Lost in Translation: Translating Generation Z Internet Slang using Machine Learning

A Special Problem

Presented to

the Faculty of the Division of Physical Sciences and Mathematics

College of Arts and Sciences

University of the Philippines Visayas

Miag-ao, Iloilo

In Partial Fulfillment

of the Requirements for the Degree of

Bachelor of Science in Computer Science by

FLAUTA, Neil Bryan

GIMENO, Ashley Joy

GIMENO, Carl Jorenz

Francis DIMZON, Ph.D.

Adviser

June 2025

Approval Sheet

The Division of Physical Sciences and Mathematics, College of Arts and Sciences, University of the Philippines Visayas

certifies that this is the approved version of the following special problem:

Lost in Translation: Translating Generation Z Internet Slang using Machine Learning

Approved by:

Name	**Signature**	**Date**
Francis D. Dimzon, Ph.D. (Adviser)		
Ara Abigail E. Ambita (Panel Member)		
Kent Christian A. Castor (Division Chair)		

Division of Physical Sciences and Mathematics
College of Arts and Sciences
University of the Philippines Visayas

Declaration

We, Neil Bryan Flauta, Ashley Joy Gimeno, and Carl Jorenz Gimeno, hereby certify that this Special Problem has been written by us and is the record of work carried out by us. Any significant borrowings have been properly acknowledged and referred.

Name	**Signature**	**Date**
Flauta, Neil Bryan (Student)	____________	____________
Gimeno, Ashley Joy (Student)	____________	____________
Gimeno, Carl Jorenz (Student)	____________	____________

Dedication

This study is dedicated to our loved ones, especially our loving parents, for their unwavering support throughout our academic journey and our continual source of inspiration and strength when we were on the verge of giving up.

To our dear friends, we are grateful for your warm presence, valuable insights, and constant encouragement, which helped us complete this study.

Finally, to our future selves, may this hard work serve as a testament to the obstacles you have overcome. Let this milestone remind you to keep learning and face the future with courage, even if the path is uncertain.

Acknowledgment

We extend our heartfelt gratitude to Dr. Francis D. Dimzon for his patient guidance throughout this study. His thoughtful mentorship in the field of machine learning contributed to the foundation and direction of this study.

Abstract

Internet slang is an informal variation of language that is prominent to the younger generation. Its widespread use has contributed to the generational divide between younger and older generations. This study aimed to develop a translation tool leveraging Large Language Models (LLMs) to bridge this divide. A dataset of Generation Z slang sentences and their formal equivalents was used to fine-tune Zephyr-7B-Beta model. The performance of the fine-tuned model was evaluated against the base model using automatic metrics (BLEU and ROUGE-L) and manual evaluations through online surveys involving Gen Z students. The BLEU and ROUGE-L scores of 0.8151 and 0.8396 respectively, indicates a high degree of similarity between the generated text and the reference, suggesting that the model produces translations that closely match the formal equivalents of the Gen Z slang sentences. Furthermore, manual evaluation results showed that 53.5% of the respondents preferred the translations produced by the fine-tuned model, supporting the results of the automatic metrics. The results suggest that fine-tuning LLMs can significantly improve their ability to translate internet slang into formal English.

Keywords: Internet Slang, Generation Z, Generational Divide, LoRA, LLM

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Language is how humans communicate and express themselves (Crystal & Robins, 2024). It evolves, adapting to the changing needs of users (Jeresano & Carretero, 2022). New words are borrowed or invented (Mantiri, 2010), and most linguistic changes are initiated by young adults and adolescents (Thump, 2016 as cited in (Jeresano & Carretero, 2022)). The younger generation demographic tends to focus on belonging to self-organized groups of peers and friends, forming what can be described as the "we" generation. Through their interactions, language changes differently, making them remarkably distinct from previous generations.

Slang is a great example of the dynamic nature of language. Slang is an informal language used by people in the same social group (Fernández-Toro, 2016). It serves multiple social purposes: identifying group members, communicating informally, and opposing established authority (McArthur, 2003). Slang is highly contextual and pervasive, even in non-standard English. Its figurative nature and how it twists the definitions of the words used make it difficult for outsiders to understand.

In recent years, the Internet has become a significant medium for the evolution and spread of language, giving rise to 'Internet slang' (J. Liu, Zhang, & Li, 2023). Internet slang is a collection of everyday language forms used by various online groups (Barseghyan, 2014). Ujang et al. (2018, as cited in (binti Sabri, bin Hamdan, Nadarajan, & Shing, 2020)) state that internet slang is not easily understood by people outside the social group or people who are not fluent in the language where the slang is used. This phenomenon is particularly prominent among the younger generation (Maulidiya, Wijaya, Mauren, Adha, & Pandin, 2021), where they use it to communicate and interact with friends.

Generation Z, individuals born between 1996 and 2009, are regarded as "digital natives" because technology is an integral part of their upbringing (Dua et al., 2024). Even the language of this generation is greatly affected by technology, where newly coined terms and phrases, called Gen Z slang, are tied to the media culture they've grown up with (Jeresano & Carretero, 2022). However, this evolution of language often creates communication barriers with older generations (Venter, 2017 as cited in (Ghazali & Abdullah, 2021)). Furthermore, studies show that even within Generation Z, people with limited exposure to social media may struggle to understand the prevalent slang (Vacalares, Salas, Babac, Cagalawan, & Calimpong, 2023).

These gaps highlight the need for a tool that can bridge the generational divide, making it easier for individuals to understand the language of Generation Z. Multiple studies have tried translating slang into a formal language using machine learning. Khazeni et al. achieved a 81.91% accuracy in translating Persian slang to formal Persian language using deep learning. Another study by Nocon et al. created a translator to translate Filipino colloquialisms into the Filipino language using Tensorflow's sequence-to-sequence model and Moses' phrase-based statistical machine translation. Furthermore, Ibrahim and Sharief developed a slang translator using models from Hugging Face.

Building on these studies, this study created a translation tool specifically to translate Gen Z slang. The tool will utilize Low Rank Adaptation (LoRA) to a selected Large Language Model (LLM). The results will be evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE).

By fostering mutual understanding, this tool aims to promote more effective and harmonious interactions across age groups, ultimately enhancing relationships and reducing miscommunication.

The main contributions of this study are as follows:

- Enhance linguistic understanding between generations by using fine-tuning a LLM to translate Gen Z slang to formal language, leveraging the strengths of advanced NLP techniques
- Bridge communication gaps between generations using the proposed model to foster better relationships
- Create a scalable framework that can be adapted to translate slang in other languages

1.2 Problem Statement

Internet slang fosters informal, relatable communication within the younger generation (Ghazali & Abdullah, 2021), especially Generation Z, but it presents challenges in understanding for people outside this demographic. The gap in comprehension with older generations widens as internet slang evolves, often leading to miscommunication affecting social relationships that contribute to the generational divide (Vacalares et al., 2023). A more specific translation tool developed using language models can be used to bridge this divide.

By leveraging the ability of LLM to generate a more nuanced and properly constructed answer, a better tool can be made to translate the slang into proper sentences. It has already been proven by the likes of GPT being modified and tailored for use in several automated chatbots to provide customer service. However, no such tool exists for slang translation of Generation Z, which arguably has the most diverse slangs compared to other generations. The creation of this tool will allow translating of such texts into formal sentences and help with bridging the generational divide between them and older people, especially teachers.

1.3 Research Objectives

1.3.1 General Objectives

This study aims to fine-tune the zephyr-7b LLM for use in the translation of Generation Z internet slang used by Filipinos in social media.

1.3.2 Specific Objectives

Specifically, the study aims to:

1. create a dataset of sentences containing Generation Z slang used in differing contexts and its formal translation,
2. create a LoRA implementation for fine-tuning an existing model,
3. fine-tune an existing LLM to translate sentences containing Generation Z slang into formal sentences, and
4. evaluate the performance of the trained model and compare it to the baseline model using several performance metrics.

