

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

3 A Special Problem Proposal  
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  
16 Adviser

17 April 29, 2025

## Abstract

19 Internet slang is an informal variation of language that is prominent to the younger  
20 generation. The usage of this language brought generational divide between them  
21 and the older generations. This study aims to develop a translation tool leveraging  
22 Large Language Models (LLMs) to bridge this issue. A dataset of Generation  
23 Alpha slang sentences and their formal equivalents will be used to fine-tune an  
24 existing LLM. The model will be trained to translate slang sentences into formal  
25 English, and will be compared against the baseline model using various evaluation  
26 metrics. The study highlights the significance of addressing communication gaps  
27 and provides insights into how technology can enhance understanding and reduce  
28 miscommunications across generations. This research contributes to the broader  
29 discourse on language adaptation and generational communication in the digital  
30 age.

31 **Keywords:** Internet Slang, Generation Alpha, Generational Divide,  
LoRA, LLM

32

# Contents

33	<b>1 Introduction</b>	<b>1</b>
34	1.1 Overview . . . . .	1
35	1.2 Problem Statement . . . . .	3
36	1.3 Research Objectives . . . . .	3
37	1.3.1 General Objectives . . . . .	3
38	1.3.2 Specific Objectives . . . . .	3
39	1.4 Scope and Limitations of the Research . . . . .	4
40	1.5 Significance of the Research . . . . .	4
41	<b>2 Review of Related Literature</b>	<b>5</b>
42	2.1 Communication Gap between Generations . . . . .	5
43	2.2 Generative AI . . . . .	6
44	2.3 Existing Studies . . . . .	6
45	2.4 LoRA for Fine Tuning . . . . .	7
46	2.5 Chapter Summary . . . . .	7
47	<b>3 Research Methodology</b>	<b>10</b>
48	3.1 Research Activities . . . . .	10

49	3.1.1	Data Gathering . . . . .	10
50	3.1.2	Data Preprocessing . . . . .	11
51	3.1.3	Model Fine-Tuning . . . . .	11
52	3.1.4	Model Evaluation . . . . .	12
53	3.2	Calendar of Activities . . . . .	12
54	<b>4</b>	<b>Results &amp; Discussions</b>	<b>14</b>
55	4.1	Dataset . . . . .	14
56	4.2	Model Evaluation . . . . .	14
57	4.2.1	Model Training . . . . .	14
58	4.2.2	Text Generation . . . . .	15
59	4.2.3	Automatic Evaluation Metrics . . . . .	15
60	4.2.4	Manual Evaluation Metrics . . . . .	15
61	<b>5</b>	<b>Conclusion</b>	<b>16</b>
62	5.1	Limitations . . . . .	16
63	5.2	Recommendations . . . . .	16
64		<b>References</b>	<b>17</b>
65	<b>A</b>	<b>Figures and Tables</b>	<b>20</b>

# 66 List of Figures

<small>67</small>	4.1 Automatic Evaluation of Dataset using BLEU and ROUGE metrics.	15
<small>68</small>	A.1 Code snippet of the prototype . . . . .	20

# 69 List of Tables

70	2.1	Summary of Existing Studies . . . . .	9
71	3.1	Timetable of Activities . . . . .	13
72	A.1	Difference between Generated Answers of the Base and LoRA Fine-	
73		Tuned Model . . . . .	21

# 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)). This 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

99 younger generation (Maulidiya, Wijaya, Mauren, Adha, & Pandin, 2021), where  
100 they use it to communicate and interact with friends.

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

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

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

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

127 The main contributions of this study are as follows:

- 128 • Enhance linguistic understanding between generations by using fine-tuning  
129 a LLM to translate Gen Z slang to formal language, leveraging the strengths  
130 of advanced NLP techniques
- 131 • Bridge communication gaps between generations using the proposed model  
132 to foster better relationships
- 133 • Create a scalable framework that can be adapted to translate slang in other  
134 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.

## 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

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

## 159    **1.4    Scope and Limitations of the Research**

160    This study focused on the use of internet slang by Filipino Generation Z, with  
161    an emphasis on the English language, as it is widely used on different digital  
162    platforms, such as social networks.

## 163    **1.5    Significance of the Research**

164    The study contributed to understanding the evolving linguistic landscape shaped  
165    by Internet slang, especially as used by Generation Z. The insights gained from  
166    this study aid educators, parents, and communication professionals in bridging  
167    inter-generational communication gaps and fostering better understanding across  
168    age groups.

## 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 (Suslak, 2009) argues that age influences language use, noting that language evolves across generations. Supporting this, a study by Teng and Joo (Teng & 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).

