and run the following command:

```
$ pip --version
```

If Pip is installed, it will display the version number. If not, you can install Pip by following the instructions provided on the official Pip website: https://pip.pypa.io/en/stable/installing/

Creating a Conda Environment

To ensure a clean and isolated environment for running NokkaewGPT, it is recommended to set up a Conda environment. Conda allows you to manage packages and dependencies separately from your system's Python installation. Follow the steps below to create a Conda environment for NokkaewGPT:

1. Open a command-line interface and navigate to the root directory of the cloned repository.

2. Create a new Conda environment using the provided environment.yml file. Run the following command:

```
$ conda env create -f environment.yml
```

This command will create a new environment named nokkaewGPT with all the required dependencies specified in the environment.yml file. The environment creation process may take a few minutes to complete.

3. Activate the newly created Conda environment. Run the following command:

```
$ conda activate nokkaewGPT
```

This will activate the nokkaewGPT environment and allow you to run the NokkaewGPT model within this isolated environment.

With the Conda environment set up and activated, you are ready to proceed with the next steps to preprocess the corpus, train the model, and generate output. Ensure that you activate the nokkaewGPT environment whenever you work on this project to ensure compatibility with the installed dependencies.

Generating an Output

To generate text using the NokkaewGPT, a series of steps need to be followed. This setup process involves three key stages: preprocessing, training, and output generation. First, the corpus needs to be preprocessed using the provided script. This step prepares the data for training by applying necessary transformations and encoding techniques. The preprocessing command can be executed in the terminal or command prompt using the following command:

```
$ python preprocess.py
```

Once the corpus is preprocessed, the training process can be initiated. The training script allows the model to learn from the data and optimize its performance. Before

training begins, the script prompts the user to decide whether to reset the model or continue training with the existing one. To start training, execute the command:

```
$ python train.py
```

Before the training begins, the training script will prompt you with a question regarding whether you wish to reset the model. If you want to start the training from scratch and reset the model, simply enter "yes" at the prompt. On the other hand, if you want to continue training with the existing model and retain the learned parameters, you can enter anything else as your response.

Once you have handled the model reset prompt, the script will proceed to ask you for the number of training steps you wish to execute. This parameter determines the duration of the training process and controls the number of iterations the model will go through to optimize its performance. To specify the desired number of training steps, simply enter the corresponding value and press Enter to proceed with the training.

It is essential to configure the number of training steps appropriately based on your specific requirements and available computational resources. This adjustment allows you to strike a balance between the training time and the model's learning capacity, enabling you to achieve the desired results effectively.

Once the model is trained, you can generate output using the trained language model. By running the command:

```
$ python generate.py
```

and providing an input prompt, the script will generate text based on the learned patterns and save the output to a file named `output_from_model.txt` in the `./output directory`.

Training Result Log

During the training of the language model, various logs are generated to track important information and monitor the performance of the model. These logs are stored in a

dedicated directory, and each log file is named with a timestamp indicating when the script was executed.

The log files capture essential details such as the training settings, hyperparameters, and training progress. They serve as a valuable resource for analyzing the training process and evaluating the model's performance. By reviewing the log files, you can gain insights into the training loss, validation loss, and other metrics that indicate the model's convergence and quality.

In this section, we present an example output from a log file to demonstrate the information recorded during the training process. We will discuss the format of the log file and interpret the key components such as the timestamp, model reset status, hyperparameters, training steps, and training/validation loss values. This information will help you gain a better understanding of the training progress and assist in assessing the performance of the trained language model.

The log files generated by the `train.py` script are stored in the directory `./log/yyyy-mm-dd-hh:mm:ss.txt`, where `yyyy-mm-dd-hh:mm:ss` represents the timestamp of when the script was executed. These log files provide valuable information about the training process and allow for detailed analysis of the model's performance.

