

1 LOST IN TRANSLATION: TRANSLATING GENERATION
2 Z INTERNET SLANG USING MACHINE LEARNING

3 A Special Problem
4 Presented to
5 the Faculty of the Division of Physical Sciences and Mathematics
6 College of Arts and Sciences
7 University of the Philippines Visayas
8 Miag-ao, Iloilo

9 In Partial Fulfillment
10 of the Requirements for the Degree of
11 Bachelor of Science in Computer Science by

12 FLAUTA, Neil Bryan
13 GIMENO, Ashley Joy
14 GIMENO, Carl Jorenz

15 Francis DIMZON, Ph.D.
16 Adviser

17 May 26, 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)

26 Division of Physical Sciences and Mathematics

27 College of Arts and Sciences

28 University of the Philippines Visayas

29 **Declaration**

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

Name

Signature

Date

Flauta, Neil Bryan

(Student)

34 Gimeno, Ashley Joy

(Student)

Gimeno, Carl Jorenz

(Student)

Dedication

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

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

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

Acknowledgment

45 We extend our heartfelt gratitude to Dr. Francis D. Dimzon for his patient
46 guidance throughout this study. His thoughtful mentorship in the field of machine
47 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

66

Contents

67	1 Introduction	1
68	1.1 Overview	1
69	1.2 Problem Statement	4
70	1.3 Research Objectives	4
71	1.3.1 General Objectives	4
72	1.3.2 Specific Objectives	5
73	1.4 Scope and Limitations of the Research	5
74	1.5 Significance of the Research	6
75	2 Review of Related Literature	9
76	2.1 Communication Gap between Generations	9
77	2.2 Generative AI	10

78	2.3 Existing Studies	10
79	2.4 LoRA for Fine Tuning	13
80	2.5 Data Augmentation through Synthetic Data Generation	13
81	2.6 Evaluation Metrics	14
82	2.7 Chapter Summary	15
83	3 Research Methodology	19
84	3.1 Research Activities	19
85	3.1.1 Data Gathering	20
86	3.1.2 Data Preprocessing	20
87	3.1.3 Model Fine-Tuning	21
88	3.1.4 Model Evaluation	22
89	4 Results and Discussions	25
90	4.1 Dataset	25
91	4.2 Model Evaluation	26
92	4.2.1 Model Training	26
93	4.2.2 Text Generation	28
94	4.2.3 Automatic Evaluation Metrics	28

95	4.2.4 Manual Evaluation Metrics	29
96	4.3 Summary	35
97	5 Conclusion	37
98	5.1 Limitations	38
99	5.2 Recommendations	38
100	6 References	39

101 List of Figures

102	3.1 Summarized Methodology	19
103	4.1 Training loss curve of the fine-tuned model across training steps .	26
104	4.2 Evaluation loss curve of the fine-tuned model across training steps	27
105	4.3 Evaluated using BLEU metric	27
106	4.4 Evaluated using ROUGE-L metric	28
107	4.5 Form 1 Evaluation	30
108	4.6 Form 2 Evaluation	31
109	4.7 Form 3 Evaluation	31
110	4.8 Form 4 Evaluation	32
111	4.9 Form 5 Evaluation	33
112	4.10 Form 6 Evaluation	34
113	4.11 Summary Evaluation	34

¹¹⁴ List of Tables

¹¹⁵	2.1 Summary of Existing Studies	17
----------------	---	----

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 in-

130 formally, and opposing established authority (McArthur, 2003). Slang is highly
131 contextual and pervasive, even in non-standard English. Its figurative nature and
132 how it twists the definitions of the words used make it difficult for outsiders to
133 understand.

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

143 Generation Z, individuals born between 1996 and 2009, are regarded as “digital
144 natives” because technology is an integral part of their upbringing (Dua et al.,
145 2024). Even the language of this generation is greatly affected by technology,
146 where newly coined terms and phrases, called Gen Z slang, are tied to the me-
147 dia culture they’ve grown up with (Jeresano & Carretero, 2022). However, this
148 evolution of language often creates communication barriers with older generations
149 (Venter, 2017 as cited in (Ghazali & Abdullah, 2021)). Furthermore, studies show
150 that even within Generation Z, people with limited exposure to social media may
151 struggle to understand the prevalent slang (Vacalares, Salas, Babac, Cagalawan,
152 & Calimpong, 2023).

153 These gaps highlight the need for a tool that can bridge the generational divide,

154 making it easier for individuals to understand the language of Generation Z. Mul-
155 tiple studies have tried translating slang into a formal language using machine
156 learning. Khazeni et al. achieved a 81.91% accuracy in translating Persian slang
157 to formal Persian language using deep learning. Another study by Nocon et al.
158 created a translator to translate Filipino colloquialisms into the Filipino language
159 using Tensorflow's sequence-to-sequence model and Moses' phrase-based statis-
160 tical machine translation. Furthermore, Ibrahim and Sharief developed a slang
161 translator using models from Hugging Face.

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

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

169 The main contributions of this study are as follows:

- 170 • Enhance linguistic understanding between generations by using fine-tuning
171 a LLM to translate Gen Z slang to formal language, leveraging the strengths
172 of advanced NLP techniques
- 173 • Bridge communication gaps between generations using the proposed model
174 to foster better relationships
- 175 • Create a scalable framework that can be adapted to translate slang in other
176 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

217 status as a marker of education and social standing in the Philippines. As noted
218 by Mateo (2018) cited by (Esquivel, 2019), the widespread use of English in social
219 media underscores its significance in Filipino society, where proficiency in English
220 is often associated with educational attainment and social mobility.

