

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	1 Introduction	1
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	2 Review of Related Literature	5
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	3 Research Methodology	10
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	4	Results & Discussions	14
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	5	Conclusion	16
62	5.1	Limitations	16
63	5.2	Recommendations	16
64		References	17
65	A	Figures and Tables	20

66 List of Figures

<small>67</small>	A.1 Code snippet of the prototype	20
-------------------	---	----

68 List of Tables

69	2.1	Summary of Existing Studies	9
70	3.1	Timetable of Activities	13
71	A.1	Difference between Generated Answers of the Base and LoRA Fine-	
72		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

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

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

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

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

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

126 The main contributions of this study are as follows:

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

158 **1.4 Scope and Limitations of the Research**

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

162 **1.5 Significance of the Research**

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

190 2.2 Generative AI

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

199 2.3 Existing Studies

200 Vergho et al.(?, ?) used multiple open source LLMs and compared them with the
201 latest ersion of GPT-3.5 and 4.0 models at that time. They determined zephyr-
202 7b-beta is a viable open-source alternative to these models and is comparable with
203 the latest GPT-4.0 model.

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

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

225 based statistical MT in tackling slang translation challenges.

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

236 2.4 LoRA for Fine Tuning

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

250 2.5 Chapter Summary

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

258 2024) used deep learning to translate Persian slang, while Nocon et al. (Nocon et
259 al., 2018) created a Filipino slang translator using statistical models. Moreover,
260 Ibrahim and Mustafa (Ibrahim & Sharief, 2023) fine-tuned pre-trained models to
261 learn slang meanings. One promising technique for this is Low Rank Adaptation
262 (LoRA), which is a fine-tuning method that keeps the original model stable while
263 using less storage. Studies by Zhao et al. (Zhao et al., 2024) and Nguyen et al.
264 (Nguyen et al., 2023) show that LoRA models are not only efficient but can even
265 outperform advanced models like GPT-4 when it comes to slang translation and
266 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.

286 3.1.2 Data Preprocessing

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

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

301 3.1.3 Model Fine-Tuning

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

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

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

315 To evaluate the model training process and ensure that the model is not overfit-
316 ting, Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy
317 for Gisting Evaluation (ROUGE) are used. BLEU is used to measure the pre-
318 cision of the model by determining how much of the generated text appear in
319 the reference text while ROUGE is used to measure recall as it determines how

320 much of the reference text is in the generated text. These metrics use n-grams,
321 making them superior to standard recall and precision metrics as they take into
322 account the positioning of the words. These two metrics were implemented using
323 the Evaluate library by HuggingFace, making it easier to integrate with the rest
324 of the model training process. These metrics was calculated at every epoch of the
325 training process and is used for an early stopping callback to immediately stop
326 the model training if the model seems to be overfitting.

327 The model was then trained using SFTTrainer library to simplify the training
328 process. The model was trained with the following parameters: batch size of 8,
329 learning rate of 1e-5, and maximum number of epochs of 50. These parameters
330 were chosen based on the GPU provided in Colab and the default parameters of
331 SFTTrainer.

332 3.1.4 Model Evaluation

333 The model was evaluated using both automatic and manual evaluation metrics.
334 The model was then prompted to generate a formal sentence for each sentence in
335 the test dataset. The generated sentences were then compared to the formal trans-
336 lation of the sentence using BLEU and ROUGE metrics. The base zephyr-7b-beta
337 model was also prompted to generate sentences for the BLEU and ROUGE metric
338 and the pairwise comparison for human evaluation. Identical answers between the
339 finetuned and the base model were removed to in the test set to ensure that the
340 model is evaluated properly. A total of 144 sentences were used to evaluate the
341 model.

342 A survey was conducted to compare the finetuned model to the base model
343 to determine if the finetuning was effective. A total of participants were involved
344 in the survey, and the survey was conducted online using Google Forms. The
345 survey asked the participants to pick which of the following sentences is the more
346 accurate translation of the given sentence. The order in which sentences from the
347 two models were shown was randomly selected to avoid bias.

348 3.2 Calendar of Activities

349 Table 3.1 shows a Gantt chart of the activities. Each bullet represents approxi-
350 mately 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.