Here is an example of the generated log file:

```
Ran at 2023-05-17-13:31:19

Reset: False

Hyperparameters:
batch_size: 16
block_size: 32
n_embd: 64
n_head: 4
```

```
n_layer: 4
dropout: 0.1
vocab_size: 2146

Training steps: 100

The step below is saved!0/100
0/100
Step 0: Training Loss: 3.3713910579681396, Validation Loss:
3.6702888011932373

100/100
Step 100: Training Loss: 3.3713910579681396, Validation Loss:
3.6702888011932373

Test Loss: 1.0674601793289185
```

In this example, the log file begins with the timestamp indicating when the script was executed. It is followed by information about whether the model was reset or not. Next, the hyperparameters used for the training are listed, providing insight into the configuration of the model.

The log also includes the number of training steps that were executed. This is followed by the training progress, displaying the current step out of the total number of steps. For each step, the training loss and validation loss values are recorded, reflecting the model's performance at that particular point in the training process.

Every time the training is conducted, the model would validate the final checkpoint by evaluating the loss value using the test dataset. The result would be recorded at the end of the created log file.

Analyzing the log files allows you to monitor the training progress, evaluate the loss values, and make informed decisions about the model's training and performance. It provides valuable insights for further optimization and fine-tuning of the language model.

Interpreting the log files is an essential part of analyzing and understanding the training process of the language model. The training loss and validation loss values recorded in the log files provide valuable insights into the model's performance and its ability to learn from the data.

The training loss represents the average loss computed during each training step. A lower training loss indicates that the model is successfully learning from the training data and making progress towards minimizing the error. On the other hand, a higher training loss suggests that the model might be struggling to capture the patterns in the data or encountering difficulties in convergence.

The validation loss, on the other hand, measures the model's performance on a separate validation dataset that is not used for training. It serves as an indicator of how well the model generalizes to unseen data. Monitoring the validation loss helps in identifying overfitting or underfitting. Overfitting occurs when the model performs well on the training data but fails to generalize to new data. Underfitting, on the other hand, happens when the model is too simplistic to capture the underlying patterns in the data.

By examining the training and validation loss values over different training steps, you can gain insights into the model's convergence, stability, and generalization capabilities. Analyzing the trends and patterns in the loss values helps in determining the optimal training duration and identifying potential issues or areas for improvement.

It is important to note that the interpretation of the log files should be done in conjunction with domain knowledge and specific project requirements. Different tasks and datasets may have unique characteristics and performance metrics that need to be considered. Therefore, it is recommended to perform thorough analysis and experimentation to fine-tune the model and optimize its performance based on your specific needs.

PROJECT DISCUSSION

This chapter is the section where we discuss our issues, solutions, and discoveries during our project processing.

Runtime Error

This section is a discussion about the runtime error that we encounter during the training process in APEX. The error shown below as an imager Figure 7.1 indicates that “Runtime Error: Expected all tensors to be on the same device”.