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

224 1.5 Significance of the Research

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

232 The findings of this study offer practical benefits for various stakeholders. For edu-
233 cators, the insights can support the development of more inclusive and responsive
234 classroom communication strategies, enabling them to better understand and en-
235 gage with their students' language use and cultural context. For parents, the study
236 provides a framework for interpreting the language their children use online and
237 in casual conversations, helping in bridging communication gaps and improving
238 parent-child relationships. For media practitioners and digital marketers, under-

239 standing the patterns and meanings behind Gen Z slang can inform the creation of
240 more relatable and culturally relevant content, enhancing audience engagement.

241 By addressing the communicative divide brought about by generational language
242 differences, this research encourages a more informed approach to language vari-
243 ation in contemporary digital spaces. Ultimately, the study underscores the im-
244 portance of adapting to linguistic change in order to foster clearer, more effective
245 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 lin-

260 guistic familiarity as Suslak (Suslak, 2009) argues that age influences language
261 use, noting that language evolves across generations. Supporting this, a study by
262 Teng and Joo (Teng & Joo, 2023) found that the older a person is, the less likely
263 they are to understand internet language.

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

267 **2.2 Generative AI**

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

276 **2.3 Existing Studies**

277 Zephyr-7b-beta has shown performance comparable to that of larger models, most
278 notably, GPT-4 (?, ?). This is further corroborated by the study by Vergho
279 et al. (Vergho, Godbout, Rabbany, & 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.

Khazeni et al. (Heydari, Albadvi, & 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 et al. (Nocon, Kho, & 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.

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

315 While general-purpose instruction tuning is now well-documented, less attention
316 has been paid to fine-tuning LLMs for tasks involving informal or non-standard
317 language such as slang. However, studies are emerging that suggest promising
318 outcomes. For example, the SlangDIT benchmark (?, ?) developed a testbed
319 specifically for slang understanding and translation, and preliminary findings in-
320 dicate that even relatively small models fine-tuned on slang-rich datasets can
321 rival zero-shot GPT-4 performance. This supports the notion that domain adap-
322 tation—particularly to informal linguistic domains—benefits substantially from
323 task-specific training, even if the examples are synthetic. A study of Sun et al.
324 (?, ?) also showed that even a small dataset of slang sentences helped GPT 3.5
325 perform better than zero-shot GPT-4.0 at slang detection. While it is a classi-
326 fication task, this suggests a promising approach to improve the performance of
327 LLMs in slang translation tasks.

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 (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. (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 et al. (Nguyen, Wilson, & 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 (?, ?), which is

349 a synthetic dataset from data obtained from the Internet (?, ?). In addition, the
350 model showed performance comparable to larger open-source models in language
351 tasks.

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

362 2.6 Evaluation Metrics

363 Automatic evaluation metrics are essential for assessing the performance of ma-
364 chine translation systems, especially in the context of slang translation. These
365 metrics provide a quantitative measure of translation quality, allowing for efficient
366 comparison between different models and approaches. Commonly used metrics
367 include BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented
368 Understudy for Gisting Evaluation). BLEU measures the overlap between the
369 machine-generated translation and one or more reference translations, focusing
370 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 (?). 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)(?). 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, Khazeni et al. (Heydari et al., 2024) used deep learning to translate Persian slang, while Nocon et al. (Nocon et

394 al., 2018) created a Filipino slang translator using statistical models. Moreover,
395 Ibrahim and Mustafa (Ibrahim & Sharief, 2023) fine-tuned pre-trained models to
396 learn slang meanings. One promising technique for this is Low Rank Adaptation
397 (LoRA), which is a fine-tuning method that keeps the original model stable while
398 using less storage. Studies by Zhao et al. (Zhao et al., 2024) and Nguyen et al.
399 (Nguyen et al., 2023) show that LoRA models are not only efficient but can even
400 outperform advanced models like GPT-4 when it comes to slang translation and
401 text classification. However, datasets specifically for slang translation are often
402 limited, making synthetic data generation a valuable tool.

Table 2.1: Summary of Existing Studies

Author	Focus	Gaps	Problem Solved
Nocon et al.	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 et.al	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.
Khazeni et al.	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.

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

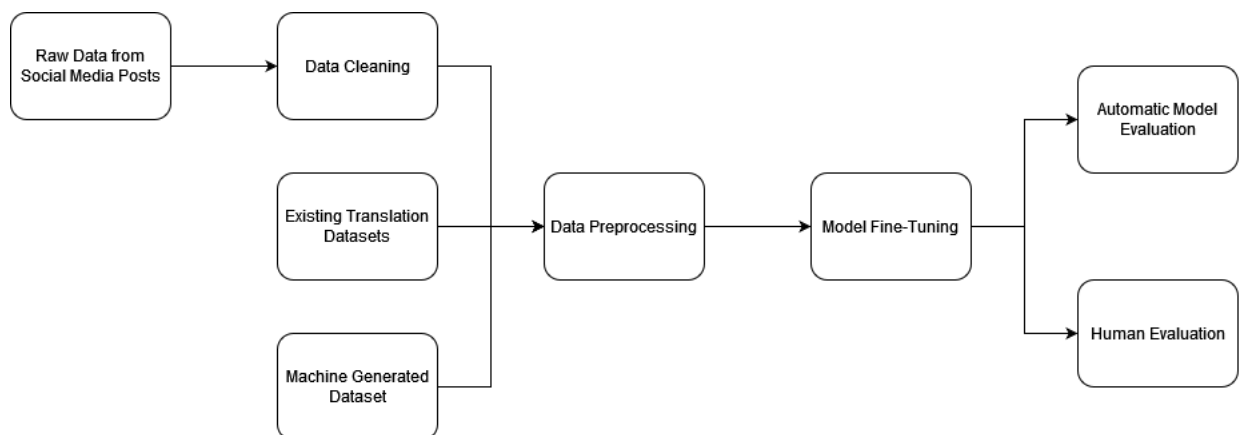


Figure 3.1: Summarized Methodology

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. 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

431 generalize to new data. The train dataset is then split again into a 90/10 ratio
432 to ensure no overfitting while still allowing the model to adapt to the pattern
433 of slang. The cleaned up dataset was then tokenized through the Transformers
434 library provided by HuggingFace as the library already has tokenizers available
435 for their pretrained models. This ensures that the data is formatted properly as
436 required by the model to be used.