1.4 Scope and Limitations of the Research

This study focused on the use of internet slang by Filipino Generation Z, with an emphasis on the English language, as it is widely used across various digital platforms, including social media. English has become a dominant medium of communication in the Philippines' digital landscape, particularly among younger demographics. According to a study by (Olobia, 2024), social media platforms serve as powerful tools for communicating in English as a second language, significantly influencing students' language use. The prevalence of English in social media facilitates learning opportunities and cross-cultural communication, highlighting its integral role in the digital communication practices of Filipino youth.

Furthermore, the extensive use of English on social media platforms reflects its status as a marker of education and social standing in the Philippines. As noted by Mateo (2018) cited by (Esquivel, 2020), the widespread use of English in social media underscores its significance in Filipino society, where proficiency in English is often associated with educational attainment and social mobility.

These findings underscore the importance of focusing on English in studies of internet slang among Filipino Generation Z, as it remains a prevalent and influential language in their digital interactions.

1.5 Significance of the Research

This study contributes to the growing body of research on the evolving linguistic landscape shaped by the use of Internet slang, highlighting the communication practices of Generation Z. As digital platforms become increasingly central to daily interactions, Generation Z continues to develop and adopt informal linguistic expressions that reflect their identity, creativity, and cultural environment. While this form of communication enhances peer connectivity, it can also create barriers for individuals outside this demographic, particularly older generations.

The findings of this study offer practical benefits for various stakeholders. For educators, the insights can support the development of more inclusive and responsive classroom communication strategies, enabling them to better understand and engage with their students' language use and cultural context. For parents, the study provides a framework for interpreting the language their children use online and in casual conversations, helping in bridging communication gaps and improving parent-child relationships. For media practitioners and digital marketers, understanding the patterns and meanings behind Gen Z slang can inform the creation of more relatable and culturally relevant content, enhancing audience engagement.

By addressing the communicative divide brought about by generational language differences, this research encourages a more informed approach to language variation in contemporary digital spaces. Ultimately, the study underscores the importance of adapting to linguistic change in order to foster clearer, more effective intergenerational communication.

Chapter 2

Review of Related Literature

2.1 Communication Gap between Generations

Language is dynamic in nature and thus, constantly evolving over time. One example of this behavior is the development of internet slang. Internet slang is a result of language variation and is often regarded as informal (S. Liu, Gui, Zuo, & Dai, 2019). In the study, *The Use of Online Slang for Independent Learning in English Vocabulary* (Ambarsari, Amrullah, & Nawawi, 2020), students used internet slang to express their feelings and emotions, and to align their communication style with their peers.

However, this development has its challenges. It is suggested that younger generation should use slang to communicate with each other instead of older generations because it might cause confusion between them (Jeresano & Carretero, 2022).

This miscommunication is prominent between generations with differences in linguistic familiarity as Suslak (2009) argues that age influences language use, noting that language evolves across generations. Supporting this, a study by Teng and Joo (2023) found that the older a person is, the less likely they are to understand internet language.

Studies have shown that using internet slang improves relationships between those who use it. However, using internet slang for inter-generational communication can be a hindrance to proper and effective communication (Gonzaga, 2025).

2.2 Generative AI

Generative AI encompasses machine learning models that create new content, such as text, images, and audio, based on patterns learned from extensive data (Euchner, 2023). These models, including LLMs like those used in ChatGPT and Bing AI, use neural networks to predict the next word or phrase in a sequence, enabling them to generate human-like text (Brynjolfsson, Li, & Raymond, 2023). The ability of generative AI to understand and produce diverse content, ranging from creative writing code, makes it potentially useful for various applications, such as language translation (Fui-Hoon Nah, Zheng, Cai, Siau, & Chen, 2023).

2.3 Existing Studies

Zephyr-7b-beta has shown performance comparable to that of larger models, most notably, GPT-4 (Tunstall et al., 2023). This is further corroborated by the study by Vergho, Godbout, Rabbany, and Pelrine (2024), which compared multiple open-source LLMs with GPT-3.5 and GPT-4.0 models at that time. They found that zephyr-7b-beta is a viable open-source alternative to these models and is comparable with the latest GPT-4.0 model.

Heydari, Albadvi, and Khazeni (2024) used deep learning to create a model for translating Persian slang text into formal ones. The researchers explored the challenges of translating Persian slang into English within the context of film subtitling, specifically focusing on the performance of three neural machine translation (NMT) systems, namely Google Translate, Targoman, and Farazin. The primary interest of the paper lies in the understanding of how these NMT systems handle the complexities of slang translation. It was revealed that the NMT systems often struggle to capture the nuances of slang, leading to unnatural and inaccurate translations. Targoman performed best in naturalness, but it fell short of human translation quality. This implies the need for specialized algorithms or training data suitable for slang, and potentially human post-editing, to achieve accurate and culturally appropriate translations in this domain.

The study by Nocon, Kho, and Arroyo (2018) explored translating Filipino colloquialisms, such as Conyo and Datkilab, into standardized Filipino, addressing comprehension barriers for non-familiar speakers. Two machine translation (MT) approaches were evaluated: Tensorflow's Sequence-to-Sequence model using Recurrent Neural Networks (RNNs) and Moses' Phrase-based Statistical MT. Moses outperformed Tensorflow on test data due to its handling of phrase combinations and unfamiliar words, while Tensorflow excelled on training data, indicating potential with refinement and more training data. The research underscores the need for robust datasets and highlights the strengths of phrase-based statistical MT in tackling slang translation challenges.

Ibrahim and Sharief (2023) developed a system to translate slang into formal language, addressing challenges posed by slang's informality and variability. Using updated datasets of slang words, formal equivalents, and contextual sentences, they fine-tuned pre-trained models from Hugging Face's Transformer library. While the T5-base model showed promise during training, it performed poorly in testing. In contrast, the "facebook/bart-base" model excelled, demonstrating high accuracy and low loss values. The study highlights the importance of fine-tuning and updating datasets for effective slang translation and emphasizes the potential of transformer models like "facebook/bart-base" in bridging informal and formal language gaps.

While general-purpose instruction tuning is now well-documented, less attention has been paid to fine-tuning LLMs for tasks involving informal or non-standard language such as slang. However, studies are emerging that suggest promising outcomes. For example, the SlangDIT benchmark (Liang, Meng, Wang, & Zhou, 2025) developed a testbed specifically for slang understanding and translation, and preliminary findings indicate that even relatively small models fine-tuned on slang-rich datasets can rival zero-shot GPT-4 performance. This supports the notion that domain adaptation—particularly to informal linguistic domains—benefits substantially from task-specific training, even if the examples are synthetic. A study by Sun, Hu, Gupta, Zemel, and Xu (2024) also showed that even a small dataset of slang sentences helped GPT 3.5 perform better than zero-shot GPT-4.0 at slang detection. While it is a classification task, this suggests a promising approach to improve the performance of LLMs in slang translation tasks.

Table 2.1: Summary of Existing Studies

Author	Focus	Gaps	Problem Solved
Nocon et al. (2018)	Developing machine translators for Filipino colloquialisms using sequence-to-sequence models and statistical machine translation (Moses).	Tensorflow models had issues with unknown tokens and repetitions, and limited ability to generalize to unseen data.	Demonstrated the feasibility of machine translation for Filipino colloquialisms, with Moses as a viable solution.
Ibrahim and Sharief (2023)	Developing an intelligent system to transform English slang words into formal words.	The study noted that more powerful processors could improve the training and testing, and that previous datasets were outdated and needed updating.	Demonstrated an effective model for translating English slang to formal English and highlighted the importance of fine-tuning pre-trained models.
Heydari et al. (2024)	Persian slang text conversion to formal and deep learning of Persian short texts on social media	The BERT models used did not align well with the informal data used in the sentiment analysis.	Created a tool to convert Persian slang to formal text and improved sentiment analysis of short texts using deep learning.

2.4 LoRA for Fine Tuning

Low Rank Adaptation, or LoRA, is an efficient Parameter Efficient Fine Tuning (PEFT) method proposed by Hu et al. (2021). This can significantly decrease the required storage for training while producing comparable results and in some cases even outperforming other adaptation methods. In addition, it has minimal chance of catastrophic forgetting as the original weights are not being tampered with, unlike other fine-tuning methods. These factors make it a suitable option for slang translation as a quick yet accurate solution. In a study conducted by Zhao et al. (2024), they determined that some LLMs using LoRA for fine tuning can outperform GPT-4, one of the most advanced LLM models currently. A study by Nguyen, Wilson, and Dalins (2023) used LoRA in fine tuning a pre-trained Llama 2 7B model for text classification of a dataset that contains slang. They were able to create a more accurate model compared to models by existing studies at that time.