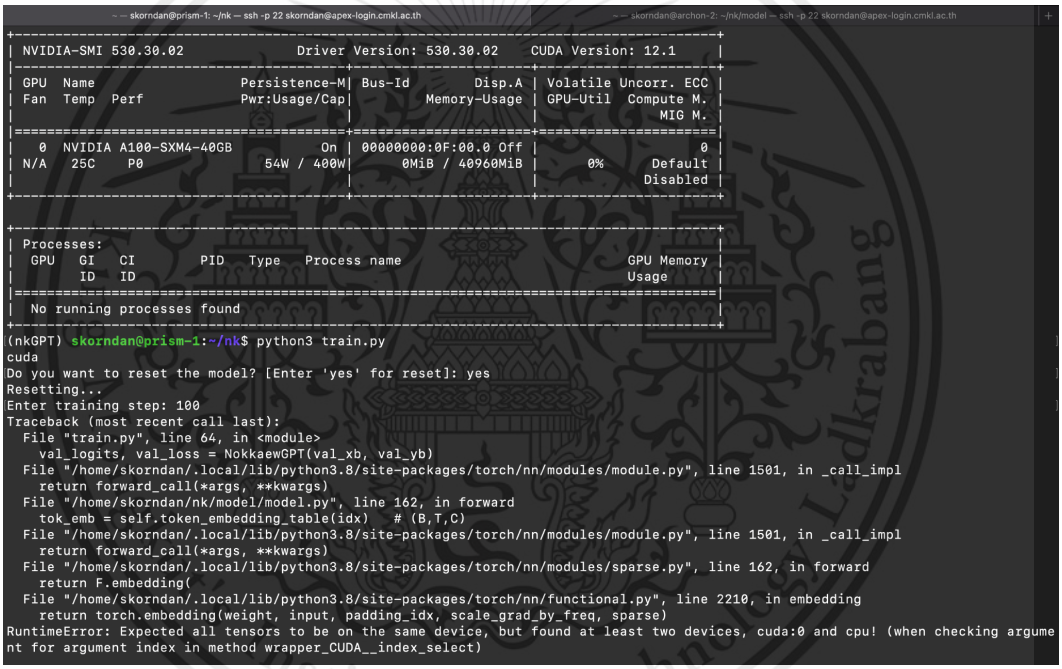

Figure 5.1 Runtime Error: Expected all tensors to be on the same device

During the training process, an error occurred when the model attempted to train while the tensors were not on the same device. Specifically, the model was assigned to the GPU (cuda) as indicated in Figure 7.1 by the line of code "m = NokkaewGPT.to(model.device)". However, when evaluating the loss, the tensors val_xb and val_yb were not on the same device as the model. To resolve this issue, it was necessary to ensure that these tensors were moved to the same device, which in this

case is the GPU. By adding ".to(model.device)" to the code, the tensors were properly assigned to the GPU device, enabling successful evaluation of the loss.

Out of Memory Error

This section is a discussion about the out of memory error that we encounter during the training process. The error shown below as an imager Figure 7.2 indicates that “torch.cuda.OutOfMemoryError: CUDA out of memory.”.

```
torch.cuda.OutOfMemoryError: CUDA out of memory. Tried to
allocate 128.00 MiB (GPU 0; 4.75 GiB total capacity; 3.71 GiB
already allocated; 114.00 MiB free; 3.84 GiB reserved in total
by PyTorch) If reserved memory is >> allocated memory try
setting max_split_size_mb to avoid fragmentation. See
documentation for Memory Management and
PYTORCH_CUDA_ALLOC_CONF
```

Figure 5.2 Out of Memory Error

During the training process, if the error "torch.cuda.outofmemoryerror: CUDA out of memory" occurs, it indicates that the GPU's available memory is insufficient for the operation. This error message provides details about the total GPU capacity, already allocated memory, and available free memory. This error involves various factors such as batch size, model complexity, memory optimization techniques, and hardware resources.

To resolve this issue, several solutions can be employed. Firstly, reducing the batch size decreases the memory requirements per training iteration. Another approach involves simplifying the model by reducing the number of layers, hidden units, or attention heads, thereby reducing memory demand. Alternatively, utilizing a GPU with a larger memory capacity or employing distributed training across multiple GPUs can provide additional memory resources. Additionally, PyTorch's memory management features, such as memory caching and releasing, can be utilized for efficient memory allocation.