## 192 2.2 Generative AI

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

## 201 2.3 Existing Studies

202 Verghe et al. (Verghe, Godbout, Rabbany, & Pelrine, 2024) used multiple open  
203 source LLMs and compared them with the latest version of GPT-3.5 and 4.0 models  
204 at that time. They determined zephyr-7b-beta is a viable open-source alternative  
205 to these models and is comparable with the latest GPT-4.0 model.

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

218 The study by Nocon et al. (Nocon, Kho, & Arroyo, 2018) explores translating  
219 Filipino colloquialisms, such as Conyo and Datkilab, into standardized Filipino,  
220 addressing comprehension barriers for non-familiar speakers. Two machine  
221 translation (MT) approaches were evaluated: Tensorflow’s Sequence-to-Sequence  
222 model using Recurrent Neural Networks (RNNs) and Moses’ Phrase-based Statistical  
223 MT. Moses outperformed Tensorflow on test data due to its handling of  
224 phrase combinations and unfamiliar words, while Tensorflow excelled on training  
225 data, indicating potential with refinement and more training data. The research  
226 underscores the need for robust datasets and highlights the strengths of phrase-

227 based statistical MT in tackling slang translation challenges.

228 Ibrahim and Mustafa (Ibrahim & Sharief, 2023) developed a system to trans-  
229 late slang into formal language, addressing challenges posed by slang’s informality  
230 and variability. Using updated datasets of slang words, formal equivalents, and  
231 contextual sentences, they fine-tuned pre-trained models from Hugging Face’s  
232 Transformer library. While the T5-base model showed promise during training,  
233 it performed poorly in testing. In contrast, the “facebook/bart-base” model ex-  
234 celled, demonstrating high accuracy and low loss values. The study highlights the  
235 importance of fine-tuning and updating datasets for effective slang translation  
236 and emphasizes the potential of transformer models like “facebook/bart-base” in  
237 bridging informal and formal language gaps.

## 238 2.4 LoRA for Fine Tuning

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

## 252 2.5 Chapter Summary

253 This chapter shows how generational differences create communication gaps, espe-  
254 cially due to internet slang. Younger people tend to use slang to express emotions  
255 and connect with friends, but this can confuse older generations who aren’t as  
256 familiar with these terms. Research shows that as language changes over time,  
257 older people are generally less likely to understand the newest internet language.  
258 To bridge this gap, some recent studies have utilized machine learning to translate  
259 slang into more standard language. For instance, Khazeni et al. (Heydari et al.,

260 2024) used deep learning to translate Persian slang, while Nocon et al. (Nocon et  
261 al., 2018) created a Filipino slang translator using statistical models. Moreover,  
262 Ibrahim and Mustafa (Ibrahim & Sharief, 2023) fine-tuned pre-trained models to  
263 learn slang meanings. One promising technique for this is Low Rank Adaptation  
264 (LoRA), which is a fine-tuning method that keeps the original model stable while  
265 using less storage. Studies by Zhao et al. (Zhao et al., 2024) and Nguyen et al.  
266 (Nguyen et al., 2023) show that LoRA models are not only efficient but can even  
267 outperform advanced models like GPT-4 when it comes to slang translation and  
268 text classification.

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

#### 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.



### 288 3.1.2 Data Preprocessing

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

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

### 303 3.1.3 Model Fine-Tuning

304 The model used in this study was zephyr-7b-beta because it is open-source and  
305 was proven to perform better than other models of the same size. In addition,  
306 it can be trained in a GPU with 16GB of VRAM, necessary as we are using the  
307 free tier of Google Colab as the platform of choice for prototype fine-tuning of the  
308 model.

309 This study used the example codes provided by HuggingFace in the documen-  
310 tation of their various libraries and sample notebook provided in the zephyr-7b-  
311 beta repository.

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

318 To evaluate the model training process and ensure that the model is not overfit-  
319 ting, Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy  
320 for Gisting Evaluation (ROUGE) are used. BLEU is used to measure the preci-  
321 sion of the model by determining how much of the generated text appear in the

reference text (Papineni, Roukos, Ward, & Zhu, 2001) while ROUGE is used to measure recall as it determines how much of the reference text is in the generated text (Lin, 2004). 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 from the TRL library of HuggingFace to simplify the training process. The model was trained with the following parameters: 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 and the default parameters of SFTTrainer.

### 3.1.4 Model Evaluation

The model was evaluated using both automatic and manual evaluation metrics. The model was then prompted to generate a formal sentence for each sentence 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. Identical answers between the finetuned and the base model were removed to in the test set to ensure that the model is evaluated properly. A total of 144 sentences were used to evaluate the model.

A survey was conducted to compare the finetuned model to the base model to determine if the finetuning was effective. The survey was conducted online using Google Forms asked the participants to pick which of the following sentences is the more accurate translation of the given sentence based on accuracy, naturalness, and context. The order in which sentences from the two models were shown was randomly selected to avoid bias.