2.5 Data Augmentation through Synthetic Data Generation

Datasets specifically of slang sentences are hard to come by especially ones dedicated to a certain group. This is where synthetic data generation comes into play. Modern LLMs fine-tuning leverages synthetic data generation in many ways. A good example of which is the model we are using, zephyr-7b-beta. This model is fine-tuned from Mistral 7B and was trained on ultrachat dataset (Tunstall et al., 2023), which is a synthetic dataset from data obtained from the Internet (Ding et al., 2023). In addition, the model showed performance comparable to larger open-source models in language tasks.

Synthetic data on its own is not enough to create a model that can perform well in slang translation tasks. A study by Liang et al. (2025) showed that even a small dataset of slang sentences can help improve the performance of LLMs in slang translation tasks. This suggests that domain adaptation, particularly to informal linguistic domains, benefits substantially from task-specific training, even if the examples are synthetic. Nadas, Diosan, and Tomescu (2025) also showed that synthetic data generation can be used to create a synthetic dataset. The measures they used made sure that the dataset is almost as good as a dataset of real slang sentences, especially when augmenting a small dataset. This is particularly useful for slang translation tasks, where datasets are often limited and hard to come by.

2.6 Evaluation Metrics

Automatic evaluation metrics are essential for assessing the performance of machine translation systems, especially in the context of slang translation. These metrics provide a quantitative measure of translation quality, allowing for efficient comparison between different models and approaches. Commonly used metrics include BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation). BLEU measures the overlap between the machine-generated translation and one or more reference translations, focusing on n-gram precision (Papineni, Roukos, Ward, & Zhu, 2001). ROUGE, on the other hand, evaluates the quality of summaries by comparing them to reference summaries, emphasizing recall and precision (Lin, 2004). For slang translation, these metrics can be particularly useful in assessing how well a model captures the nuances and informal expressions characteristic of slang. However, it is important to note that while these metrics provide valuable insights, they may not fully capture the semantic richness and cultural context inherent in slang expressions (Liang et al., 2025). Therefore, human evaluation is often recommended to complement automatic metrics, ensuring a more comprehensive assessment of translation quality. As such, a pairwise comparison of the generated translations against a reference translation is often used to evaluate the performance of LLMs, as it is done with other studies (Zhao et al., 2024)(Chiang et al., 2024). This method allows for a more nuanced understanding of how well a model captures the informal expressions and cultural context inherent in slang, providing a more comprehensive assessment of translation quality.

2.7 Chapter Summary

This chapter shows how generational differences create communication gaps, especially due to internet slang. Younger people tend to use slang to express emotions and connect with friends, but this can confuse older generations who aren't as familiar with these terms. Research shows that as language changes over time, older people are generally less likely to understand the newest internet language. To bridge this gap, some recent studies have utilized machine learning to translate slang into more standard language. For instance, Heydari et al. (2024) used deep learning to translate Persian slang, while Nocon et al. (2018) created a Filipino slang translator using statistical models. Moreover, Ibrahim and Sharief (2023) fine-tuned pre-trained models to learn slang meanings. One promising technique for this is Low Rank Adaptation (LoRA), which is a fine-tuning method that keeps the original model stable while using less storage. Studies by Zhao et al. (2024) and Nguyen et al. (2023) show that LoRA models are not only efficient but can even outperform advanced models like GPT-4 when it comes to slang translation and text classification. However, datasets specifically for slang translation are often limited, making synthetic data generation a valuable tool.

Chapter 3

Research Methodology

This chapter lists and discusses the specific steps and activities that will be performed to accomplish the project. The discussion covers the activities from pre-proposal to Final SP Writing.

3.1 Research Activities

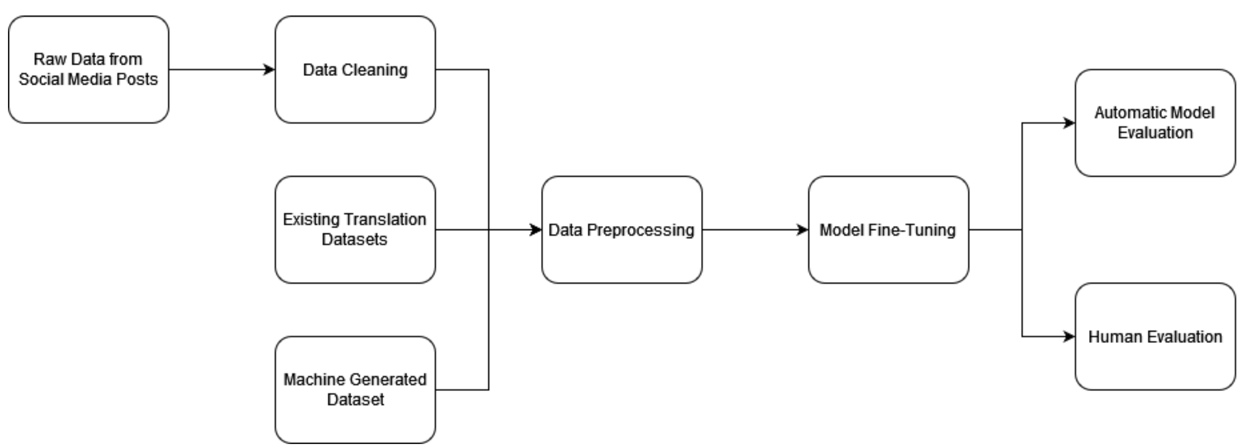

Figure 3.1: Summarized Methodology

3.1.1 Data Gathering

A dataset of sentences containing Generation Z slang and its formal translation was used in this study. This dataset was created using several source: data obtained from social media posts and manually translated by the researchers, existing datasets from HuggingFace, and machine generated and translated sentences using GPT-4o from OpenAI.

The data obtained from social media posts were from verified users of X whose ages are within the Generation Z, so that the dataset is accurate. The data was manually translated by the researchers to ensure that the translation is accurate and reflective of the target demographic. This study also takes into account the regional, cultural, and temporal variations inherent in Gen Z internet slang. The dataset was sourced mostly from individuals whose slang expressions are heavily influenced by contemporary pop culture, including social media trends, music, gaming, and online communities. Temporally, the dataset covers slang used from 2020 through 2024, capturing the evolution of language over recent years and allowing the model to account for shifts in slang popularity and meaning within this period. We, then, used the consensus of the translators and the surrounding environment in which it was created. For social media posts, we considered other comments made by the poster to determine the context in which the word is used in. Its translation is then created using the translator's own experience. To ensure correctness and accuracy of the translations, each term was cross-referenced with reliable online sources, such as online slang dictionaries, forums, and context-specific usage on social media platforms. This process helped confirm that the translated meanings aligned with commonly accepted interpretations and real-world usage, thus enhancing the validity of the dataset.

Data obtained from existing datasets and GPT-4o was checked manually to check if whether the sentence is one used by Generation Z. These processes ensured that the dataset is of high quality and representative of what and how Generation Z slang is used.

3.1.2 Data Preprocessing

The dataset used for the fine-tuning of the model was preprocessed to ensure optimal performance of the model. Unnecessary information such as email addresses and URLs was removed. The data was then manually cleaned up to remove unnecessary characters such as emojis and fixed issues such as typos. A similar approach was done with existing and machine generated datasets to ensure consistency within the training dataset.

The dataset is then split into train and test datasets in a 90/10 ratio to maximize the data learned by the model without compromising on the model's ability to generalize to new data. The train dataset is then split again into a 90/10 ratio to ensure no overfitting while still allowing the model to adapt to the pattern of slang. The cleaned up dataset was then tokenized through the Transformers library provided by HuggingFace as the library already has tokenizers available for their pretrained models. This ensures that the data is formatted properly as required by the model to be used.