437 3.1.3 Model Fine-Tuning

438 The model used in this study was zephyr-7b-beta because it is open-source and
439 was proven to perform better than other models of the same size. In addition, it
440 can be trained in a GPU with 16GB of VRAM, necessary as we are using the free
441 plan of Google Colab as the platform of choice for prototype fine-tuning of the
442 model. However, during the training process with the full dataset, the Pro+ plan
443 of Google Colab was used for faster training time and background execution of the
444 training process, allowing the training to continue uninterrupted regardless of the
445 network connection. This study used the example codes provided by HuggingFace
446 in the documentation of their various libraries and sample notebook provided in
447 the zephyr-7b-beta repository.

448 The model was loaded using the Transformers library and was quantized into 4
449 bits through BitsandBytes library to fit the entire model in the allocated resources
450 while having enough headroom for training. In addition, the Unsloth library was
451 used to speed up the training time and reduce the resources used even more
452 (Daniel Han & team, 2023). A LoRA adapter was then attached to the model to
453 further reduce the parameters to be trained.

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

462 The model was then trained using SFTTrainer class from the Transformer Rein-
463 forcement Learning (TRL) library of HuggingFace to simplify the training process
464 (?, ?). The model was trained with the following parameters: optimizer is paged
465 4bit AdamW, batch size of 8, learning rate of 2e-5, and maximum number of
466 epochs of 50. These parameters were chosen based on the GPU provided in Co-
467 lab, the test notebook by HuggingFace and the default parameters of SFTTrainer.

468 3.1.4 Model Evaluation

469 The model was evaluated using both automatic and manual evaluation metrics.
470 Identical answers and answers with minimal difference, such as punctuation, be-
471 tween the fine-tuned and the base model were removed in the test set to ensure
472 that the model is evaluated properly. After filtering, a total of 143 sentences
473 were used to evaluate the model. The model was then prompted to generate a
474 formal sentence for 170 sentences in the test dataset. The generated sentences
475 were then compared to the formal translation of the sentence using BLEU and
476 ROUGE metrics. The base zephyr-7b-beta model was also prompted to gener-

477 ate sentences for the BLEU and ROUGE metric and the pairwise comparison for
478 human evaluation.

479 An online survey was conducted using Google Forms to compare the outputs of the
480 fine-tuned model and the base model in order to evaluate the effectiveness of the
481 fine-tuning process. Participants were presented with sentence pairs generated
482 by both models and were asked to choose the more accurate translation of a
483 given Generation Z slang sentence based on accuracy, naturalness, and contextual
484 appropriateness. To minimize potential ordering bias, the sequence in which the
485 outputs from the two models were displayed was randomized for each pair. The
486 researchers implemented a Split Questionnaire Design (SQD) by dividing the full
487 survey into multiple sets to improve response rates and reduce respondent fatigue
488 (? , ?). A total of 143 questions was unevenly distributed into six forms. In
489 addition, the number of responses per form varied which leads to an unbalanced
490 where direct comparisons between all items were not possible.

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

499 This method offered a simple yet effective way to integrate responses from an SQD
500 structure without requiring overlap or complex modeling assumptions. It also

501 ensured that items answered by more respondents contributed more substantially
502 to the overall evaluation while avoiding bias from unequal form lengths.

503 Chapter 4

504 Results and Discussions

505 4.1 Dataset

506 We built a dataset containing a total of 1155 Gen Z internet slang sentences and
507 their corresponding formal translations. The created dataset was then combined
508 with another dataset from Hugging Face that contains 548 Gen Z internet slang
509 and their corresponding formal translation for a total of 1703 sentence pairs.
510 The dataset was then split into training, validation, and test sets with a ratio of
511 81:9:10. The training set contains 1380 sentence pairs, the validation set contains
512 153 sentence pairs, and the test set contains 170 sentence pairs. The dataset was
513 then tokenized using the tokenizer of the base model, zephyr-7b-beta, to prepare
514 it for training. The tokenized dataset was then saved in a JSON format to be
515 used for training the model.

4.2 Model Evaluation

4.2.1 Model Training

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 overfitting seen in the following figure. Figure 4.1 shows that the training loss is decreasing and in Figure 4.2 the validation loss is increasing and other metrics are not improving. These indicate that the model is overfitting 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 amongst the epochs, the one with the lowest validation loss and highest metrics, was chosen as the final model.