## 3.2 Calendar of Activities

Table 3.1 shows a Gantt chart of the activities. Each bullet represents approximately one week's worth of activity.

Table 3.1: Timetable of Activities

Activities (2024-2025)	Dec	Jan	Feb	Mar	Apr	May	Jun
Creation of the dataset	•						
Identification of potential LLM to be used	•						
Lookup on available GPU on demand services	•						
Study on LoRA implementation for LLM	•						
Preprocessing of data	•••						
Prototype implementation of LoRA	•	••••					
Implementation of LoRA on selected model			••				
Implementation of LLM Evaluation Metrics			••				
Model Evaluation and Analysis of Results				••••			
Documentation	••	••••	••••	••••	••••		

## Chapter 4

# Results & 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 698 Gen Z internet slang and their corresponding formal translation.

### 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.

Here, we can see that the while the training loss is decreasing, the validation loss is increasing and other metrics are not improving. This indicates 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: Automatic Evaluation of Dataset using BLEU and ROUGE metrics.

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	RougeL	RougeLsum	Bleu
1	2.303800	1.268362	0.672177	0.467319	0.641333	0.669745	0.497899
2	0.793800	0.637890	0.760193	0.593917	0.728189	0.757228	0.595286
3	0.606600	0.588256	0.775976	0.610833	0.748757	0.773886	0.615690
4	0.543100	0.571621	0.784202	0.621421	0.758633	0.782125	0.620381
5	0.474800	0.565389	0.783062	0.623826	0.758653	0.780710	0.620449
6	0.408300	0.582480	0.782192	0.623898	0.759282	0.779676	0.612101
7	0.331900	0.613242	0.778451	0.619570	0.758207	0.775850	0.614556

### 4.2.2 Text Generation

A total of 197 sentences were translated using both the base zephyr-7b-beta model and the finetuned model. These served as the dataset used to evaluate the performance of the model and comparing it with the other base model.

### 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 0.8112 and ROUGE-L Score of 0.8390 and the finetuned model obtained a BLEU score of 0.8125 and ROUGE-L Score of 0.8412. While the difference between the models are very 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 finetuned model, while not significantly better than the base model, is close to the reference model.

### 4.2.4 Manual Evaluation Metrics

## 390 Chapter 5

## 391 Conclusion

### 392 5.1 Limitations

393 Language is dynamic and constantly evolving, making it difficult to establish clear  
394 boundaries on when slang terms form or end within a generation.

### 395 5.2 Recommendations

396 Future researchers are encouraged to expand the vocabulary of slang terms on the  
397 Internet and explore more recent trends, taking into account the dynamic nature  
398 of language.

## References

- 400 Ambarsari, S., Amrullah, A., & Nawawi, N. (2020, Aug). The use of online  
401 slang for independent learning in english vocabulary. *Proceedings of the 1st*  
402 *Annual Conference on Education and Social Sciences (ACCESS 2019)*, 465,  
403 295–297. doi: 10.2991/assehr.k.200827.074
- 404 Barseghyan, L. (2014). *On some aspects of internet slang*. Retrieved from  
405 <https://api.semanticscholar.org/CorpusID:51730779>
- 406 binti Sabri, N. A., bin Hamdan, S., Nadarajan, N.-T. M., & Shing, S. R. (2020,  
407 Jun). The usage of english internet slang among malaysians in social media.  
408 *Selangor Humaniora Review*, 4(1), 16–17.
- 409 Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). *Generative ai at work* (Tech.  
410 Rep.). National Bureau of Economic Research.
- 411 Crystal, D., & Robins, R. H. (2024, Oct). *Language*. Encyclopædia Britannica,  
412 inc. Retrieved from <https://www.britannica.com/topic/language>
- 413 Daniel Han, M. H., & team, U. (2023). *Unsloth*. Retrieved from [http://github](http://github.com/unslothai/unsloth)  
414 [.com/unslothai/unsloth](http://github.com/unslothai/unsloth)
- 415 Dua, A., Jacobson, R., Ellingrud, K., Enomoto, K., Cordina, J., Coe, E. H.,  
416 & Finneman, B. (2024, Aug). *What is gen z?* McKinsey & Com-  
417 pany. Retrieved from [https://www.mckinsey.com/featured-insights/](https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z)  
418 [mckinsey-explainers/what-is-gen-z](https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z)
- 419 Euchner, J. (2023). Generative ai. *Research-Technology Management*, 66(3),  
420 71–74.
- 421 Fernández-Toro, M. (2016, Jun). *Exploring languages and cultures*. Re-  
422 trieved from [https://www.open.edu/openlearn/languages/exploring](https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2)  
423 [-languages-and-cultures/content-section-3.2](https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2)
- 424 Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). *Generative ai*  
425 *and chatgpt: Applications, challenges, and ai-human collaboration* (Vol. 25)  
426 (No. 3). Taylor & Francis.
- 427 Ghazali, N. M., & Abdullah, N. N. (2021, Dec). Slang language use  
428 in social media among malaysian youths: A sociolinguistic per-  
429 spective. *International Young Scholars Journal of Languages*,  
430 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](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](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 transform 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>