3.1.3 Model Fine-Tuning

The model used in this study was zephyr-7b-beta because it is open-source and was proven to perform better than other models of the same size. The LLM is capable of understanding how a slang word is used through the surrounding words. This ensures that as long as the word is used within the same context, it will have the correct interpretation. In addition, it can be trained in a GPU with 16GB of VRAM, necessary as we are using the free plan of Google Colab as the platform of choice for prototype fine-tuning of the model. However, during the training process with the full dataset, the Pro+ plan of Google Colab was used for faster training time and background execution of the training process, allowing the training to continue uninterrupted regardless of the network connection. This study used the example codes provided by HuggingFace in the documentation of their various libraries and sample notebook provided in the zephyr-7b-beta repository.

The SFTTrainer has EarlyStoppingCallback built in that stops the model training when the evaluation criteria set for the callback stop improving more than the specified threshold for a specified number of epochs regardless of if the training loss is still lowering. After it stops the model training, it will load the model with the best score in terms of the evaluation criteria. This ensures that no overfitting occurs as the validation dataset is independent of the training dataset.

The model was loaded using the Transformers library and was quantized into 4 bits through BitsandBytes library to fit the entire model in the allocated resources while having enough headroom for training. In addition, the Unsloth library was used to speed up the training time and reduce the resources used even more

(Daniel Han & Team, 2023). A LoRA adapter was then attached to the model to further reduce the parameters to be trained.

To evaluate the model training process and ensure that the model is not overfitting, BLEU and ROUGE will be used. These metrics use n-grams, making them superior to standard recall and precision metrics as they take into account the positioning of the words. These two metrics were implemented using the Evaluate library by HuggingFace, making it easier to integrate with the rest of the model training process. These metrics was calculated at every epoch of the training process and is used for an early stopping callback to immediately stop the model training if the model seems to be overfitting.

The model was then trained using SFTTrainer class from the Transformer Reinforcement Learning (TRL) library of HuggingFace to simplify the training process (von Werra et al., 2020). The model was trained with the following parameters: optimizer is paged 4bit AdamW, batch size of 8, learning rate of 2e-5, and maximum number of epochs of 50. These parameters were chosen based on the GPU provided in Colab, the test notebook by HuggingFace and the default parameters of SFTTrainer.

3.1.4 Model Evaluation

The model was evaluated using both automatic and manual evaluation metrics. Identical answers and answers with minimal difference, such as punctuation, between the fine-tuned and the base model were removed in the test set to ensure that the model is evaluated properly. After filtering, a total of 143 sentences were used to evaluate the model. The model was then prompted to generate a formal sentence for 170 sentences in the test dataset. The generated sentences were then compared to the formal translation of the sentence using BLEU and ROUGE metrics. The base zephyr-7b-beta model was also prompted to generate sentences for the BLEU and ROUGE metric and the pairwise comparison for human evaluation.

An online survey was conducted using Google Forms to compare the outputs of the fine-tuned model and the base model in order to evaluate the effectiveness of the fine-tuning process. Participants were presented with sentence pairs generated by both models and were asked to choose the more accurate translation of a given Generation Z slang sentence based on accuracy, naturalness, and contextual appropriateness. To minimize potential ordering bias, the sequence in which the outputs from the two models were displayed was randomized for each pair. The researchers implemented a Split Questionnaire Design (SQD) by dividing the full survey into multiple sets to improve response rates and reduce respondent fatigue (Peytchev & Peytcheva, 2017). A total of 143 questions was unevenly distributed into six forms. In addition, the number of responses per form varied which leads to an unbalanced results with some items being evaluated more than others.

To address these challenges, aggregated weighted average was utilized. In weighted average, the results of each form was weighted so that responses are represented proportionately (Ganti, 2024). Specifically, the responses to each item were first summarized using their average scores. These scores were then weighted by the number of respondents per item to account for variations in form size and respondent count. This weighting approach allowed us to combine results from the six forms in a way that gave appropriate emphasis to the sample size behind each item's score, providing a fair and interpretable basis for comparison across all 143 questions.

This method offered a simple yet effective way to integrate responses from an SQD structure without requiring overlap or complex modeling assumptions. It also ensured that items answered by more respondents contributed more substantially to the overall evaluation while avoiding bias from unequal form lengths.

Chapter 4

Results and Discussions

4.1 Dataset

We built a dataset containing a total of 1155 Gen Z internet slang sentences and their corresponding formal translations. The created dataset was then combined with another dataset from Hugging Face that contains 548 Gen Z internet slang and their corresponding formal translation for a total of 1703 sentence pairs. The dataset was then split into training, validation, and test sets with a ratio of 81:9:10. The training set contains 1380 sentence pairs, the validation set contains 153 sentence pairs, and the test set contains 170 sentence pairs. The dataset was then tokenized using the tokenizer of the base model, zephyr-7b-beta, to prepare it for training. The tokenized dataset was then saved in a JSON format to be used for training the model.

4.2 Model Evaluation

4.2.1 Model Training

The model is built to be highly adaptive to the ever-evolving slang terminology. This model was tested on the free tier of Google Colab ensuring that anyone with access to the service can easily replicate our training process with an updated dataset to update the model. The model was trained for 7 epochs before the early stopping callback was triggered because the evaluation metrics has not improved by at least 0.01 for 3 consecutive epochs. This prevented the over-fitting seen in the following figure. Figure 4.1 shows that the training loss is decreasing and the validation loss is increasing and other metrics are not improving. These indicate that the model is over-fitting to the training data and may not generalize well to new data. The model training was stopped in just 7 epochs and the best model among the epochs, the one with the lowest validation loss and highest metrics, was chosen as the final model.

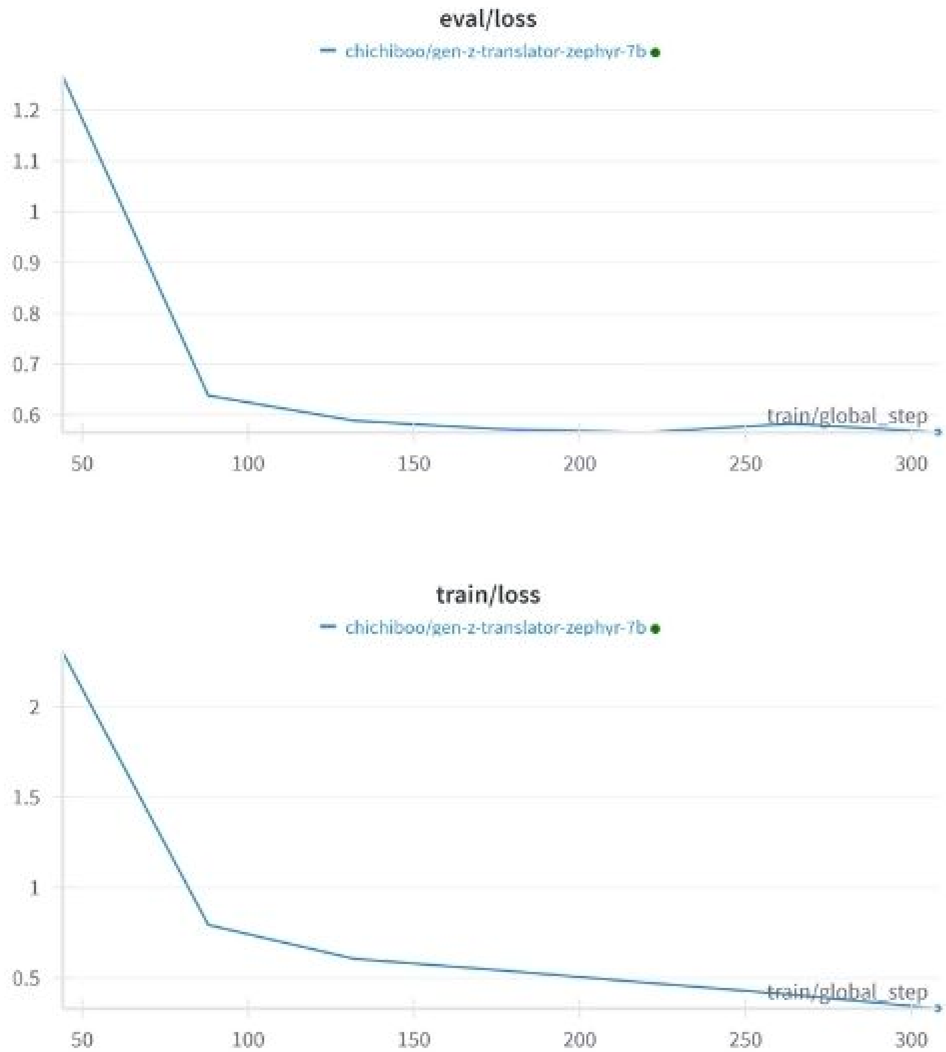

Figure 4.1: Training and evaluation loss curves of the fine-tuned model across training steps