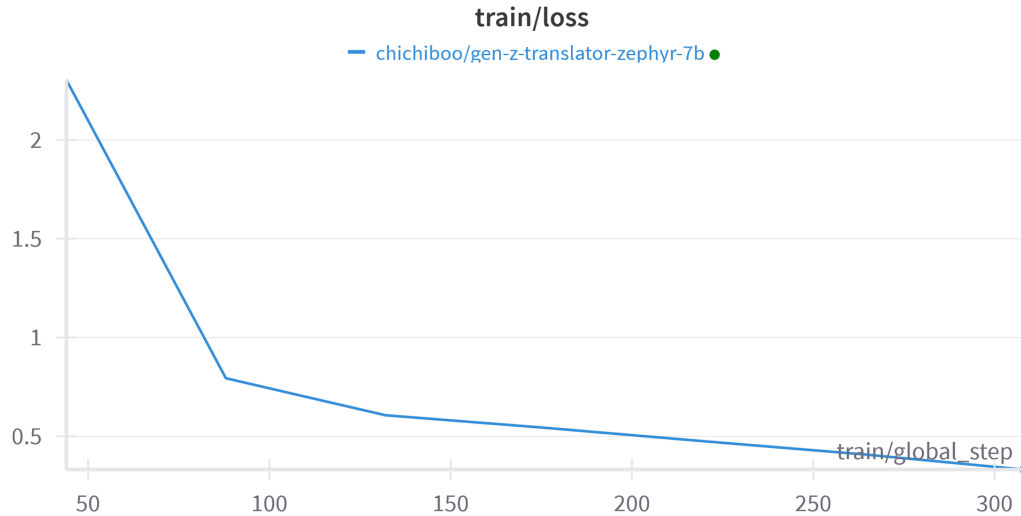


Figure 4.1: Training loss curve of the fine-tuned model across training steps

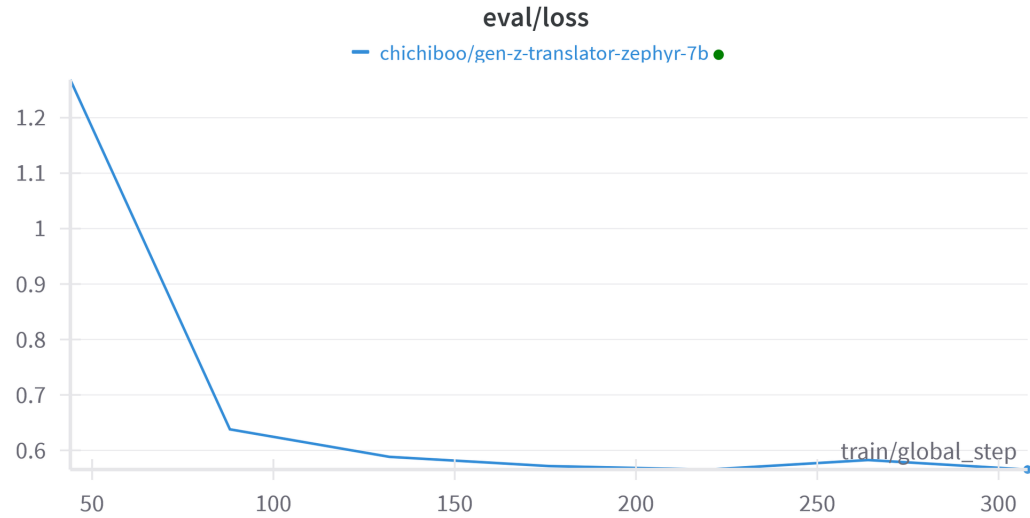


Figure 4.2: Evaluation loss curve of the fine-tuned model across training steps

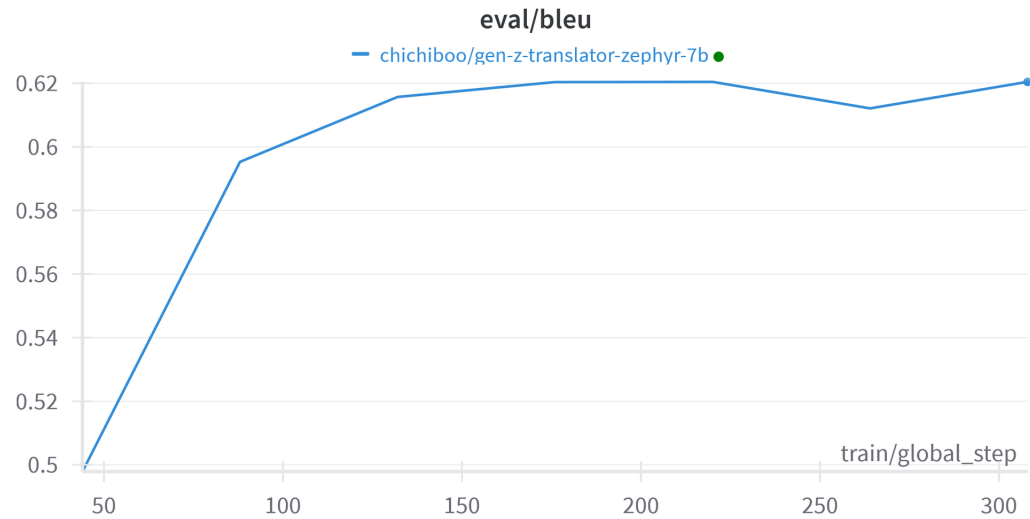


Figure 4.3: Evaluated using BLEU metric

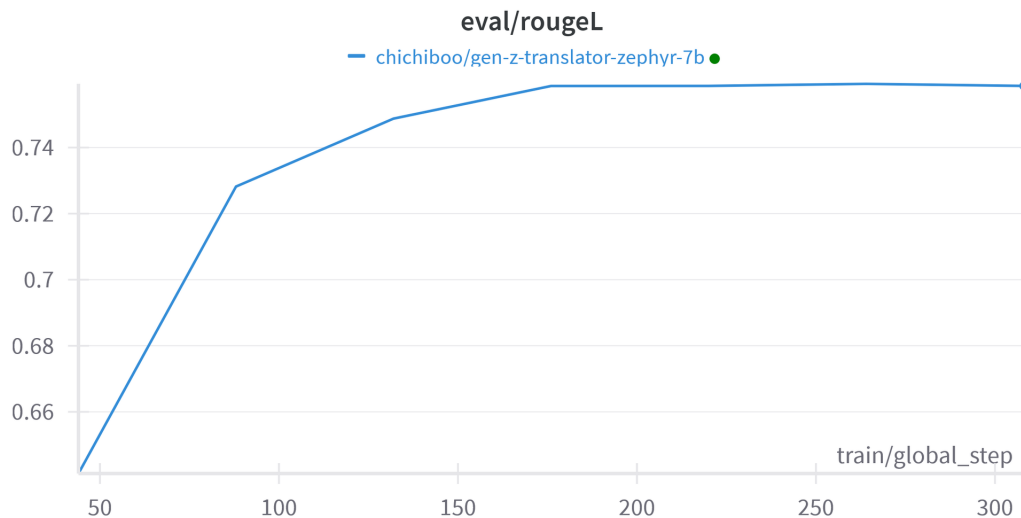


Figure 4.4: Evaluated using ROUGE-L metric

527 4.2.2 Text Generation

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

534 4.2.3 Automatic Evaluation Metrics

535 The dataset was automatically evaluated using BLEU and ROUGE metrics, specif-
 536 ically the ROUGE-L metric as the dataset do not contain newlines that ROUGE-
 537 Lsum uses to separate the input with. These scores were then averaged to deter-
 538 mine the score of the models. The base model obtained a BLEU score of 0.8099

