

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 21, 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, whose
37 unwavering support throughout our academic journey and our continual source of
38 inspiration and strength, especially 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. The usage of this language brought a generational divide between them and the older generations. This study aimed to develop a translation tool leveraging Large Language Models (LLMs) to bridge this issue. 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. Results showed that the fine-tuned model only slightly outperformed the base model in terms of automatic metrics, and it was generally preferred by human evaluators. These results indicate the fine-tuned model's effectiveness in producing more contextually appropriate and user-aligned formal translations.

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

62

Contents

63	1 Introduction	1
64	1.1 Overview	1
65	1.2 Problem Statement	4
66	1.3 Research Objectives	4
67	1.3.1 General Objectives	4
68	1.3.2 Specific Objectives	5
69	1.4 Scope and Limitations of the Research	5
70	1.5 Significance of the Research	5
71	2 Review of Related Literature	7
72	2.1 Communication Gap between Generations	7
73	2.2 Generative AI	8

74	2.3 Existing Studies	8
75	2.4 LoRA for Fine Tuning	10
76	2.5 Chapter Summary	11
77	3 Research Methodology	13
78	3.1 Research Activities	13
79	3.1.1 Data Gathering	14
80	3.1.2 Data Preprocessing	14
81	3.1.3 Model Fine-Tuning	15
82	3.1.4 Model Evaluation	16
83	4 Results and Discussions	19
84	4.1 Dataset	19
85	4.2 Model Evaluation	19
86	4.2.1 Model Training	19
87	4.2.2 Text Generation	22
88	4.2.3 Automatic Evaluation Metrics	22
89	4.2.4 Manual Evaluation Metrics	23
90	4.3 Summary	29

91	5 Conclusion	31
92	5.1 Limitations	32
93	5.2 Recommendations	32
94	6 References	33

95 List of Figures

96	3.1 Summarized Methodology	13
97	4.1 Training loss curve of the fine-tuned model across training steps .	20
98	4.2 Evaluation loss curve of the fine-tuned model across training steps	21
99	4.3 Evaluated using BLEU metric	21
100	4.4 Evaluated using ROUGE-L metric	22
101	4.5 Form 1 Evaluation	24
102	4.6 Form 2 Evaluation	25
103	4.7 Form 3 Evaluation	26
104	4.8 Form 4 Evaluation	27
105	4.9 Form 5 Evaluation	28
106	4.10 Summary Evaluation	29

¹⁰⁷ **List of Tables**

¹⁰⁸	2.1 Summary of Existing Studies	12
----------------	---	----

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-

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

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

136 Generation Z, individuals born between 1996 and 2009, are regarded as “digital
137 natives” because technology is an integral part of their upbringing (Dua et al.,
138 2024). Even the language of this generation is greatly affected by technology,
139 where newly coined terms and phrases, called Gen Z slang, are tied to the me-
140 dia culture they’ve grown up with (Jeresano & Carretero, 2022). However, this
141 evolution of language often creates communication barriers with older generations
142 (Venter, 2017 as cited in (Ghazali & Abdullah, 2021)). Furthermore, studies show
143 that even within Generation Z, people with limited exposure to social media may
144 struggle to understand the prevalent slang (Vacalares, Salas, Babac, Cagalawan,
145 & Calimpong, 2023).

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

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

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

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

162 The main contributions of this study are as follows:

- 163 • Enhance linguistic understanding between generations by using fine-tuning
164 a LLM to translate Gen Z slang to formal language, leveraging the strengths
165 of advanced NLP techniques
- 166 • Bridge communication gaps between generations using the proposed model
167 to foster better relationships
- 168 • Create a scalable framework that can be adapted to translate slang in other
169 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.

190 1.3.2 Specific Objectives

191 Specifically, the study aims to:

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

199 1.4 Scope and Limitations of the Research

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

203 1.5 Significance of the Research

204 The study contributed to understanding the evolving linguistic landscape shaped
205 by Internet slang, especially as used by Generation Z. The insights gained from
206 this study aid educators, parents, and communication professionals in bridging

207 inter-generational communication gaps and fostering better understanding across
208 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 lin-

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

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

230 **2.2 Generative AI**

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

239 **2.3 Existing Studies**

240 Vergho et al. (Vergho, Godbout, Rabbany, & Pelrine, 2024) used multiple open
241 source LLMs and compared them with the latest ersion of GPT-3.5 and 4.0 models
242 at that time. They determined zephyr-7b-beta is a viable open-source alternative

243 to these models and is comparable with the latest GPT-4.0 model.