Weird Number Text Generation

This section is a discussion about the weird number text generation that we encounter when generating the output. The error shown below as an imager Figure 7.3 indicates the number mixing with thai word.

```
type your content:
จากกรณีมีภาพเผยแพร่ทางโซเชียลมีเดีย พบวัตถุปริศนาเปล่งแสงคล้ายดาวมองเห็นชัดในตอนกลางวัน ในพื้นที่ จ.มุกดาหาร ช่วงเวลาประมาณ 16.00 น. ของวันที่ 22 ธ.ค. ที่ผ่านมา วันนี้ (23 ธ.ค.2564) นายศุภฤกษ์ คฤหานนท์ หัวหน้างานบริการวิชาการดาราศาสตร์ สถาบันวิจัยดารา ศาสตร์แห่งชาติ กล่าวว่า จากคลิปวิดีโอดังกล่าว เป็นวัตถุที่เคลื่อนที่อย่างช้า ๆ บนท้องฟ้า มีลักษณะเป็นสีขาวสะท้อนแสงดวงอาทิตย์เบื้องต้น สันนิษฐานว่าเป็นวัตถุลอย เช่น โคมลอย หรือลูกโป่ง การที่เราเห็นวัตถุมีความสว่างตามทิศทางของแสง แสดงให้เห็นว่ามีการสะท้อนแสงจากดวง อาทิตย์ การสะท้อนแสงของสีขาวเป็นเรื่องปกติที่สามารถสังเกตได้จากภาพถ่ายต่าง ๆ ซึ่งมีความแตกต่างอย่างชัดเจนจากวัตถุประเภท "ดาวตก" จากคลิปวิดีโอ เราค่อนข้างมั่นใจว่าไม่ใช่ดาวตก เนื่องจากวัตถุที่เป็นประเภท "ดาวตก" จะมีลักษณะการเคลื่อนที่เข้ามาในชั้นบรรยากาศของโลก เป็นแนวเส้นตรงด้วยความเร็วค่อนข้างสูง และจะเสียดสีในชั้นบรรยากาศลุกไหม้เป็นแสงวาบให้เห็นได้อย่างชัดเจน แตกต่างจากภาพที่เห็นในคลิป วิดีโอดังกล่าว ที่ไม่มีการเสียดสีลุกไหม้กันในชั้นบรรยากาศ และไม่มีแสงวาบแต่อย่างใด สำหรับข้อสันนิษฐานว่าเป็น "ดาวเทียม" ไม่น่าจะเป็นไปได้ เนื่องจากดาวเทียมเคลื่อนที่ด้วยอัตราเร็วพอสมควร และเราจะสังเกตเห็นดาวเทียมจากการสะท้อนแสงของแผงโซลาร์เซลล์ที่สะท้อนแสงดวง อาทิตย์กลับมายังโลก เป็นจุดแสงที่มีความสว่างในระดับหนึ่งเท่านั้น นอกจากนี้ ดาวเทียมประเภทวงโคจรต่ำ จะโคจรอยู่ที่ความสูงจากพื้นโลก ระหว่าง 350-2,000 กิโลเมตร อยู่ที่ระดับความสูงมากกว่าการเสียดของพวกดาวตก วัตถุประเภทดาวตก จะเสียดสีในชั้นบรรยากาศที่ระดับความ สูงประมาณ 80-120 กิโลเมตร ดังนั้น การที่จะสามารถมองเห็นวัตถุประเภทดาวเทียมได้ในเวลากลางวันด้วยตาเปล่านั้น เป็นไปได้ยากมาก

Content tokenize:
จา ก ก ร ณี มี ภา พ เผ ย แพ ร่ ทา ง โซ เชีย ล มี เดีย พ บ วั ต ถุ ป ริ ศ นา เป ล่ ง แส ง ค ล้ า ย ดา ว ม อ ง เห็น ชั ด ใน ต อ น ก ลา ง วั น ใน พื้น ที่ จ . มุ ก ดา หา ร ช่ ว ง เว ลา ป ระ มา ณ 16 . 00 น . ข อ ง วั น ที่ 22 ธ . ค . ที่ ผ่ า น มา วั น นี้ ( 23 ธ . ค . 2564 ) นา ย ศุ ภ ฤกษ์ ค ฤ หา นนท์ หั ว หน้ า งา น บ ริ กา ร วิ ชา กา ร ดา รา ศา สตร์ ส ถา บั น วิ จั ย ดา รา ศา สตร์ แห่ ง ชา ติ ก ล่ า ว ว่ า จา ก ค ลิ ป วิ ดี โอ ดั ง ก ล่ า ว เป็ น วั ต ถุ ที่ เค ลื่ อ น ที่ อ ย่ า ง ช้ า ๆ บ น ท้ อ ง ฟ้า มี ลั ก ษ ณะ เป็ น สี ขา ว สะ ท้ อ น แส ง ด ว ง อา ทิตย์ เบื ้ อ ง ต้ น สั น นิ ษ ฐา น ว่า เป็น วั ต ถุ ล อ ย เช่ น โค ม ล อ ย ห รือ ลู ก โป่ ง กา ร ที่ เรา เห็น วั ต ถุ มี ค วา ม ส ว่ า ง ตา ม ทิ ศ ทา ง ข อ ง แส ง แส ด ง ให้ เห็น ว่า มี กา ร สะ ท้ อ น แส ง จา ก ด ว ง อา ทิตย์ กา ร สะ ท้ อ น แส ง ข อ ง สี ขา ว เป็น เรื ่ อ ง ป ก ติ ที่ สา มา ร ถ สั ง เก ต ได้ จา ก ภา พ ถ่ า ย ต่ า ง ๆ ซึ่ ง มี ค วา ม แต ก ต่ า ง อ ย่ า ง ชั ด เจ น จา ก วั ต ถุ ป ระ เภ ท " ดา ว ต ก " จา ก ค ลิ ป วิ ดี โอ เรา ค่ อ น ข้ า ง มั่น ใจ ว่า ไม่ ใช่ ดา ว ต ก เนื ่ อ ง จา ก วั ต ถุ ที่ เป็ น ป ระ เภ ท " ดา ว ต ก " จะ มี ลั ก ษ ณะ กา ร เค ลื่ อ น ที่ เข้ า มา ใน ชั้ น บ ร ร ยา กา ศ ข อ ง โล ก เป็ น แน ว เส้ น ต ร ง ด้ ว ย ค วา ม เร็ว ค่ อ น ข้ า ง สู ง และ จะ เสี ย ด สี ใน ชั้ น บ ร ร ยา กา ศ ลุ ก ไห ม้ เป็ น แส ง วา บ ให้ เห็น ได้ อ ย่ า ง ชั ด เจ น แต ก ต่ า ง จา ก ภา พ ที่ เห็น ใน ค ลิ ป วิ ดี โอ ดั ง ก ล่ า ว ที่ ไม่ มี กา ร เสี ย ด สี ลุ ก ไห ม้ กั น ใน ชั้ น บ ร ร ยา กา ศ และ ไม่ มี แส ง วา บ แต่ อ ย่ า ง ใด สำ ห รั บ ข้ อ สั น นิ ษ ฐา น ว่า เป็น " ดา ว เทีย ม " ไม่ น่า จะ เป็น ไป ได้ เนื ่ อ ง จา ก ดา ว เทีย ม เค ลื่ อ น ที่ ด้ ว ย อั ต รา เร็ว พ อ ส ม ค ว ร และ เรา จะ สั ง เก ต เห็น ดา ว เทีย ม จา ก กา ร สะ ท้ อ น แส ง ข อ ง แผ ง โซ ลา ร์ เซลล์ ที่ สะ ท้ อ น แส ง ด ว ง อา ทิตย์ ก ลั บ มา ยั ง โล ก เป็น จุ ด แส