EVALUATING LLM COMPREHENSION OF SPANGLISH

Kailen Todd McCauley
Georgia Institute of Technology College of Computing
kailen.mccauley@gatech.edu

Scott Lenney
Georgia Institute of Technology College of Computing
slenney3@gatech.edu

1 INTRODUCTION

The continuing growing Hispanic population in the United States has led to shifts in multilingualism and cultural diversity not previously seen. With many of the older generations that have immigrated or migrated to the mainland, formal English education is limited, often resulting in code-mixing into an interlanguage known as Spanglish. Spanglish use is widespread among U.S. Hispanics with 63% reporting speaking it on occasion, and 40% using the interlanguage often (Pew Research, 2023).

This growth of Spanglish can be seen to have grown over the generations as well. In a survey published by the U.S. Hispanic Chamber of Commerce and the marketing agency Chemistry Cultura, roughly 20% of Gen Z Latinos in the United States reported speaking comfortably in Spanglish frequently. This is a large increase, as ”only 14% of millennials, 10% of Gen Xers and 5% of boomers said they were most comfortable speaking Spanglish most of the time” (Franco, 2023). It's clear that Spanglish is on the rise.

In tandem with this, we have seen the recent rise of Large Language Models (LLMs) in many facets. With these LLMs being integrated as tools in many sectors, questions arise about their effectiveness in handling these complicated interlanguages, such as Spanglish, as they often have undefined grammatical rules and vocabulary, such as Spanglish. Traditional LLMs, which are predominantly trained on monolingual data (especially English), often struggle with mixed-language contexts while showing bias towards English-centric abilities (Yuemei et al., 2024).

This paper aims to explore the ability of LLMs to process Spanglish without favoring either language and contribute to the efforts of making NLP more representative of linguistic diversity.

2 PRIOR RELATED WORK

The challenges of processing code-mixed language, particularly Spanglish, have garnered increasing attention in natural language processing (NLP) research. This was most notably seen in SemEval-2020 Task 9, which introduced the SentiMix dataset comprising annotated Hinglish and Spanglish tweets for sentiment analysis (Patwa et al., 2020). They highlighted that transformer-based models, like XLM-RoBERTa, along with ensemble methods were the top performers, achieving F1 scores up to 80.6% on the Spanglish subset (Patwa et al., 2020). These results underscore the efficacy of multilingual transformers in handling code-mixed data, yet they also highlight the persistent challenges due to the scarcity of high-quality, annotated code-mixed corpora.

Recent studies have delved into enhancing LLMs' capabilities to manage code-mixing. Zhang et al. proposed a new approach that utilized reinforcement learning from AI feedback called CHAI (Code Mixed Understanding via Hybrid AI Instruction) to fine-tune multilingual LLMs on code-mixed tasks (Wenbow et al., 2024). Through the CHAI framework, they were able to improve model performance on code-mixed inputs by 25.66% in code-mixed translation tasks [5].

The limitations of LLMs in multilingual contexts extend beyond code-mixing. Studies have shown that models like ChatGPT exhibit significant performance disparities across languages, often favoring English due to the predominance of English data in training corpora (Wenhao et al., 2023). This bias not only affects model accuracy but also raises concerns about cultural representation and fairness in AI applications.

One of the most relevant papers to our work is by Syamkumar et al. (2024), which focuses on the challenges large language models face when dealing with bilingual or code-mixed inputs in educational settings. Their work evaluates how well these models can understand and assess student responses written in English, Spanish, or Spanglish—specifically looking at explanations in Science and Social Science. Utilizing both LLama 3.1 and Minstral Nemo, English student responses were graded the best (0.90 vs 0.84), followed by Spanish (0.83 vs 0.73), and Spanglish (0.74 vs 0.69) (Syamkumar et al., 2024). Overall, there was a bias towards monolingual modes, particularly English, when grading. When fine-tuning models using Spanglish examples, the accuracy went up noticeably across all three languages, signifying not only an improvement in the target domain (Spanglish) but also strong cross-language transfer in English and Spanish (Syamkumar et al., 2024). The paper points out a major issue: a lack of high-quality, Spanglish data to be utilized for various NLP tasks (Syamkumar et al., 2024). While synthetic data from LLMs might help fill the gap, that Syamkumar et al. generated for their experimentation, it risks reinforcing the same kinds of biases that already exist.

Across these works, the necessity for developing and fine-tuning LLMs that can handle the complexities of code-mixed languages is clear. They also underscore the broader implications of language biases in AI, advocating for more inclusive and representative training methodologies to ensure equitable AI applications across diverse linguistic contexts. Lastly, there is a clear call to gather and annotate more interlanguage and code-mixed datasets for other NLP tasks.

3 Data

Due to the fact we are solely testing these models for our research, we opted to use the Spanglish set provided by the SemEval-2020 Task 9 paper (Patwa et al., 2020). This data was originally collected from Twitter based on bilingual regions and code-mixing behavior, particularly focusing on cities in Texas and California, in addition to New York City and Miami (Patwa et al., 2020). The dataset was curated to prioritize tweets with code-mixed content, forming the SentiMix corpus, and includes a small number of monolingual English and Spanish tweets for balance (Patwa et al., 2020).