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

256 The study by Nocon et al. (Nocon, Kho, & Arroyo, 2018) explores translating
257 Filipino colloquialisms, such as Conyo and Datkilab, into standardized Filipino,
258 addressing comprehension barriers for non-familiar speakers. Two machine trans-
259 lation (MT) approaches were evaluated: Tensorflow’s Sequence-to-Sequence model
260 using Recurrent Neural Networks (RNNs) and Moses’ Phrase-based Statistical
261 MT. Moses outperformed Tensorflow on test data due to its handling of phrase
262 combinations and unfamiliar words, while Tensorflow excelled on training data,
263 indicating potential with refinement and more training data. The research under-
264 scores the need for robust datasets and highlights the strengths of phrase-based
265 statistical MT in tackling slang translation challenges.

266 Ibrahim and Mustafa (Ibrahim & Sharief, 2023) developed a system to translate

267 slang into formal language, addressing challenges posed by slang’s informality
268 and variability. Using updated datasets of slang words, formal equivalents, and
269 contextual sentences, they fine-tuned pre-trained models from Hugging Face’s
270 Transformer library. While the T5-base model showed promise during training,
271 it performed poorly in testing. In contrast, the “facebook/bart-base” model ex-
272 celled, demonstrating high accuracy and low loss values. The study highlights the
273 importance of fine-tuning and updating datasets for effective slang translation
274 and emphasizes the potential of transformer models like “facebook/bart-base” in
275 bridging informal and formal language gaps.

276 **2.4 LoRA for Fine Tuning**

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

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

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

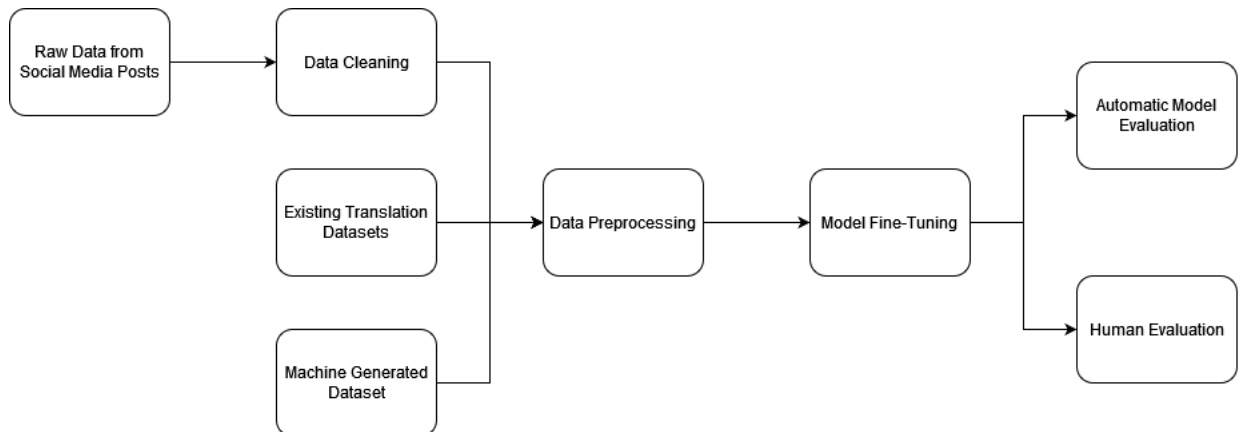


Figure 3.1: Summarized Methodology

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

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

341 3.1.3 Model Fine-Tuning

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

347 This study used the example codes provided by HuggingFace in the documentation
348 of their various libraries and sample notebook provided in the zephyr-7b-beta
349 repository.

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

356 To evaluate the model training process and ensure that the model is not overfitting,

357 Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for
358 Gisting Evaluation (ROUGE) are used. BLEU is used to measure the precision of
359 the model by determining how much of the generated text appear in the reference
360 text (Papineni, Roukos, Ward, & Zhu, 2001) while ROUGE is used to measure
361 recall as it determines how much of the reference text is in the generated text (Lin,
362 2004). These metrics use n-grams, making them superior to standard recall and
363 precision metrics as they take into account the positioning of the words. These
364 two metrics were implemented using the Evaluate library by HuggingFace, making
365 it easier to integrate with the rest of the model training process. These metrics
366 was calculated at every epoch of the training process and is used for an early
367 stopping callback to immediately stop the model training if the model seems to
368 be overfitting.

369 The model was then trained using SFTTrainer from the TRL library of Hugging-
370 Face to simplify the training process. The model was trained with the following
371 parameters: optimizer is paged 4bit AdamW, batch size of 8, learning rate of 2e-5,
372 and maximum number of epochs of 50. These parameters were chosen based on
373 the GPU provided in Colab, the test notebook by HuggingFace and the default
374 parameters of SFTTrainer.