ง ที่ มี ค วา ม ส ว่ า ง ใน ระ ดั บ ห นึ่ง เท่ า นั้ น น อ ก จา ก นี้ ดา ว เทีย ม ป ระ เภ ท ว ง โค จ ร ต่ า จะ โค จ ร อ ยู่ ที่ ค วา ม สู ง จา ก พื้น โล ก ระ ห ว่ า ง 350 - 2 , 000 กิ โล เม ต ร อ ยู่ ที่ ระ ดั บ ค วา ม สู ง มา ก ก ว่ า กา ร เสี ย ด ข อ ง พ ว ก ดา ว ต ก วั ต ถุ ป ระ เภ ท ดา ว ต ก จะ เสี ย ด สี ใน ชั้ น บ ร ร ยา กา ศ ที่ ระ ดั บ ค วา ม สู ง ป ระ ม <h>

Generate:
ลคุ์ ลคุ์ 0847 ลคุ์ 0847 ลคุ์ 1799 72151 มี 72815 8787 ลคุ์ 0847 0847 ลคุ์ ลคุ์ 1629000 ไกลน์ 0847 ลคุ์ 0847 1214571 ลคุ์ ลคุ์ ลคุ์ ๔๔๒ ลคุ์ 0847 ลคุ์ ลคุ์ 0847 ญษ์ 80301 มุห์ ลคุ์ ลคุ์ 1300600 ลคุ์ 0847 3893 ลาว์ 是 ๒๕๕๗๒๑ ลคุ์ 0847 1300600 ลคุ์ 145060 167559 53101 80301 80301 5035755 1799 58217 ๒๕๔๐ แดะ 112044 แฮ็ช 8746 ลคุ์ แดะ ทีส ลคุ์ 1799 มี ลคุ์ ลคุ์ 58217 ลคุ์ 0847 58217 397581 อุดร์ ริกซ์ ลคุ์ ลคุ์ สรน์ ลาว์ ลคุ์ 1300600 ลคุ์ 0847 80301 ลคุ์ 1214571 ไกลน์ 1300600 ลคุ์ ลคุ์ มะฮ์ ลคุ์ 0847 5035755 8746 80301 6108 5035755 1799 ลคุ์ 145060 ลคุ์ 145060 ค้า 646928 72815 22755000 4333 0949520668 5371 ลคุ์ 0847 ลคุ์ ลคุ์ 0847 ลคุ์ 0847 ลคุ์ 145185000 ลคุ์ ลคุ์ ลคุ์ 0847 58217 ลคุ์ 0847 1300600 ลคุ์ ลคุ์ ลคุ์ 1799 1300600 เร็ธ Þ 1799 ลคุ์ 137 มนย์ ษ์ แฮ็ช ๒๕๕๗๒๑ 閧 ลคุ์ 1799 0847 5809 มุห์ ลคุ์ 255319 มีป ๒๕๕๗๒๑ 80301 ลคุ์ 58217 ลคุ์ 255319 5035755 ลคุ์ ลคุ์ 1300600 ลคุ์ 1799 1300600 ลคุ์ 1300600 ลคุ์ 0847 1214571 ลคุ์ ลคุ์ แดะ 22755000 มุห์ ลคุ์ ลคุ์ 0847 361090 มนย์ 150050 มุห์ 1300600 ลคุ์ ลคุ์ 0847 ลคุ์ 44750 ลคุ์ ลคุ์ 1799 ลคุ์ 0847 ลคุ์ 0847 ลคุ์ 58217 0847 ลคุ์ ลคุ์ 0847 1300600 มุห์ ลคุ์ 0847 1300600 0847 委 ลคุ์ 0847 0847 มีป ลคุ์ 1799 0910503 ลคุ์ 0847 1799 1799 0847 0847 มีป ลคุ์ 0847 0847 1300600 สรน์ 14140 0847 0847 ลคุ์ 0847 ลคุ์ 0847 ลคุ์ 0847 ลคุ์ 0847 ลคุ์ 145060 0847 1300600 1300600 0847 2807 ลคุ์ 0847 ค้า ลคุ์ 1300600 8854072 0847 0847 2807 37350000 9073 ลคุ์ 0847 ลคุ์ 0847 ลคุ์ 0847 2807 1300600
```