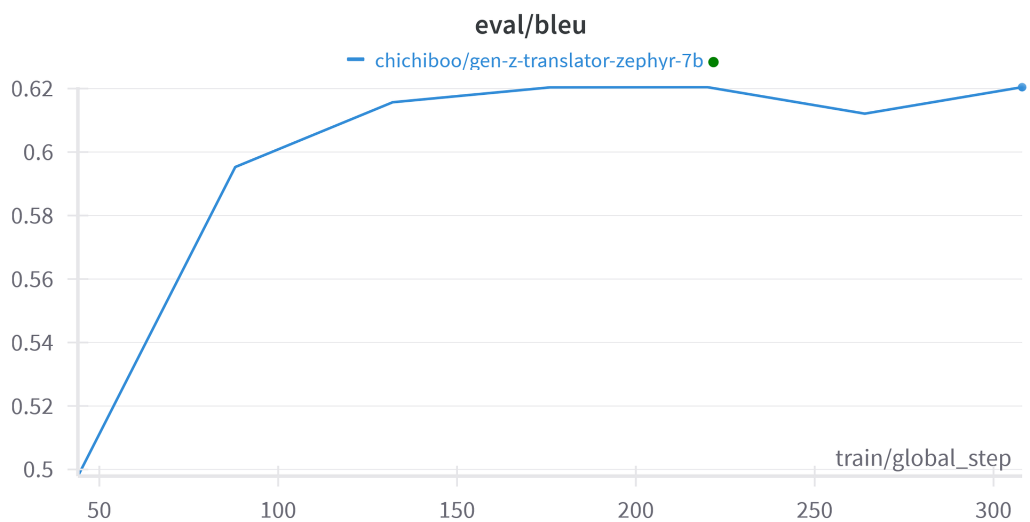

Figure 4.2: Evaluated using BLEU metric

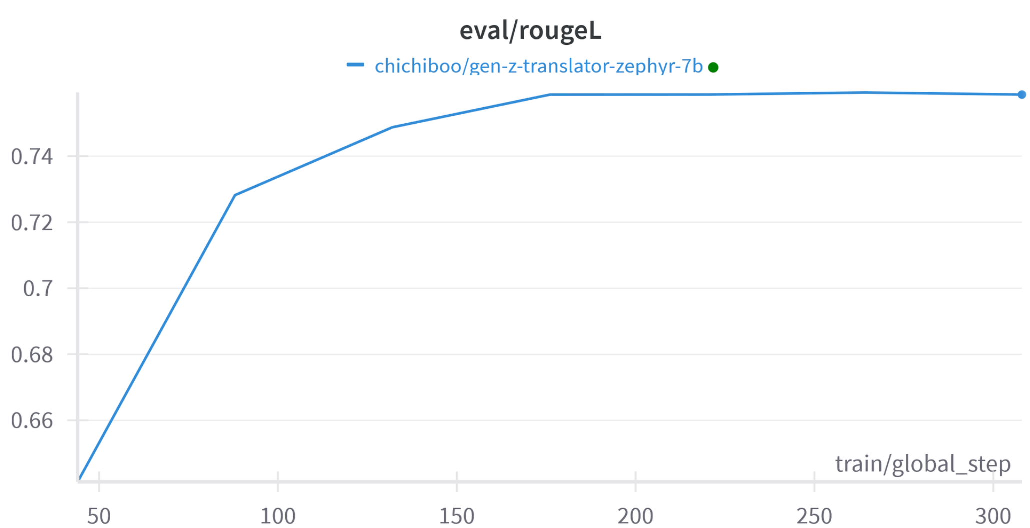

Figure 4.3: Evaluated using ROUGE-L metric

4.2.2 Text Generation

A total of 170 sentences were translated using both the base zephyr-7b-beta model and the finetuned model. The translations are then filtered to remove duplicate answers between models or has minor differences such as punctuation or filler words that does not contribute to the meaning of the sentence. A total of 143 sentences then served as the dataset used to evaluate the performance of the model and comparing it with the other base model.

While the model successfully translated the sentence structure and preserved much of the semantic content, it underperformed by injecting additional commentary not present in the source. This kind of over-interpretation reflects a critical challenge in slang translation tasks where preserving tone and intent is essential.

4.2.3 Automatic Evaluation Metrics

The dataset was automatically evaluated using BLEU and ROUGE metrics, specifically the ROUGE-L metric as the dataset do not contain newlines that ROUGE-Lsum uses to separate the input with. These scores were then averaged to determine the score of the models. The base model obtained a BLEU score of 0.8099 and ROUGE-L Score of 0.8336 and the fine-tuned model obtained a BLEU score of 0.8151 and ROUGE-L Score of 0.8396. While the difference between the models is minimal, this does not completely represent the performance of the models as these metrics are only used to determine if the generated text is close to the reference text, regardless of the context and the overall quality of the generated text. However, it does show that the fine-tuned model has little improvement over the base model.

4.2.4 Manual Evaluation Metrics

A manual evaluation was conducted by the researchers through a survey administered via Google Forms to determine which of the two models is preferred by Generation Z students at University of the Philippines Visayas (UPV). The knowledge of respondents answering the survey ranges from people who knew some slangs, to people who use slangs in their everyday conversations. The survey comprised a total of 143 questions, which were distributed across five separate forms. The first form contained 20 questions, the second form contained 19, the third form contained 20, the fourth form contained 20, the fifth form contained 14, and the sixth form contained 50 amounting to 143 questions in total. Each question presented two translation options: one generated by the fine-tuned model and the other by the base model. Respondents were asked to select the translation they preferred in each case. A total of 135 responses were gathered in the survey, with 29, 22, 22, 21, 20, and 21 responses completing Forms 1 through 6, respectively.

The data presented below illustrate respondent preferences between the base and fine-tuned models across the six survey forms, as well as the overall summary of the results. Each graph visualizes the outcomes for an individual form, specifically indicating both the raw number of responses and the corresponding percentages favoring each model. A systematic evaluation for each graph is provided as follows:

Figure 4.4 shows that among the 29 responses, 306 responses or 52.8 percent preferred the base model, while 274 responses or 47.2 percent favored the fine-tuned

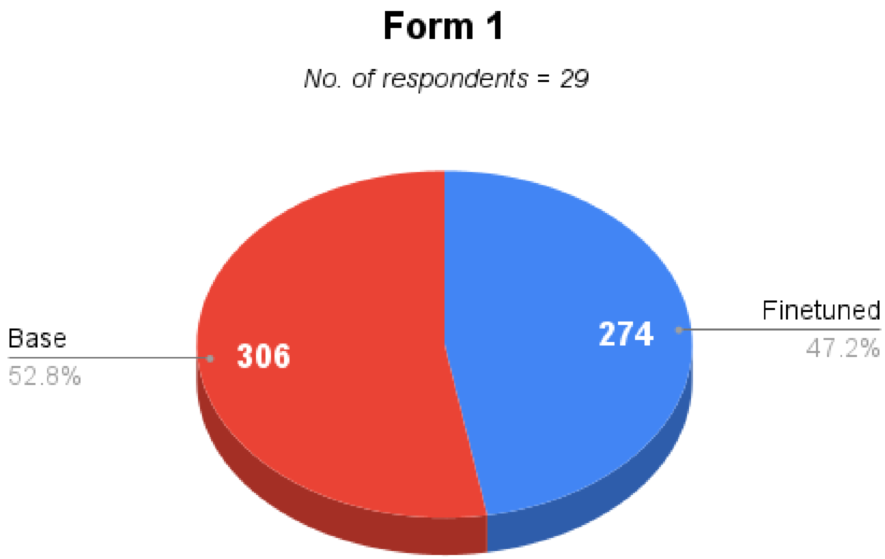

Figure 4.4: Form 1 Evaluation

model. This indicates a slight preference for the base model in this particular form. Notably, this result deviates from the overall trend observed in the other four forms, where the fine-tuned model tends to be favored. Form 1 is the only instance in which the base model outperformed the fine-tuned model, suggesting that specific characteristics of this form may have influenced the preferences of the respondents.