375 3.1.4 Model Evaluation

376 The model was evaluated using both automatic and manual evaluation metrics.
377 The model was then prompted to generate a formal sentence for each sentence in
378 the test dataset. The generated sentences were then compared to the formal trans-
379 lation of the sentence using BLEU and ROUGE metrics. The base zephyr-7b-beta

380 model was also prompted to generate sentences for the BLEU and ROUGE metric
381 and the pairwise comparison for human evaluation. Identical answers between the
382 finetuned and the base model were removed to in the test set to ensure that the
383 model is evaluated properly. A total of 144 sentences were used to evaluate the
384 model.

385 A survey was conducted to compare the finetuned model to the base model to
386 determine if the finetuning was effective. The survey was conducted online using
387 Google Forms asked the participants to pick which of the following sentences is the
388 more accurate translation of the given sentence based on accuracy, naturalness,
389 and context. The order in which sentences from the two models were shown was
390 randomly selected to avoid bias. To improve the response rate of the survey,
391 the survey was split into multiple sets, answered by the same groups of people,
392 allowing them to answer any or all of the survey forms.

393 Chapter 4

394 Results and Discussions

395 4.1 Dataset

396 We built a dataset containing a total of 1155 Gen Z internet slang sentences and
397 their corresponding formal translations. The created dataset was then combined
398 with another dataset from Hugging Face that contains 548 Gen Z internet slang
399 and their corresponding formal translation.

400 4.2 Model Evaluation

401 4.2.1 Model Training

402 The model was trained for 7 epochs before the early stopping callback was trig-
403 gered 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.

?? 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.