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](https://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



- 473 Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). Bleu: a method for  
 474 automatic evaluation of machine translation. *Proceedings of the 40th Annual*  
 475 *Meeting on Association for Computational Linguistics - ACL '02*. Retrieved  
 476 from <https://dl.acm.org/citation.cfm?id=1073135> doi: [https://doi](https://doi.org/10.3115/1073083.1073135)  
 477 [.org/10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135)
- 478 Suslak, D. F. (2009). The sociolinguistic problem of generations. *Lan-*  
 479 *guage & Communication*, 29(3), 199–209. Retrieved from [https://www](https://www.sciencedirect.com/science/article/pii/S0271530909000196)  
 480 [.sciencedirect.com/science/article/pii/S0271530909000196](https://www.sciencedirect.com/science/article/pii/S0271530909000196) (Re-  
 481 flecting on language and culture fieldwork in the early 21st century) doi:  
 482 <https://doi.org/10.1016/j.langcom.2009.02.003>
- 483 Teng, C. E., & Joo, T. M. (2023). Is internet language a destroyer to communica-  
 484 tion? In X.-S. Yang, R. S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of*  
 485 *eighth international congress on information and communication technology*  
 486 (pp. 527–536). Singapore: Springer Nature Singapore.
- 487 Vacalares, S. T., Salas, A. F. R., Babac, B. J. S., Cagalawan, A. L., & Calimpong,  
 488 C. D. (2023, Jun). The intelligibility of internet slangs between millennials  
 489 and gen zers: A comparative study. *International Journal of Science and*  
 490 *Research Archive*, 9(1), 400–409. doi: 10.30574/ijrsra.2023.9.1.0456
- 491 Verghe, T., Godbout, J.-F., Rabbany, R., & Pelrine, K. (2024). *Comparing gpt-4*  
 492 *and open-source language models in misinformation mitigation*. Retrieved  
 493 from <https://arxiv.org/abs/2401.06920>
- 494 Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., ... Rishi, D.  
 495 (2024). *Lora land: 310 fine-tuned llms that rival gpt-4, a technical report*.  
 496 Retrieved from <https://arxiv.org/abs/2405.00732>

## <sup>497</sup> **Appendix A**

## <sup>498</sup> **Figures and Tables**

Figure A.1: Code snippet of the prototype

```
from transformers import TrainingArguments

YOUR_HF_USERNAME = "chichiboo"

output_dir = f"{YOUR_HF_USERNAME}/zephyr-7b-qlora-ultrachat"
per_device_train_batch_size = 4
gradient_accumulation_steps = 4
optim = "paged_adamw_32bit"
save_steps = 10
logging_steps = 10
learning_rate = 2e-4
max_grad_norm = 0.3
max_steps = 100
warmup_ratio = 0.03
lr_scheduler_type = "constant"

training_arguments = TrainingArguments(
    output_dir=output_dir,
    per_device_train_batch_size=per_device_train_batch_size,
    gradient_accumulation_steps=gradient_accumulation_steps,
    optim=optim,
    save_steps=save_steps,
    logging_steps=logging_steps,
    learning_rate=learning_rate,
    max_grad_norm=max_grad_norm,
    max_steps=max_steps,
    warmup_ratio=warmup_ratio,
    lr_scheduler_type=lr_scheduler_type,
```

Table A.1: Difference between Generated Answers of the Base and LoRA Fine-Tuned Model

---

**Before LoRA:**

USER: Can you explain contrastive learning in machine learning in simple terms for someone new to the field of ML?

ASSISTANT: Contrastive learning is a machine learning technique that involves training a model to learn from data that is similar to the target data. The model is trained to identify patterns in the data that are similar to the target data, and to use those patterns to make predictions about new data.

---

**After LoRA:**

USER: Can you explain contrastive learning in machine learning in simple terms for someone new to the field of ML?

ASSISTANT: Contrastive learning is a technique in machine learning that uses a technique called contrastive loss to train a model. The idea is to train a model to distinguish between similar and dissimilar data. In contrastive learning, a model is trained to distinguish between a pair of similar data, such as a pair of images, and a pair of dissimilar data, such as a pair of different images. The model is trained to minimize the loss, which is the difference between the similarity and dissimilarity of the data. In contrast, traditional machine learning techniques, such as supervised learning, are trained to distinguish between different types of data, such as a classification or regression task. In summary, contrastive learning is a technique in machine learning that uses a technique called contrastive loss to train a model to distinguish between similar and dissimilar data, which is different from traditional machine learning techniques, such as supervised learning.

---