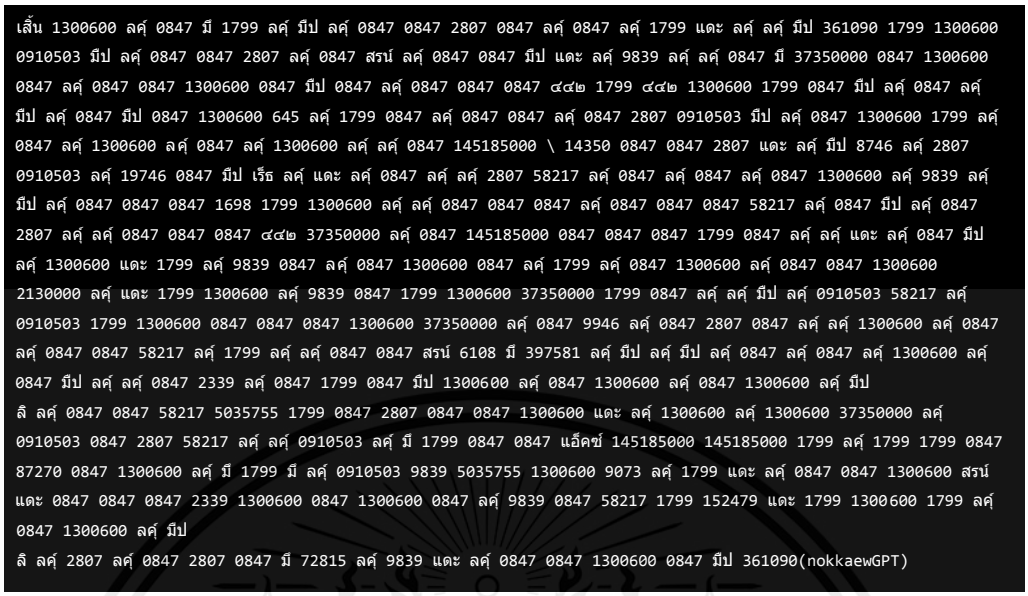

Figure 5.3 Weird Number Text Generation

From figure 7.3 above, our text generation model encountered an issue where numbers and words were being mixed together in the generated text. This problem was identified to be associated with the tokenization process, particularly the handling of numbers. The model was treating numerical values from the input article, like prices, years, and phone numbers, as vocabulary items and incorporating them into the training process. As a result, the output text contained a combination of Thai characters and numbers. To resolve this, we refined the tokenization rules and configurations to treat numbers separately and ensure they are not included in the generated text.

PROJECT CONCLUSION

In conclusion, the NokkaewGPT model, built on the GPT-2 architecture tailored for the Thai language, has demonstrated impressive capabilities in natural language generation tasks. Leveraging the power of the GPT-2 architecture, which has been widely successful in various language models, NokkaewGPT extends its effectiveness to the Thai language domain.

The preprocessing stage plays a crucial role in shaping the performance of NokkaewGPT, as it employs various tokenization techniques such as character, subword, and word level tokenization. These tokenization approaches have a direct impact on the model's understanding of the Thai language and its ability to generate coherent and appropriate text.

Character-level tokenization treats individual characters as distinct tokens, allowing the model to capture fine-grained details and relationships between characters. Subword-level tokenization, on the other hand, breaks the text into smaller meaningful units, enabling the model to handle more complex words and expressions. Lastly, word-level tokenization treats complete words as tokens, emphasizing semantic understanding at a higher level.

By experimenting with these tokenization levels, NokkaewGPT explores the trade-offs between granularity and generalization in the generated outputs. Each approach has its strengths and limitations, impacting factors such as fluency, vocabulary coverage, and contextual coherence. Thus, the choice of tokenization level significantly influences the model's overall performance and the quality of generated text in the Thai language.

As such, understanding the tokenization strategy in accordance with the specific task requirements and linguistic characteristics of Thai is crucial for achieving the desired results with NokkaewGPT. Further research and experimentation in tokenization techniques can lead to enhancements in the model's capabilities and its ability to generate text that aligns more closely with human-like linguistic patterns in the Thai language context.

The training process further optimizes the model parameters through an iterative optimization algorithm. As the model progresses through the training iterations, it learns to better capture the statistical patterns and semantic structures present in the Thai language data. This optimization leads to a gradual decrease in the loss values observed during the training, validation, and testing phases.

The lower loss values obtained signify that the model becomes more adept at understanding the intricacies of Thai language and generating high-quality text outputs. The model's ability to minimize the loss function reflects its proficiency in capturing the underlying patterns and dependencies in the training data, which in turn enables it to generate meaningful and contextually relevant text.