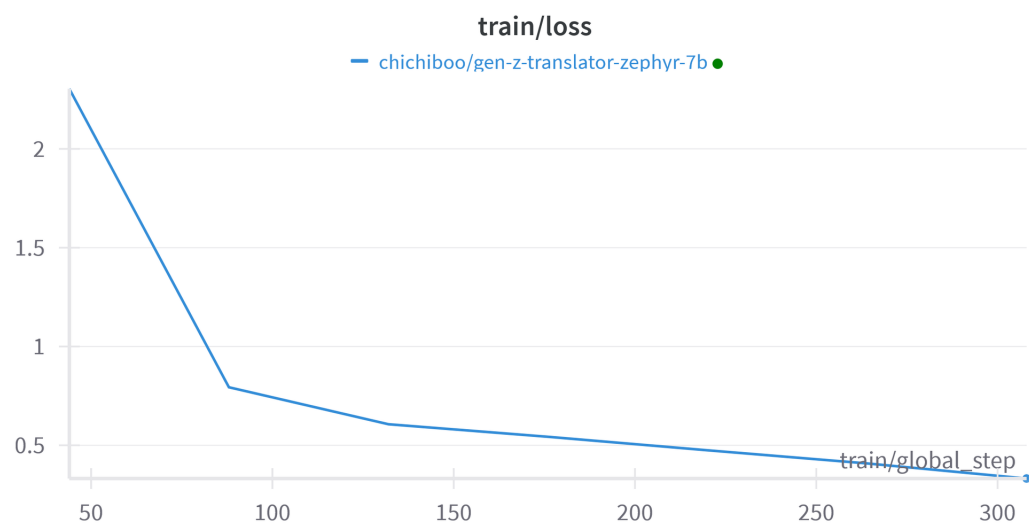


Figure 4.1: Training loss curve of the fine-tuned model across training steps

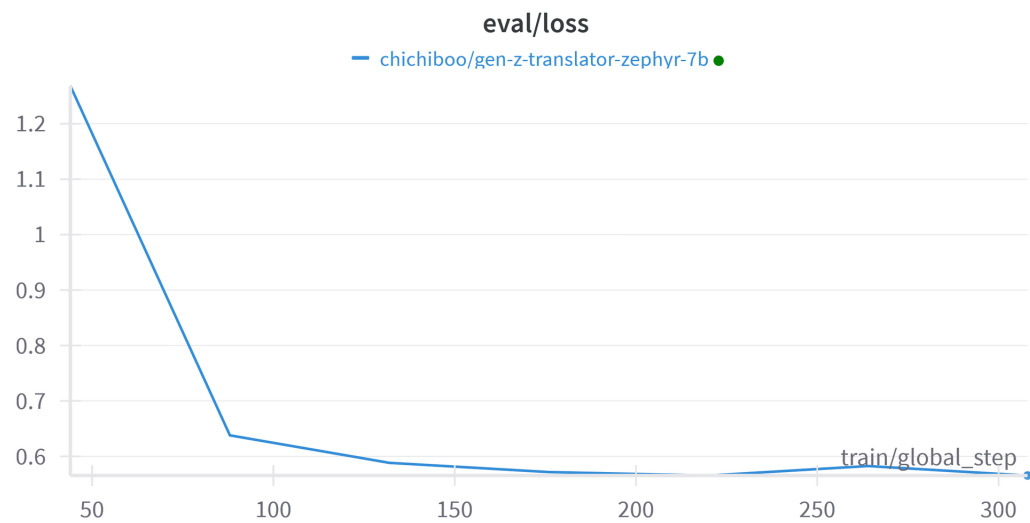


Figure 4.2: Evaluation loss curve of the fine-tuned model across training steps

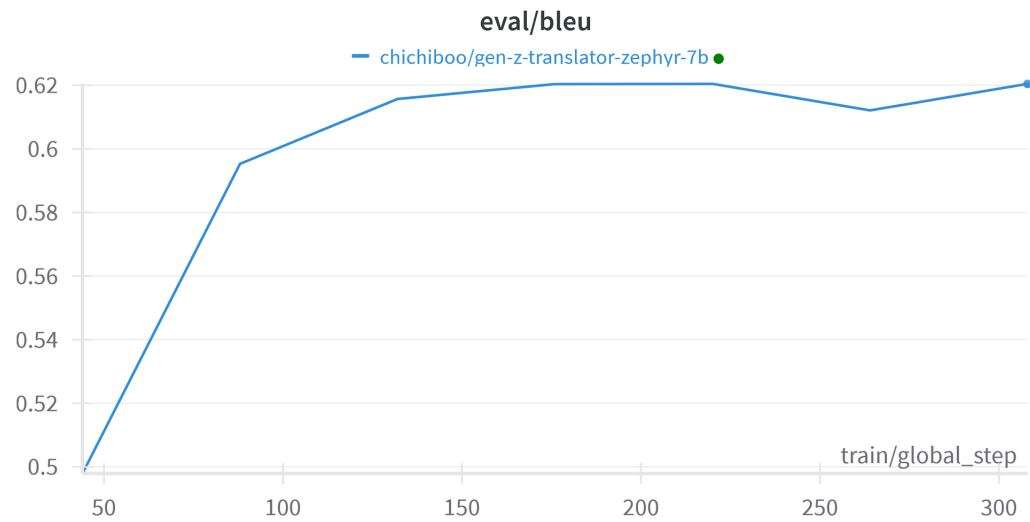


Figure 4.3: Evaluated using BLEU metric

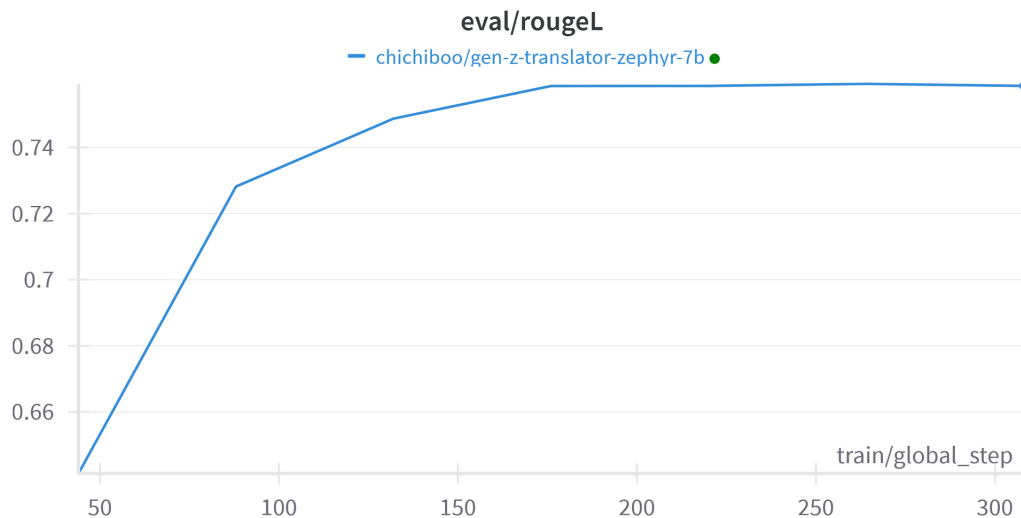


Figure 4.4: Evaluated using ROUGE-L metric

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 0.8099 and ROUGE-L Score of 0.8336 and the finetuned model obtained a BLEU score of 0.8151 and ROUGE-L Score of 0.8396. While the difference between the models

421 is minimal, this does not completely represent the performance of the models as
422 these metrics are only used to determine if the generated text is close to the refer-
423 ence text, regardless of the context and the overall quality of the generated text.
424 However, it does show that the finetuned model, while not significantly better
425 than the base model, is close to the reference model.