Figure 4.5 implies that among 22 responses, 236 responses, or 56.5 percent, favored the fine-tuned model, while 182 responses, or 43.5 percent, preferred the base model. This 13 percent margin reflects the clear preference for the fine-tuned model, which is consistent with the overall trend observed across the other forms.

Figure 4.6 illustrates that among the 22 responses, the fine-tuned model received a significantly higher preference, with 259 responses or 60.2 percent, compared

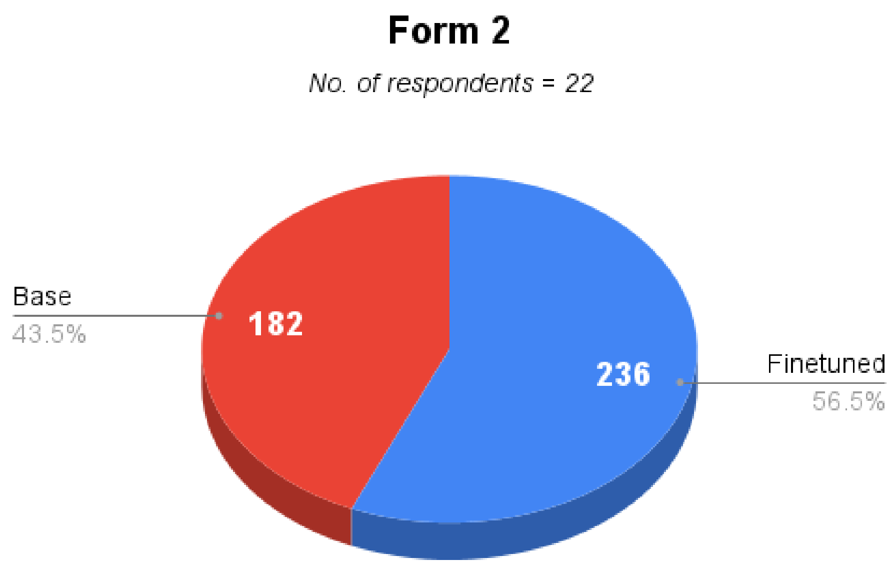

Figure 4.5: Form 2 Evaluation

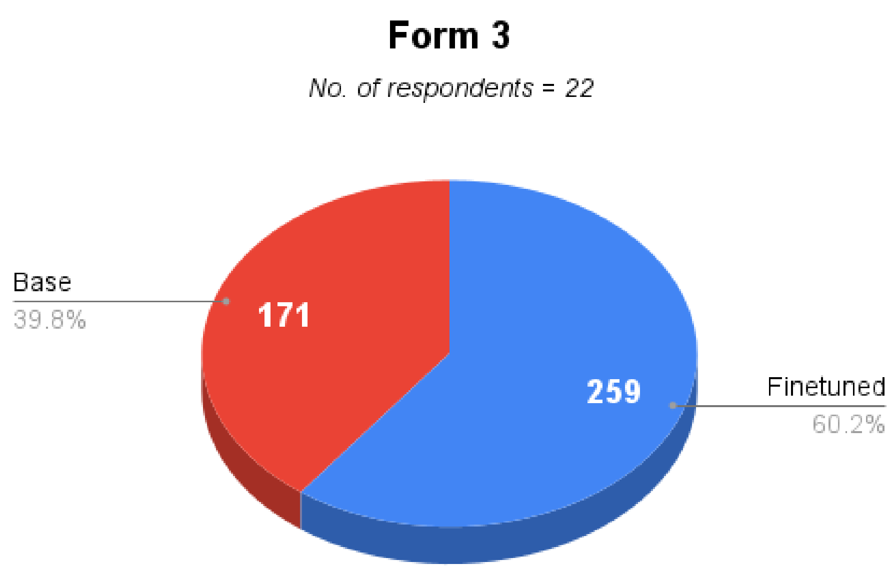

Figure 4.6: Form 3 Evaluation

to the base model with 171 responses or 29.8 percent. This 20.4 percent margin represents the widest gap among all forms. This strongly indicates the superior performance of the fine-tuned model on translating, presented in Form 3.

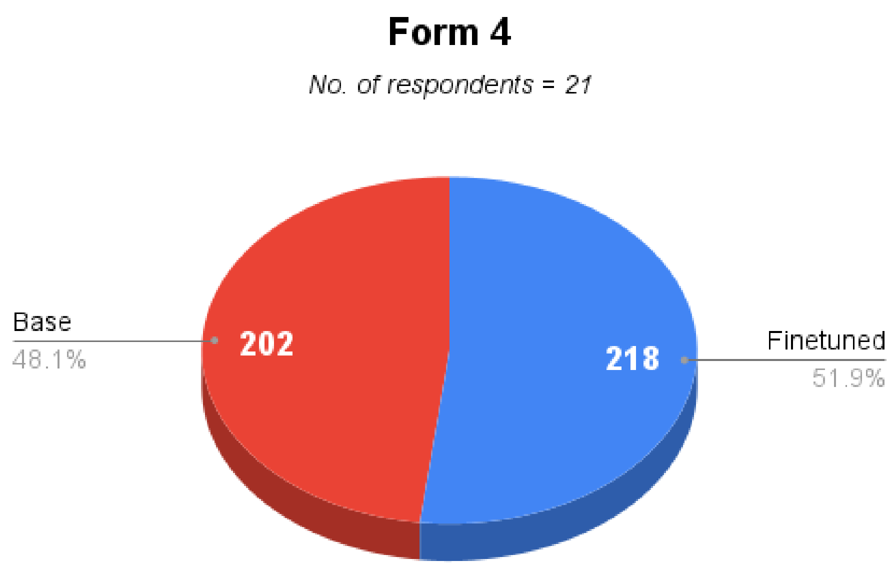

Figure 4.7: Form 4 Evaluation

Figure 4.7 highlights that the 21 responses in Form 4 yielded a nearly even distribution of preferences, with 218 responses or 51.9 percent favoring the fined-tuned model and 202 responses or 48.1 percent preferring the base model. This narrow 3.8 percent difference suggests a comparable level of performance between the two models in this particular form.

Figure 4.8 conveys that among the 20 responses in Form 5, 152 responses or 54.3 percent selected the fine-tuned model, while 128 responses or 45.7 percent chose the base model. This 8.6 percent margin reinforces the general trend toward the fine-tuned model across all forms.