539 and ROUGE-L Score of 0.8336 and the fine-tuned model obtained a BLEU score
540 of 0.8151 and ROUGE-L Score of 0.8396. While the difference between the mod-
541 els is minimal, this does not completely represent the performance of the models
542 as these metrics are only used to determine if the generated text is close to the
543 reference text, regardless of the context and the overall quality of the generated
544 text. However, it does show that the fine-tuned model has little improvement over
545 the base model.

546 4.2.4 Manual Evaluation Metrics

547 A manual evaluation was conducted by the researchers through a survey admin-
548 istered via Google Forms to determine which of the two models is preferred by
549 Generation Z students at University of the Philippines Visayas (UPV). The sur-
550 vey comprised a total of 144 questions, which were distributed across five sepa-
551 rate forms. The first form contained 20 questions, the second 19, the third 20,
552 the fourth 20, the fifth 14, **and the sixth 50 amounting to 143 questions**
553 in total. Each question presented two translation options: one generated by the
554 fine-tuned model and the other by the base model. Respondents were asked to
555 select the translation they preferred in each case. **A total of 114 individu-**
556 **als participated in the survey, with 29, 22, 22, 21, and 20 respondents**
557 **completing Forms 1 through 5, respectively.**

558 The data presented below illustrate respondent preferences between the base and
559 fine-tuned models across the six survey forms, as well as the overall summary of
560 the results. Each graph visualizes the outcomes for an individual form, specifically
561 indicating both the raw number of responses and the corresponding percentages

562 favoring each model. A systematic evaluation for each graph is provided as follows:

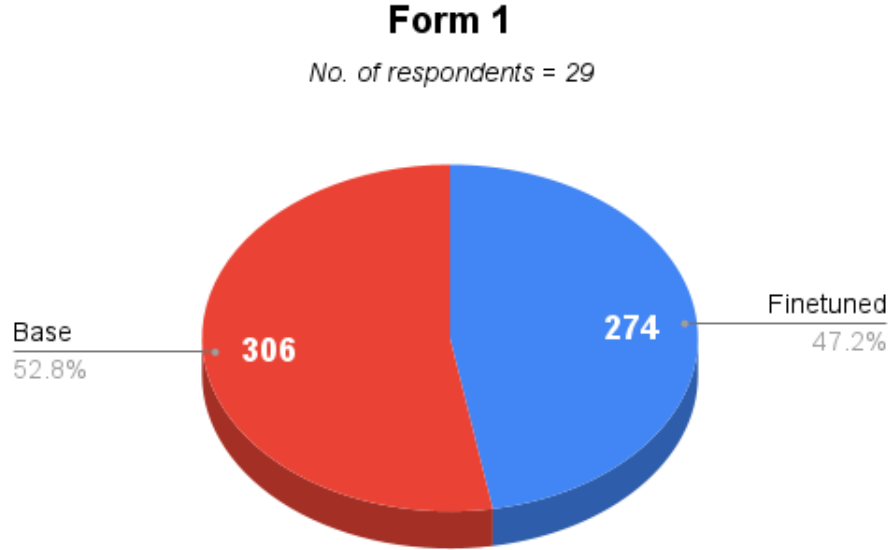


Figure 4.5: Form 1 Evaluation

563 Figure 4.5 shows that among the 29 respondents, 306 responses or 52.8 percent pre-
 564 ferred the base model, while 274 responses or 47.2 percent favored the fine-tuned
 565 model. This indicates a slight preference for the base model in this particular
 566 form. Notably, this result deviates from the overall trend observed in the other
 567 four forms, where the fine-tuned model tends to be favored. Form 1 is the only
 568 instance in which the base model outperformed the fine-tuned model, suggesting
 569 that specific characteristics of this form may have influenced the preferences of
 570 the respondents.

571 Figure 4.6 implies that among 22 respondents, 236 responses, or 56.5 percent,
 572 favored the fine-tuned model, while 182 responses, or 43.5 percent, preferred the
 573 base model. This 13 percent margin reflects the clear preference for the fine-tuned
 574 model, which is consistent with the overall trend observed across the other forms.