426 4.2.4 Manual Evaluation Metrics

427 A manual evaluation was conducted by the researchers through a survey admin-
428 istered via Google Forms to determine which of the two models is preferred by
429 Generation Z students at UPV. The survey comprised a total of 93 questions,
430 which were distributed across five separate forms. The first form contained 20
431 questions, the second 19, the third 20, the fourth 20, and the fifth 14, amounting
432 to 93 questions in total. Each question presented two translation options: one
433 generated by the fine-tuned model and the other by the base model. Respondents
434 were asked to select the translation they preferred in each case. A total of 114
435 individuals participated in the survey, with 29, 22, 22, 21, and 20 respondents
436 completing Forms 1 through 5, respectively.

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

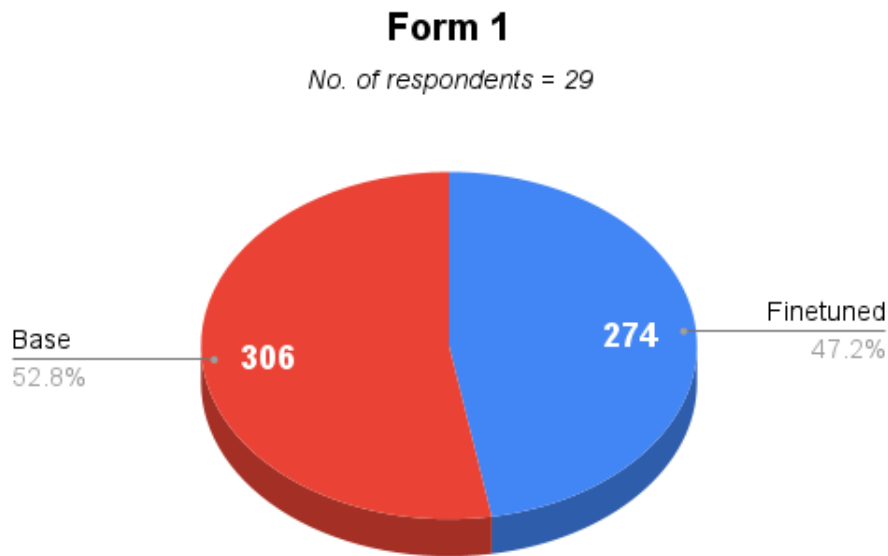


Figure 4.5: Form 1 Evaluation

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

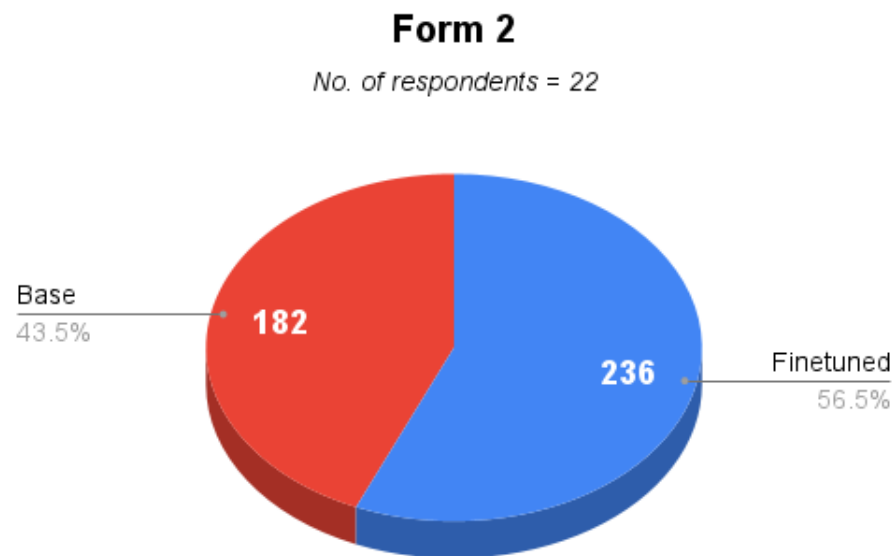


Figure 4.6: Form 2 Evaluation

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

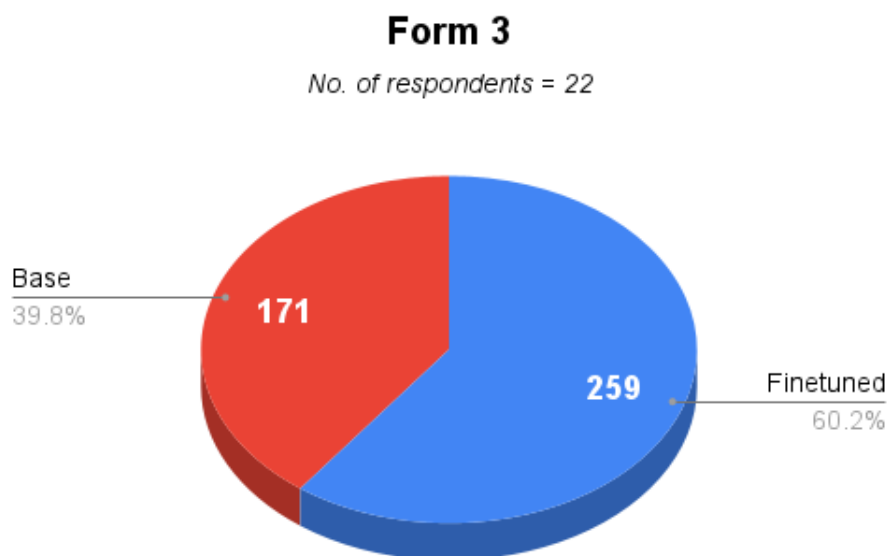


Figure 4.7: Form 3 Evaluation

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

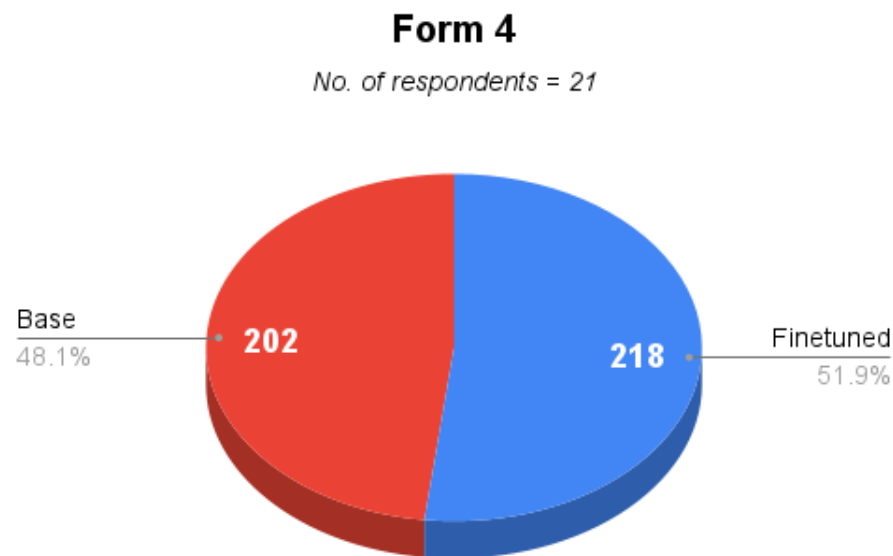


Figure 4.8: Form 4 Evaluation

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

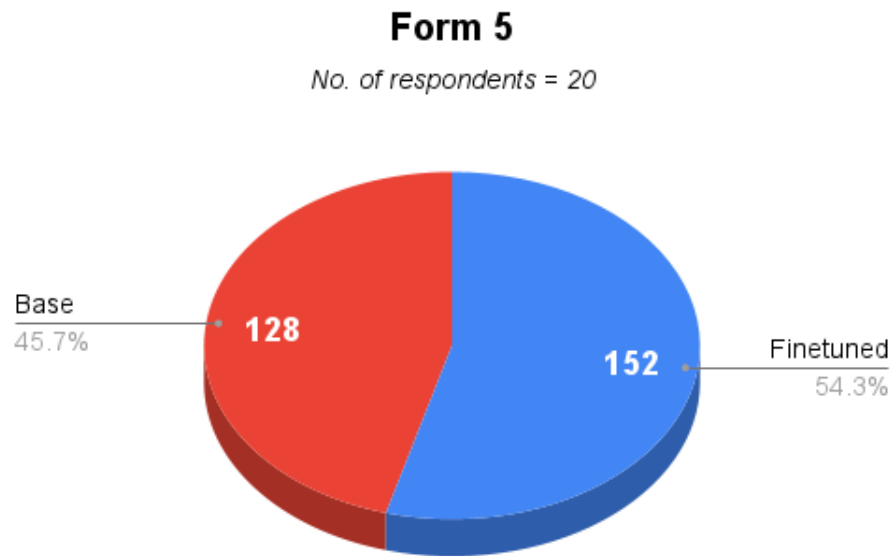


Figure 4.9: Form 5 Evaluation

464 Figure 4.9 conveys that among the 20 respondents in Form 5, 152 responses or
465 54.3 percent selected the fine-tuned model, while 128 responses or 45.7 percent
466 chose the base model. This 8.6 percent margin reinforces the general trend toward
467 the fine-tuned model across all forms.