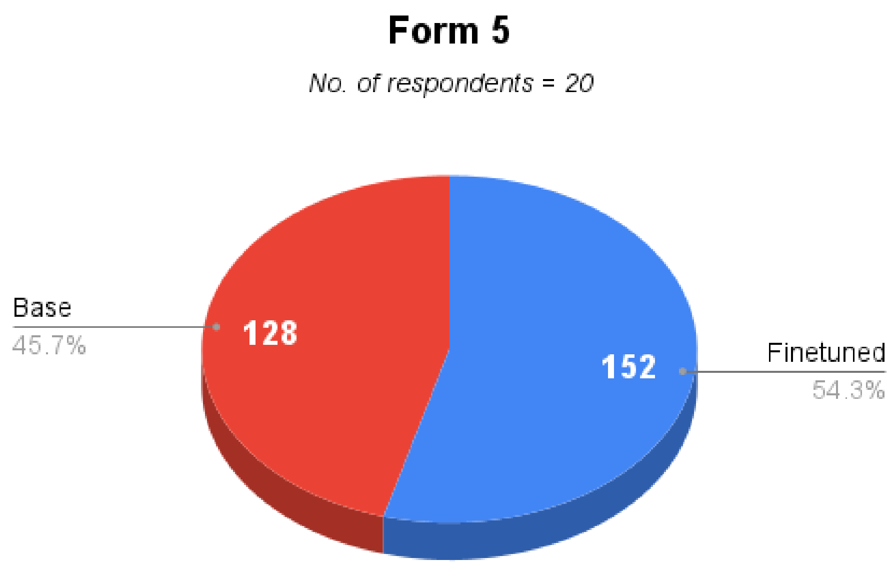

Figure 4.8: Form 5 Evaluation

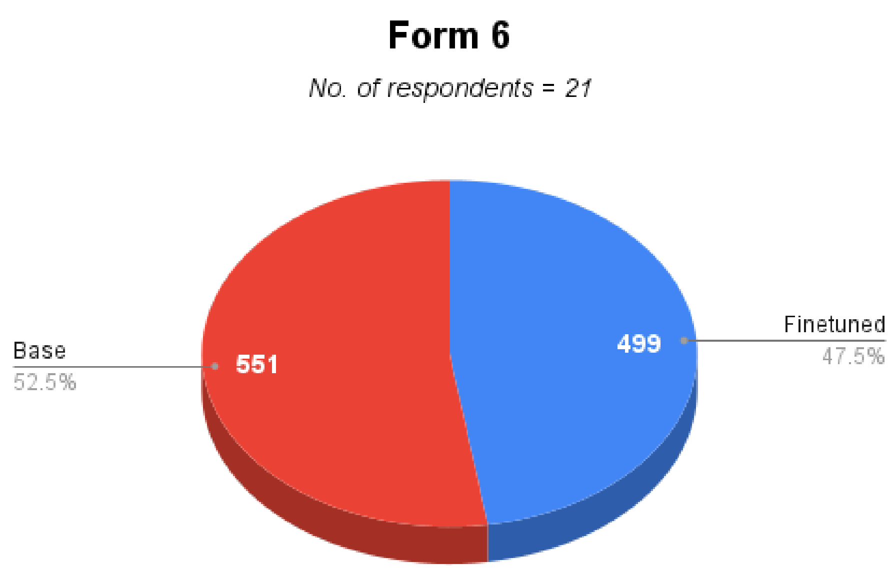

Figure 4.9: Form 6 Evaluation

Figure 4.9 indicates the results of the sixth form. 21 responses in Form 6 showed a slight preference for the base model, garnering 52.5%, over the fine-tuned model, with 47.5%. Along with Form 1, this result contrasts with the overall trend observed across all gathered data.

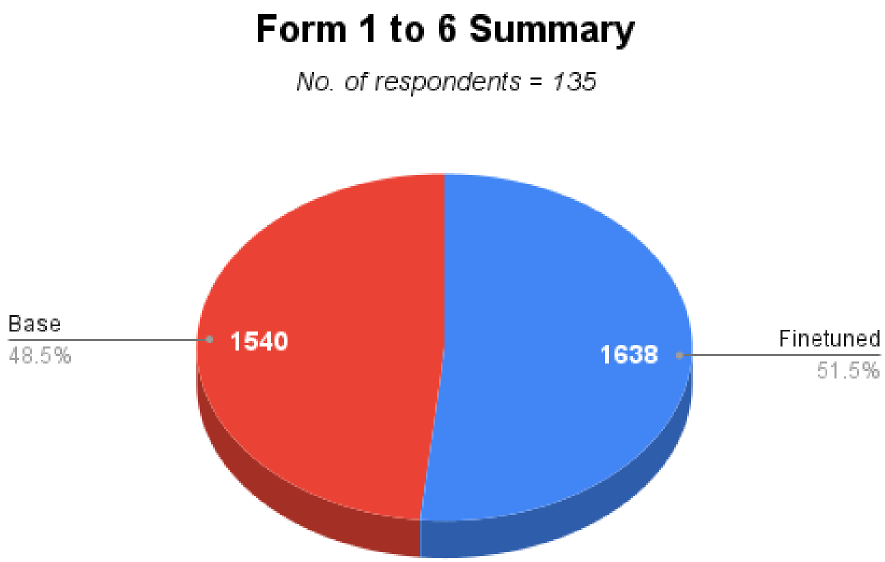

Figure 4.10: Summary Evaluation

Figure 4.10 presents the overall summary across all five forms, with a total of 135 responses garnered in the survey. In total, the fine-tuned model received 51.5%, while the base model garnered 989 preferences or 49.5%. The resulting 7% margin between the two model indicates a moderate overall preference among Gen Z students at UPV for the fine-tuned model, suggesting its relatively better performance in meeting the participants' expectations for translation quality.

Table 4.1: Manual Evaluation Results

Form	**Responses**	**Base Model Preference**	**Fine-Tuned Model Preference**	**Interpretation**
1	29	52.8%	47.2%	Responses showed more preference to base model translation
2	29	43.5%	56.5%	Clear fine-tuned preference.
3	23	50.0%	50.0%	No preference evident.
4	20	48.6%	51.4%	Slight fine-tuned preference.
5	20	45.7%	54.3%	Moderate preference for fine-tuned model.
6	21	52.5%	47.5%	Slight base model preference.

4.3 Summary

The chapter presented the evaluation results and discussions on the performance of the fine-tuned language model for translating Gen Z internet slang into their formal translations. The dataset used for training consisted of 1,703 sentence pairs, combining original and publicly available data. The model was trained for seven epochs, with early stopping employed to prevent overfitting, which was evident from the divergence between training and validation losses.

Evaluation was conducted using both automatic and manual methods. The automatic evaluation, using BLEU and ROUGE-L metrics, showed marginal improvements in the fine-tuned model compared to the base model, suggesting slightly better alignment with reference translations.

To support the results of automatic evaluation metrics, a manual evaluation was carried out through online surveys among Generation Z students at UPV. Participants compared translations from both models across six forms. Results showed a moderate overall preference for the fine-tuned model, with 51.5% of responses in its favor. While one form showed a slight preference for the base model, the fine-tuned model was generally preferred, especially in Form 3 where it showed the largest margin.

In summary, the findings indicate that the fine-tuned model slightly outperformed the base model in terms of automatic metrics and showed a modest but consistent preference among target users, supporting its effectiveness in translating Gen Z slang into more formal language.

Chapter 5

Conclusion

In this study, we constructed a dataset, containing 1,703 pairs of Gen Z internet slang sentences and their corresponding formal translations. We fine-tuned a zephyr-7B-Beta model and evaluated its performance against the base model. Model training was stopped early to prevent overfitting, and the best model was selected based on validation performance. Both automatic and manual evaluation methods were employed to assess translation quality. Automatic metrics, using BLEU and ROUGE-L, showed that the fine-tuned model slightly outperformed the base model with scores of 0.8151 and 0.8396. Manual evaluation, conducted via online surveys with Generation Z students at UPV, indicated a moderate overall preference for the fine-tuned model, which received 51.5% of the total responses. These results suggest that while the improvement in performance was not drastic, the fine-tuned model better aligned with the expectations and preferences of the target demographic.

5.1 Limitations

Language is dynamic and constantly evolving, making it difficult to establish clear boundaries on when slang terms emerge or fade within a generation. Additionally, the dataset created for this study was relatively small, and the number of evaluators involved was limited. In addition, as stated in Section 3.1.3, the computational constraints posed a challenge—loading a model with 7 billion parameters requires approximately 56 GB of memory, while Google Colab provided 16GB of VRAM which is insufficient for high-capacity models.

5.2 Recommendations

Future researchers are encouraged to expand the vocabulary of slang terms used on the Internet and explore more recent trends, taking into account the dynamic nature of language. It is also recommended that future studies utilize a larger and more diverse dataset to improve the robustness of the findings. Additionally, the future researchers could explore the use of the model for translating Filipino slang or different slang languages, enhancing further the understanding and cross-cultural communication.