These lower loss values, combined with the evaluation metrics and qualitative analysis of the generated text, demonstrate the model's effectiveness and its potential as a valuable asset for various applications that require natural language generation in the Thai language context. The ability to produce coherent, fluent, and contextually appropriate text outputs in Thai opens up possibilities for applications such as chatbots, content generation, language assistance, and more. The NokkaewGPT model contributes to advancing the field of Thai language processing and provides a powerful tool for generating high-quality text in the Thai language, enhancing communication and information processing in the Thai-speaking community.

However, it is important to acknowledge that further evaluation and fine-tuning are required to enhance the model's performance and address potential limitations specific to the Thai language. Continued research and development in this area will contribute to the advancement of Thai natural language processing techniques and their practical applications.

In summary, the NokkaewGPT model stands as a testament to the remarkable progress in Thai language generation enabled by advanced deep learning techniques. Through meticulous training, evaluation, and experimentation, the model has demonstrated its proficiency in generating coherent and contextually relevant text in Thai. By leveraging the power of language modeling and fine-tuning strategies, NokkaewGPT opens up new possibilities for natural language generation in the Thai language context. With further

research and refinement, this model has the potential to revolutionize Thai language processing and contribute to a wide range of applications, from chatbots and content generation to language assistance and creative writing support. The journey towards more fluent and culturally appropriate Thai language generation continues, and the advancements made by NokkaewGPT serve as a foundation for future innovations in Thai natural language processing.

The NokkaewGPT model and its associated codebase are openly available and hosted on GitHub at https://github.com/cuberetry/nokkaewGPT. Researchers, developers, and enthusiasts interested in exploring the model, replicating the experiments, or building upon this work are encouraged to visit the repository. The GitHub platform provides an ideal space for further discussions, collaborations, and inquiries related to the NokkaewGPT project. Feel free to engage with the community, submit issues, contribute improvements, or share your experiences and findings to foster ongoing advancements in Thai language generation and natural language processing.

APPENDICES

APPENDIX

APPENDIX A

APPENDIX

APPENDIX B

REFERENCE

- Karpathy / micrograd. (No date). Available at: https://github.com/karpathy/micrograd.
- Kumar, V. (2019). PyTorch Autograd. Medium. Available at: https://towardsdatascience.com/pytorch-autograd-understanding-the-heart-of-pytorchs-magic-2686cd94ec95.
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3, 1137-1155. Available at: https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf.
- CMKL Apex. (No date). Apex Documentation. Available at: https://cmkl-apex.github.io/apex-documentation/docs/getting-started.
- ScrapeOps. (No date). Python Scrapy Playbook - Scrapy Beginner's Guide. Available at: https://scrapeops.io/python-scrapy-playbook/scrapy-beginners-guide.
- Codex. (No date). Scraping and Analyzing News Articles with Scrapy and Selenium. Medium. Available at: https://medium.com/codex/scraping-and-analyzing-news-articles-with-scrapy-and-selenium-cbbd94381d78.
- Skeptric. (No date). WAT WET WARC - Common Crawl Archives. Available at: https://skeptric.com/notebooks/WAT%20WET%20WARC%20-%20Common%20Crawl%20Archives.html.
- csebuetnlp. (No date). xl-sum. GitHub. Available at: https://github.com/csebuetnlp/xl-sum?fbclid=IwAR2IHCmYbHKJg91-eNInRPW75guUfLKFafo6A_2Lt6QWCi9tMnaXLNgrFkY.
- nakhunchumpolsathien. (No date). ThaiSumDataset3GB. GitHub. Available at: https://github.com/nakhunchumpolsathien/ThaiSum.
- Common Crawl. (No date). Getting Started. Available at: https://commoncrawl.org/the-data/get-started/.
- Jalammar, J. (No date). The Illustrated GPT-2. Available at: https://jalammar.github.io/illustrated-gpt2/.