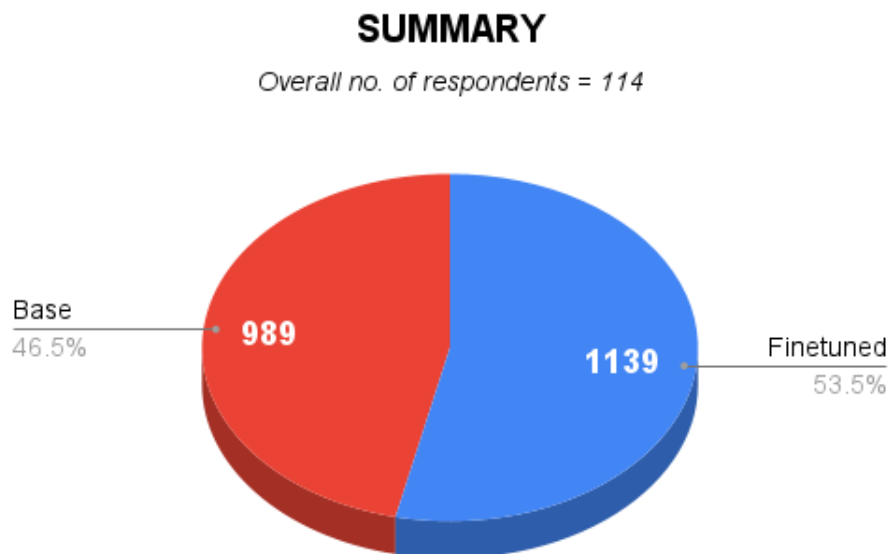


Figure 4.10: Summary Evaluation

468 Figure 4.10 presents the overall summary across all five forms, with a total of 114
469 respondents participating in the survey. In total, the fine-tuned model received
470 1,139 preferences or 53.5 percent, while the base model garnered 989 preferences
471 or 46.5 percent. The resulting 7 percent margin between the two model indicates
472 a moderate overall preference among Gen Z students at UPV for the fine-tuned
473 model, suggesting its relatively better performance in meeting the participants'
474 expectations for translation quality.

475 4.3 Summary

476 The chapter presented the evaluation results and discussions on the performance
477 of the fine-tuned language model for translating Gen Z internet slang into their

478 formal translations. The dataset used for training consisted of 1,703 sentence
479 pairs, combining original and publicly available data. The model was trained
480 for seven epochs, with early stopping employed to prevent overfitting, which was
481 evident from the divergence between training and validation losses.

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

486 To complement the results of automatic evaluation metrics, a manual evaluation
487 was carried out through online surveys among Generation Z students at UPV.
488 Participants compared translations from both models across five forms. Results
489 showed a moderate overall preference for the fine-tuned model, with 53.5% of re-
490 sponses in its favor. While one form showed a slight preference for the base model,
491 the fine-tuned model was generally preferred in the remaining forms, especially in
492 Form 3 where it showed the largest margin.

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

Chapter 5

Conclusion

In this study, we constructed 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. 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 votes. 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/unslowth>