Chapter 6

References

Ambarsari, S., Amrullah, A., & Nawawi, N. (2020, Aug). The Use of Online Slang for Independent Learning in English Vocabulary. *Proceedings of the 1st Annual Conference on Education and Social Sciences (ACCESS 2019)*, *465*, 295–297. doi: 10.2991/assehr.k.200827.074

Barseghyan, L. (2014). *On Some Aspects of Internet Slang.* Retrieved from `https://api.semanticscholar.org/CorpusID:51730779`

binti Sabri, N. A., bin Hamdan, S., Nadarajan, N.-T. M., & Shing, S. R. (2020, Jun). The Usage of English Internet Slang among Malaysians in Social Media. *Selangor Humaniora Review*, *4*(1), 16–17.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative AI at work.*

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... Stoica, I. (2024). *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference.*

Crystal, D., & Robins, R. H. (2024, Oct). *Language.* Encyclopædia Britannica, inc. Retrieved from `https://www.britannica.com/topic/language`

Daniel Han, M. H., & Team, U. (2023). *Unsloth.* Retrieved from http://github.com/unslothai/unsloth

Ding, N., Chen, Y., Xu, B., Qin, Y., Zheng, Z., Hu, S., … Zhou, B. (2023). Enhancing Chat Language Models by Scaling High-Quality Instructional Conversations. *arXiv preprint arXiv:2305.14233*.

Dua, A., Jacobson, R., Ellingrud, K., Enomoto, K., Cordina, J., Coe, E. H., & Finneman, B. (2024, Aug). *What is Gen Z?* McKinsey & Company. Retrieved from https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z

Esquivel, O. J. (2020). *Exploring the Filipinization of the English Language in a Digital Age: An Identity Apart from Other World Englishes.* Retrieved from https://files.eric.ed.gov/fulltext/ED475048.pdf

Euchner, J. (2023). Generative AI. *Research-Technology Management*, *66*(3), 71–74.

Fernández-Toro, M. (2016, Jun). *Exploring Languages and Cultures.* Retrieved from https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). *Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration* (Vol. 25) (No. 3). Taylor & Francis.

Ganti, A. (2024). *Weighted Average: Definition and How It is Calculated and Used.* Investopedia. Retrieved from https://www.investopedia.com/terms/w/weightedaverage.asp

Ghazali, N. M., & Abdullah, N. N. (2021, Dec). Slang Language Use in Social Media among Malaysian Youths: A Sociolinguistic Perspective. *International Young Scholars Journal of Lan-*

guages, *4*(2), 69. Retrieved from https://www.iium.edu.my/media/77652/Slang%20Language%20Use%20in%20Social%20Media%20Among%20Malaysian%20Youths_A%20Sociolinguistic%20Perspective.pdf

Gonzaga, M. (2025, Feb). *"Forda Convo ang Ferson": Analysis of Gen Z Slang in the Lens of BatStateU Faculty Members.* Retrieved from https://www.academia.edu/102575643/_FORDA_CONVO_ANG_FERSON_ANALYSIS_OF_GEN_Z_SLANG_IN_THE_LENS_OF_BATSTATEU_FACULTY_MEMBERS

Heydari, M., Albadvi, A., & Khazeni, M. (2024). Persian Slang Text Conversion to Formal and Deep Learning of Persian Short Texts on Social Media for Sentiment Classification. *Journal of Electrical and Computer Engineering Innovations (JECEI)*. Retrieved from https://jecei.sru.ac.ir/article_2172.html doi: 10.22061/jecei.2024.10745.731

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., … Chen, W. (2021). *Lora: Low-Rank Adaptation of Large Language Models.* Retrieved from https://arxiv.org/abs/2106.09685

Ibrahim, A., & Sharief, B. (2023, 10). Intelligent System to Transformer Slang Words into Formal Words. *NTU Journal of Engineering and Technology*, *2*. doi: 10.56286/ntujet.v2i2.689

Jeresano, E., & Carretero, M. (2022, Feb). Digital Culture and Social Media Slang of Gen Z. *United International Journal for Research & Technology*, *3*(4), 11–25. doi: http://dx.doi.org/10.1314/RG.2.2.36361.93285

Liang, Y., Meng, F., Wang, J., & Zhou, J. (2025). *Slangdit: Benchmarking LLMs in Interpretative Slang Translation.* Retrieved from https://arxiv.org/abs/2505.14181

Lin, C.-Y. (2004, Jul). ROUGE: A Package for Automatic Evaluation of Summaries. *Meeting of the Association for Computational Linguistics*, 74–81.

Liu, J., Zhang, X., & Li, H. (2023, Aug). Analysis of Language Phenomena in Internet Slang: A Case Study of Internet Dirty Language. *Open Access Library Journal*, *10*(08), 1–12. doi: 10.4236/oalib.1110484

Liu, S., Gui, D.-Y., Zuo, Y., & Dai, Y. (2019, Jun). Good Slang or Bad Slang? Embedding Internet Slang in Persuasive Advertising. *Frontiers in Psychology*, *10*. doi: 10.3389/fpsyg.2019.01251

Mantiri, O. (2010, 03). Factors affecting Language Change. *http://ssrn.com/abstract=2566128*. doi: 10.2139/ssrn.2566128

Maulidiya, R., Wijaya, S. E., Mauren, C., Adha, T. P., & Pandin, M. G. R. (2021, Dec). *Language Development of Slang in the Younger Generation in the Digital Era.* OSF Preprints. Retrieved from `osf.io/xs7kd` doi: 10.31219/osf.io/xs7kd

McArthur, T. (2003). *Concise Oxford Companion to the English Language* (1st ed.). Oxford University Press.

Nadas, M., Diosan, L., & Tomescu, A. (2025). *Synthetic Data Generation using Large Language Models: Advances in Text and Code.* Retrieved from `https://arxiv.org/abs/2503.14023`

Nguyen, T. T., Wilson, C., & Dalins, J. (2023). *Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts.* Retrieved from `https://arxiv.org/abs/2308.14683`

Nocon, N., Kho, N. M., & Arroyo, J. (2018, Oct). Building a Filipino Colloquialism Translator using Sequence-to-Sequence Model. *TENCON 2018 - 2018 IEEE Region 10 Conference*, 2199–2204. doi: 10.1109/tencon.2018.8650118

Olobia, L. (2024, Jul). *Utilizing Social Media in Communicating in English as a Second Language.* Retrieved from `https://doi.org/10.37237/150103`

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*. Retrieved from https://dl.acm.org/citation.cfm?id=1073135 doi: https://doi.org/10.3115/1073083.1073135

Peytchev, A., & Peytcheva, E. (2017). *Reduction of Measurement Error due to Survey Length: Evaluation of the Split Questionnaire Design Approach.* Retrieved from https://ojs.ub.uni-konstanz.de/srm/article/view/7145/0

Sun, Z., Hu, Q., Gupta, R., Zemel, R., & Xu, Y. (2024). *Toward Informal Language Processing: Knowledge of Slang in Large Language Models.* Retrieved from https://arxiv.org/abs/2404.02323

Suslak, D. F. (2009). The Sociolinguistic Problem of Generations. *Language & Communication*, *29*(3), 199–209. Retrieved from https://www.sciencedirect.com/science/article/pii/S0271530909000196 (Reflecting on language and culture fieldwork in the early 21st century) doi: https://doi.org/10.1016/j.langcom.2009.02.003

Teng, C. E., & Joo, T. M. (2023). Is Internet Language a Destroyer to Communication? In X.-S. Yang, R. S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of eighth international congress on information and communication technology* (pp. 527–536). Singapore: Springer Nature Singapore.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., ... Wolf, T. (2023). *Zephyr: Direct Distillation of LM Alignment.*

Vacalares, S. T., Salas, A. F. R., Babac, B. J. S., Cagalawan, A. L., & Calimpong, C. D. (2023, Jun). The Intelligibility of Internet Slangs between Millennials and Gen Zers: A Comparative Study. *International Journal of Science and Research Archive*, *9*(1), 400–409. doi: 10.30574/ijsra.2023.9.1.0456

Vergho, T., Godbout, J.-F., Rabbany, R., & Pelrine, K. (2024). *Comparing GPT-4 and Open-Source Language Models in Misinformation Mitigation.* Retrieved from `https://arxiv.org/abs/2401.06920`

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., ... Gallouédec, Q. (2020). *TRL: Transformer Reinforcement Learning.* `https://github.com/huggingface/trl`. GitHub.

Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., ... Rishi, D. (2024). *Lora Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report.* Retrieved from `https://arxiv.org/abs/2405.00732`