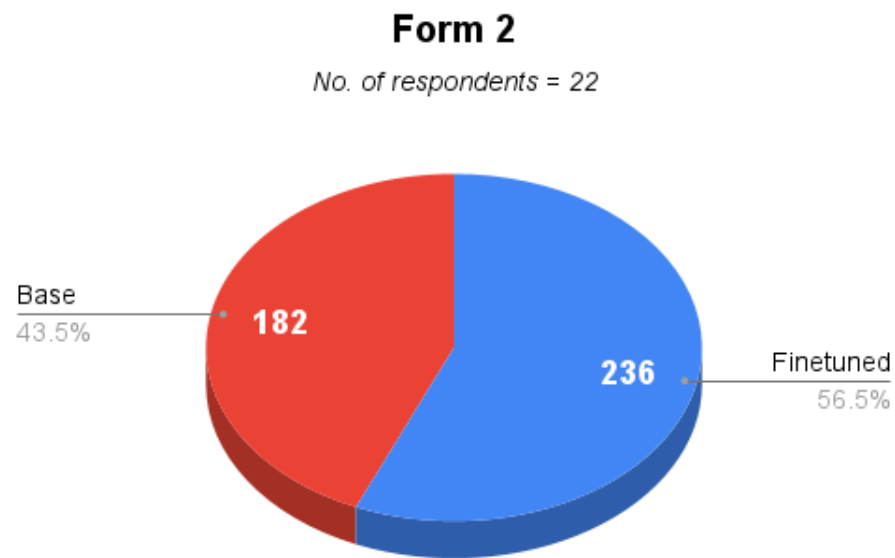


Figure 4.6: Form 2 Evaluation

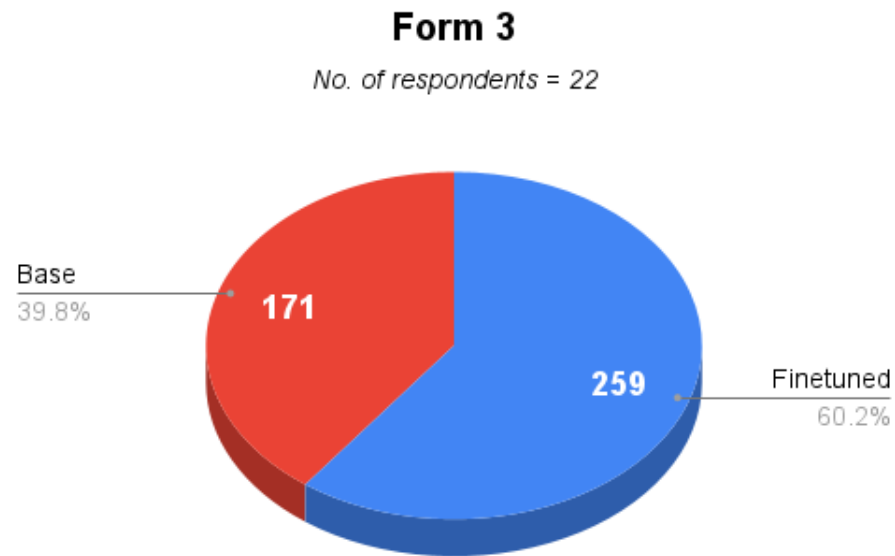


Figure 4.7: Form 3 Evaluation

575 Figure 4.7 illustrates that among the 22 respondents, the fine-tuned model received
576 a significantly higher preference, with 259 responses or 60.2 percent, compared to
577 the base model with 171 responses or 29.8 percent. This 20.4 percent margin
578 represents the widest gap among all forms. This strongly indicates the superior
579 performance of the fine-tuned model on translating, presented in Form 3.

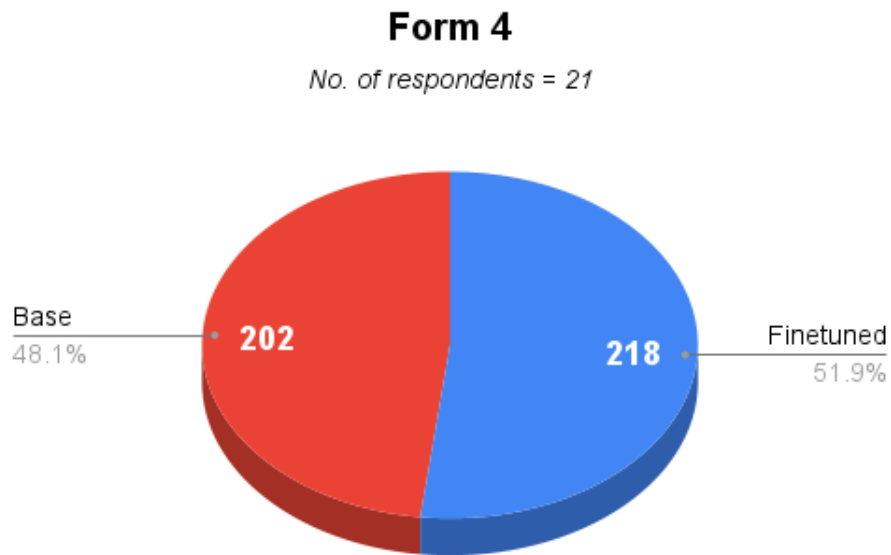


Figure 4.8: Form 4 Evaluation

580 Figure 4.8 highlights that the 21 respondents in Form 4 yielded a nearly even
581 distribution of preferences, with 218 responses or 51.9 percent favoring the fined-
582 tuned model and 202 responses or 48.1 percent preferring the base model. This
583 narrow 3.8 percent difference suggests a comparable level of performance between
584 the two models in this particular form.

585 Figure 4.9 conveys that among the 20 respondents in Form 5, 152 responses or
586 54.3 percent selected the fine-tuned model, while 128 responses or 45.7 percent
587 chose the base model. This 8.6 percent margin reinforces the general trend toward

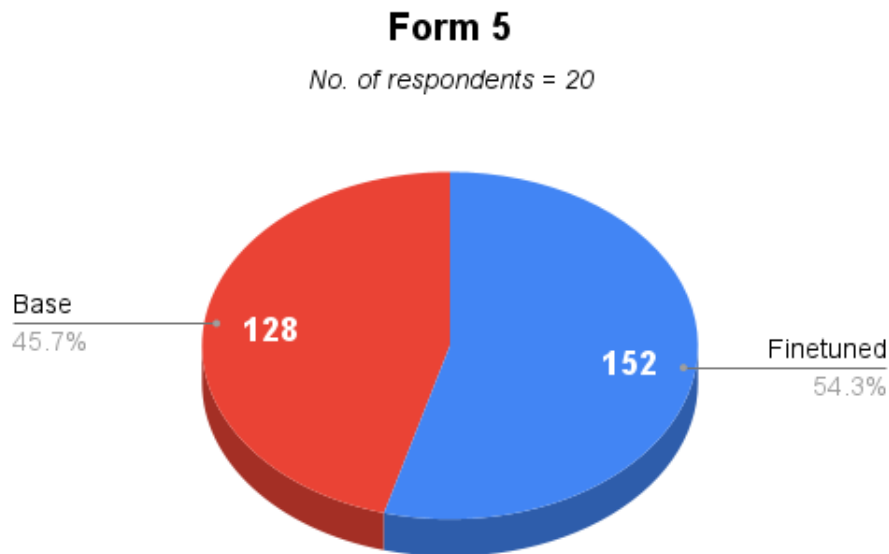


Figure 4.9: Form 5 Evaluation

the fine-tuned model across all forms.