- 541 Dua, A., Jacobson, R., Ellingrud, K., Enomoto, K., Cordina, J., Coe, E. H.,
 542 & Finneman, B. (2024, Aug). *What is gen z?* McKinsey Com-
 543 pany. Retrieved from [https://www.mckinsey.com/featured-insights/](https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z)
 544 [mckinsey-explainers/what-is-gen-z](https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-gen-z)
- 545 Euchner, J. (2023). Generative ai. *Research-Technology Management*, 66(3),
 546 71–74.
- 547 Fernández-Toro, M. (2016, Jun). *Exploring languages and cultures*. Re-
 548 trieved from [https://www.open.edu/openlearn/languages/exploring](https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2)
 549 [-languages-and-cultures/content-section-3.2](https://www.open.edu/openlearn/languages/exploring-languages-and-cultures/content-section-3.2)
- 550 Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). *Generative ai*
 551 *and chatgpt: Applications, challenges, and ai-human collaboration* (Vol. 25)
 552 (No. 3). Taylor & Francis.
- 553 Ghazali, N. M., & Abdullah, N. N. (2021, Dec). Slang language use
 554 in social media among malaysian youths: A sociolinguistic per-
 555 spective. *International Young Scholars Journal of Languages*,
 556 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)
 557 [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)
 558 [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)
- 559 Gonzaga, M. (2025, Feb). *“forda convo ang ferson”: Analysis of*
 560 *gen z slang in the lens of batstateu faculty members*. Retrieved
 561 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)
 562 [_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)
- 563 Heydari, M., Albadvi, A., & Khazeni, M. (2024). Persian slang text conversion to
 564 formal and deep learning of persian short texts on social media for sentiment
 565 classification. *Journal of Electrical and Computer Engineering Innovations*
 566 *(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>
- Lin, C.-Y. (2004, Jul). Rouge: A package for automatic evaluation of summaries. *Meeting of the Association for Computational Linguistics*, 74–81.
- Liu, J., Zhang, X., & Li, H. (2023, Aug). Analysis of language phenomena in internet slang: A case study of internet dirty language. *Open Access Library Journal*, 10(08), 1–12. doi: 10.4236/oalib.1110484
- Liu, S., Gui, D.-Y., Zuo, Y., & Dai, Y. (2019, Jun). Good slang or bad slang? embedding internet slang in persuasive advertising. *Frontiers in Psychology*, 10. doi: 10.3389/fpsyg.2019.01251
- Mantiri, O. (2010, 03). Factors affecting language change. <http://ssrn.com/abstract=2566128>. doi: 10.2139/ssrn.2566128
- Maulidiya, R., Wijaya, S. E., Mauren, C., Adha, T. P., & Pandin, M. G. R. (2021, Dec). *Language development of slang in the younger generation in the digital era*. OSF Preprints. Retrieved from osf.io/xs7kd doi: 10.31219/osf.io/xs7kd
- McArthur, T. (2003). *Concise oxford companion to the english language* (1st ed.). Oxford University Press.

- 593 Nguyen, T. T., Wilson, C., & Dalins, J. (2023). *Fine-tuning llama 2 large lan-*
 594 *guage models for detecting online sexual predatory chats and abusive texts.*
 595 Retrieved from <https://arxiv.org/abs/2308.14683>
- 596 Nocon, N., Kho, N. M., & Arroyo, J. (2018, Oct). Building a filipino colloquialism
 597 translator using sequence-to-sequence model. *TENCON 2018 - 2018 IEEE*
 598 *Region 10 Conference*, 2199–2204. doi: 10.1109/tencon.2018.8650118
- 599 Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). Bleu: a method for
 600 automatic evaluation of machine translation. *Proceedings of the 40th Annual*
 601 *Meeting on Association for Computational Linguistics - ACL '02*. Retrieved
 602 from <https://dl.acm.org/citation.cfm?id=1073135> doi: [https://doi](https://doi.org/10.3115/1073083.1073135)
 603 [.org/10.3115/1073083.1073135](https://doi.org/10.3115/1073083.1073135)
- 604 Suslak, D. F. (2009). The sociolinguistic problem of generations. *Language Com-*
 605 *munication*, 29(3), 199–209. Retrieved from [https://www.sciencedirect](https://www.sciencedirect.com/science/article/pii/S0271530909000196)
 606 [.com/science/article/pii/S0271530909000196](https://www.sciencedirect.com/science/article/pii/S0271530909000196) (Reflecting on language
 607 and culture fieldwork in the early 21st century) doi: [https://doi.org/](https://doi.org/10.1016/j.langcom.2009.02.003)
 608 [10.1016/j.langcom.2009.02.003](https://doi.org/10.1016/j.langcom.2009.02.003)
- 609 Teng, C. E., & Joo, T. M. (2023). Is internet language a destroyer to communica-
 610 tion? In X.-S. Yang, R. S. Sherratt, N. Dey, & A. Joshi (Eds.), *Proceedings of*
 611 *eighth international congress on information and communication technology*
 612 (pp. 527–536). Singapore: Springer Nature Singapore.
- 613 Vacalares, S. T., Salas, A. F. R., Babac, B. J. S., Cagalawan, A. L., & Calimpong,
 614 C. D. (2023, Jun). The intelligibility of internet slangs between millennials
 615 and gen zers: A comparative study. *International Journal of Science and*
 616 *Research Archive*, 9(1), 400–409. doi: 10.30574/ijjsra.2023.9.1.0456
- 617 Vergo, T., Godbout, J.-F., Rabbany, R., & Pelrine, K. (2024). *Comparing gpt-4*
 618 *and open-source language models in misinformation mitigation*. Retrieved

619 from <https://arxiv.org/abs/2401.06920>
620 Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., ... Rishi, D.
621 (2024). *Lora land: 310 fine-tuned llms that rival gpt-4, a technical report*.
622 Retrieved from <https://arxiv.org/abs/2405.00732>