369 4.2.2 Text Generation

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

373 4.2.3 Automatic Evaluation Metrics

374 The dataset was automatically evaluated using BLEU and ROUGE metrics, specif-
375 ically the ROUGE-L metric as the dataset do not contain newlines that ROUGE-
376 Lsum uses to separate the input with. These scores were then averaged to deter-
377 mine the score of the models. The base model obtained a BLEU score of 0.8112
378 and ROUGE-L Score of 0.8390 and the finetuned model obtained a BLEU score
379 of 0.8125 and ROUGE-L Score of 0.8412. While the difference between the mod-
380 els are very minimal, this does not completely represent the performance of the
381 models as these metrics are only used to determine if the generated text is close to
382 the reference text, regardless of the context and the overall quality of the gener-
383 ated text. However, it does show that the finetuned model, while not significantly
384 better than the base model, is close to the reference model.

385 4.2.4 Manual Evaluation Metrics

386 Chapter 5

387 Conclusion

388 5.1 Limitations

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

391 5.2 Recommendations

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

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>
- Dua, A., Jacobson, R., Ellingrud, K., Enomoto, K., Cordina, J., Coe, E. H., & Finneman, B. (2024, Aug). *What is gen z?* McKinsey Company. Retrieved from <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z>
- Euchner, J. (2023). Generative ai. *Research-Technology Management*, 66(3), 71–74.
- Fernández-Toro, M. (2016, Jun). *Exploring languages and cultures*. Retrieved from <https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2>
- Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). *Generative ai and chatgpt: Applications, challenges, and ai-human collaboration* (Vol. 25) (No. 3). Taylor & Francis.
- Ghazali, N. M., & Abdullah, N. N. (2021, Dec). Slang language use in social media among malaysian youths: A sociolinguistic perspective. *International Young Scholars Journal of Languages*, 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

- Heydari, M., Albadvi, A., & Khazeni, M. (2024). Persian slang text conversion to formal and deep learning of persian short texts on social media for sentiment classification. *Journal of Electrical and Computer Engineering Innovations (JECEI)*. Retrieved from https://jecei.sru.ac.ir/article_2172.html doi: 10.22061/jecei.2024.10745.731
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). *Lora: Low-rank adaptation of large language models*. Retrieved from <https://arxiv.org/abs/2106.09685>
- Ibrahim, A., & Sharief, B. (2023, 10). Intelligent system to 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>
- Liu, J., Zhang, X., & Li, H. (2023, Aug). Analysis of language phenomena in internet slang: A case study of internet dirty language. *Open Access Library Journal*, 10(08), 1–12. doi: 10.4236/oalib.1110484
- Liu, S., Gui, D.-Y., Zuo, Y., & Dai, Y. (2019, Jun). Good slang or bad slang? embedding internet slang in persuasive advertising. *Frontiers in Psychology*, 10. doi: 10.3389/fpsyg.2019.01251
- Mantiri, O. (2010, 03). Factors affecting language change. <http://ssrn.com/abstract=2566128>. doi: 10.2139/ssrn.2566128
- Maulidiya, R., Wijaya, S. E., Mauren, C., Adha, T. P., & Pandin, M. G. R. (2021, Dec). *Language development of slang in the younger generation in the digital era*. OSF Preprints. Retrieved from osf.io/xs7kd doi: 10.31219/osf.io/xs7kd
- McArthur, T. (2003). *Concise oxford companion to the english language* (1st ed.). Oxford University Press.
- 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
- 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*

469 (pp. 527–536). Singapore: Springer Nature Singapore.
470 Vacalares, S. T., Salas, A. F. R., Babac, B. J. S., Cagalawan, A. L., & Calimpong,
471 C. D. (2023, Jun). The intelligibility of internet slangs between millennials
472 and gen zers: A comparative study. *International Journal of Science and*
473 *Research Archive*, 9(1), 400–409. doi: 10.30574/ijrsra.2023.9.1.0456
474 Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., ... Rishi, D.
475 (2024). *Lora land: 310 fine-tuned llms that rival gpt-4, a technical report*.
476 Retrieved from <https://arxiv.org/abs/2405.00732>

477 Appendix A

478 Figures and Tables

Figure A.1: Code snippet of the prototype

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.