Figure 4.10 indicates the results of the sixth form. 21 respondents 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 4.11 presents the overall summary across all five forms, with a total of 135 responsees garnered in the survey. In total, the fine-tuned model received 53.5%, while the base model garnered 989 preferences or 46.5%. 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.

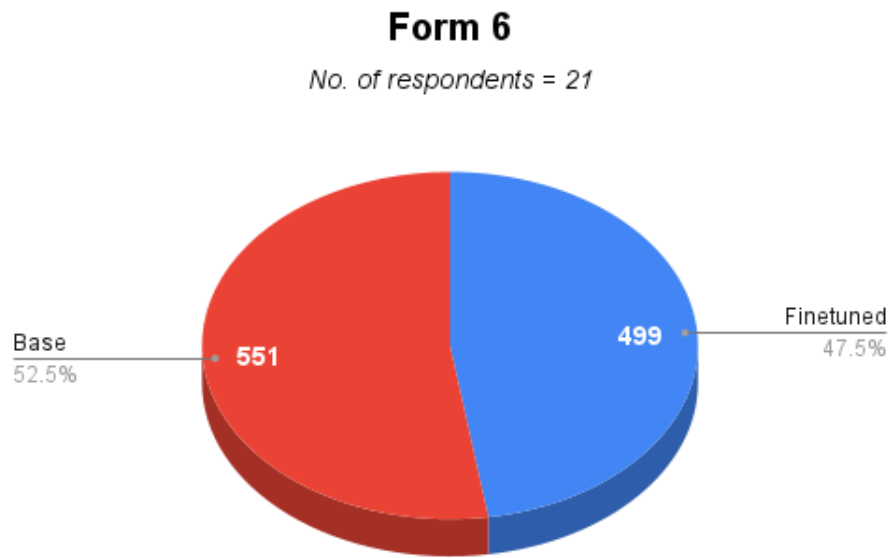


Figure 4.10: Form 6 Evaluation

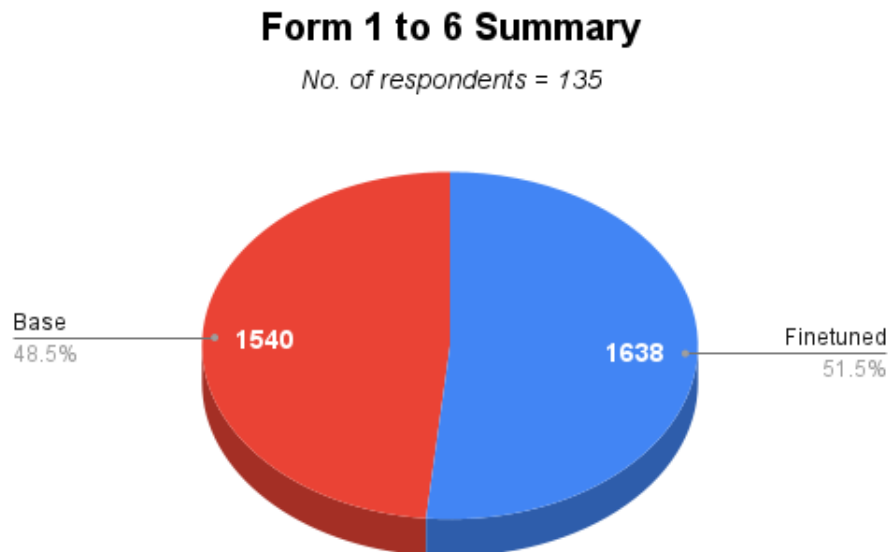


Figure 4.11: Summary Evaluation

599 4.3 Summary

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

606 Evaluation was conducted using both automatic and manual methods. The auto-
607 matic evaluation, using BLEU and ROUGE-L metrics, showed marginal improve-
608 ments in the fine-tuned model compared to the base model, suggesting slightly
609 better alignment with reference translations.

610 To support the results of automatic evaluation metrics, a manual evaluation was
611 carried out through online surveys among Generation Z students at UPV. Partic-
612 ipants compared translations from both models across six forms. Results showed
613 a moderate overall preference for the fine-tuned model, with 53.5% of responses
614 in its favor. While one form showed a slight preference for the base model, the
615 fine-tuned model was generally preferred, especially in Form 3 where it showed
616 the largest margin.

617 In summary, the findings indicate that the fine-tuned model slightly outperformed
618 the base model in terms of automatic metrics and showed a modest but consistent
619 preference among target users, supporting its effectiveness in translating Gen Z
620 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 53.5% 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 66 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.

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* (Tech. Rep.). National Bureau of Economic Research.
- 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). *Unslow*. Retrieved from <http://github.com/unslowthai/unslow>

- 666 Dua, A., Jacobson, R., Ellingrud, K., Enomoto, K., Cordina, J., Coe, E. H.,
667 & Finneman, B. (2024, Aug). *What is gen z?* McKinsey & Com-
668 pany. Retrieved from [https://www.mckinsey.com/featured-insights/](https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z)
669 [mckinsey-explainers/what-is-gen-z](https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z)
- 670 Esquivel, O. J. D. (2019). *58 exploring the filipinization of the english language in*
671 *a digital age: An identity apart from other world englishes*. Retrieved from
672 <https://files.eric.ed.gov/fulltext/EJ1244667.pdf>
- 673 Euchner, J. (2023). Generative ai. *Research-Technology Management*, 66(3),
674 71–74.
- 675 Fernández-Toro, M. (2016, Jun). *Exploring languages and cultures*. Re-
676 trieved from [https://www.open.edu/openlearn/languages/exploring](https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2)
677 [-languages-and-cultures/content-section-3.2](https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2)
- 678 Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). *Generative ai*
679 *and chatgpt: Applications, challenges, and ai-human collaboration* (Vol. 25)
680 (No. 3). Taylor & Francis.
- 681 Ghazali, N. M., & Abdullah, N. N. (2021, Dec). Slang language use
682 in social media among malaysian youths: A sociolinguistic per-
683 spective. *International Young Scholars Journal of Languages*,
684 4(2), 69. Retrieved from [https://www.iiium.edu.my/media/](https://www.iiium.edu.my/media/77652/Slang%20Language%20Use%20in%20Social%20Media%20Among%20Malaysian%20Youths_A%20Sociolinguistic%20Perspective.pdf)
685 [77652/Slang%20Language%20Use%20in%20Social%20Media%20Among%](https://www.iiium.edu.my/media/77652/Slang%20Language%20Use%20in%20Social%20Media%20Among%20Malaysian%20Youths_A%20Sociolinguistic%20Perspective.pdf)
686 [20Malaysian%20Youths_A%20Sociolinguistic%20Perspective.pdf](https://www.iiium.edu.my/media/77652/Slang%20Language%20Use%20in%20Social%20Media%20Among%20Malaysian%20Youths_A%20Sociolinguistic%20Perspective.pdf)
- 687 Gonzaga, M. (2025, Feb). *“forda convo ang ferson”: Analysis of*
688 *gen z slang in the lens of batstateu faculty members*. Retrieved
689 from [https://www.academia.edu/102575643/_FORDA_CONVO_ANG_FERSON](https://www.academia.edu/102575643/_FORDA_CONVO_ANG_FERSON_ANALYSIS_OF_GEN_Z_SLANG_IN_THE_LENS_OF_BATSTATEU_FACULTY_MEMBERS)
690 [_ANALYSIS_OF_GEN_Z_SLANG_IN_THE_LENS_OF_BATSTATEU_FACULTY_MEMBERS](https://www.academia.edu/102575643/_FORDA_CONVO_ANG_FERSON_ANALYSIS_OF_GEN_Z_SLANG_IN_THE_LENS_OF_BATSTATEU_FACULTY_MEMBERS)
- 691 Heydari, M., Albadvi, A., & Khazeni, M. (2024). Persian slang text conversion to

- 692 formal and deep learning of persian short texts on social media for sentiment
 693 classification. *Journal of Electrical and Computer Engineering Innovations*
 694 (*JECEI*). Retrieved from https://jecei.sru.ac.ir/article_2172.html
 695 doi: 10.22061/jecei.2024.10745.731
- 696 Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W.
 697 (2021). *Lora: Low-rank adaptation of large language models*. Retrieved
 698 from <https://arxiv.org/abs/2106.09685>
- 699 Ibrahim, A., & Sharief, B. (2023, 10). Intelligent system to transformer slang
 700 words into formal words. *NTU Journal of Engineering and Technology*, 2.
 701 doi: 10.56286/ntujet.v2i2.689
- 702 Jeresano, E., & Carretero, M. (2022, Feb). Digital culture and social media
 703 slang of gen z. *United International Journal for Research Technology*, 3(4),
 704 11–25. doi: <http://dx.doi.org/10.1314/RG.2.2.36361.93285>
- 705 Lin, C.-Y. (2004, Jul). Rouge: A package for automatic evaluation of summaries.
 706 *Meeting of the Association for Computational Linguistics*, 74–81.
- 707 Liu, J., Zhang, X., & Li, H. (2023, Aug). Analysis of language phenomena in
 708 internet slang: A case study of internet dirty language. *Open Access Library*
 709 *Journal*, 10(08), 1–12. doi: 10.4236/oalib.1110484
- 710 Liu, S., Gui, D.-Y., Zuo, Y., & Dai, Y. (2019, Jun). Good slang or bad slang?
 711 embedding internet slang in persuasive advertising. *Frontiers in Psychology*,
 712 10. doi: 10.3389/fpsyg.2019.01251
- 713 Mantiri, O. (2010, 03). Factors affecting language change.
 714 <http://ssrn.com/abstract=2566128>. doi: 10.2139/ssrn.2566128
- 715 Maulidiya, R., Wijaya, S. E., Mauren, C., Adha, T. P., & Pandin, M. G. R. (2021,
 716 Dec). *Language development of slang in the younger generation in the digital*
 717 *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.
- 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. P. (2024, Mar). Utilizing social media in communicating in english as a second language. *Studies in Self-Access Learning Journal*, 15(1). doi: 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>
- 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.

- 744 Vacalares, S. T., Salas, A. F. R., Babac, B. J. S., Cagalawan, A. L., & Calimpong,
 745 C. D. (2023, Jun). The intelligibility of internet slangs between millennials
 746 and gen zers: A comparative study. *International Journal of Science and*
 747 *Research Archive*, 9(1), 400–409. doi: 10.30574/ijrsra.2023.9.1.0456
- 748 Vergo, T., Godbout, J.-F., Rabbany, R., & Pelrine, K. (2024). *Comparing gpt-4*
 749 *and open-source language models in misinformation mitigation*. Retrieved
 750 from <https://arxiv.org/abs/2401.06920>
- 751 Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., ... Rishi, D.
 752 (2024). *Lora land: 310 fine-tuned llms that rival gpt-4, a technical report*.
 753 Retrieved from <https://arxiv.org/abs/2405.00732>