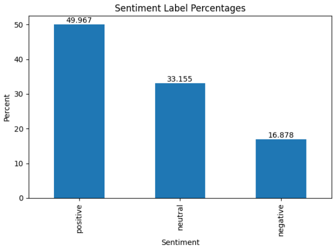

Figure 1: Sentiment Label Percentages

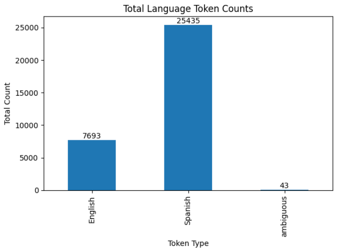

Figure 2: Token Language Counts

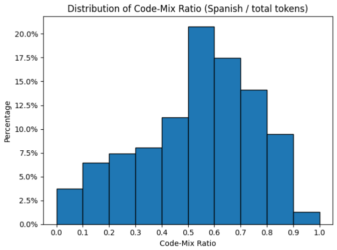

Figure 3: Code-Mix Ratio Distribution

From the exploratory data analysis of the testing data, we see a higher percentage of positive and neutral sentiment sentences from tweets. Overall the collected sentences contain a much higher count of Spanish words present as shown in Figure 2. Lastly, we wanted to highlight the ratio of Spanish to English within each sentence. We see that most sentences are predominantly Spanish words. It is important to note the extrema of sentences with less than 10% Spanish or greater than 90%. These are mostly likely examples of borrowing, in which one word from English or Spanish is utilized, whereas the other sentences with a more even distribution are code-mixing, which is a shift between speaking in both English and Spanish mid-sentence.

4 Model

We approach Spanglish sentiment classification using a causal large-language model (LLM) paired with lightweight adapter fine-tuning. Specifically, we employed a LLaMA-based 3B model from Hugging Face. In order to reduce memory footprint, we quantize the model to 4-bit precision using BitsAndBytesConfig. We then inject Low-Rank Adapters (LoRA) which we tested different rank and dropout values in addition to freezing all original model weights so that only adapter parameters are updated during training. During inference, we reload the fine-tuned adapters in inference mode and prompt the model with a structured JSON instruction to return three outputs per tweet.

- **prediction**: one of {*positive, negative, neutral*},
- **confidence**: a scalar between $[0, 1]$,
- **reason**: a brief natural-language justification.

5 Experiments

For this paper, there were three main evaluation settings being tested: zero-shot, LoRA fine-tuned, and few-shot prompt probing. All of these were run on an NVIDIA A100 GPU using Transformers and PEFT.

The Zero-shot method consisted of applying the pretrained Llama-3B-Instruct model directly to the 1,000-tweet test set, prompting each example with our JSON template and extracting the prediction and confidence level. The LoRA fine-tuned used fine-tuned adapter weights on the training split for and re-evaluated on the same test set under identical prompting. Lastly, we did few-shot prompt probing for both zero-shot and fine-tuned models. We varied the number of in-prompt demonstrations randomly to assess the impact of few-shot context on classification accuracy.

In each experiment, we computed per-class accuracy, the number of correctly labeled tweets in each sentiment category, and overall accuracy on the held-out test set. We also perform a qualitative analysis of the generated reason field to determine whether the justifications of the model align with human intuition.

6 Results

The model we developed resulted in an overall 84.46% accuracy for the test data set that we fed to it. Of the data that was fed to the model, 67% were positive sentiments, about 27% were neutral and the remaining 6% were negative.

Table 1: True vs. Correct Predictions by Sentiment Category

Sentiment Category	**Total Instances**	**Correct Predictions**
Positive	168	140
Neutral	69	62
Negative	14	10

The results showed the model to be most precise when handling positive tweets, followed by neutral tweets, and the negative tweets being the least precise by a large margin. However, the recall was relatively similar for each of the sentiments, but the neutral sentiment's recall was slightly higher. Despite the recall being higher for the negative in comparison to the precision, the F1 scores show that the model still struggles the most with negative sentiment but is able to perform about the same for positive and neutral sentiments. The confusion matrix shows that most of the false predictions are made assuming that a positive sentiment is either neutral or negative. This means that the model tends to assume that a tweet is not positive despite being classified as positive.

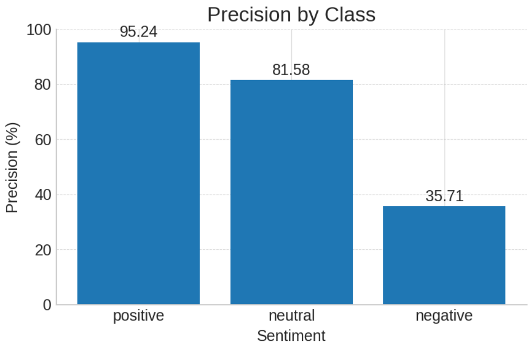

Figure 4: Precision

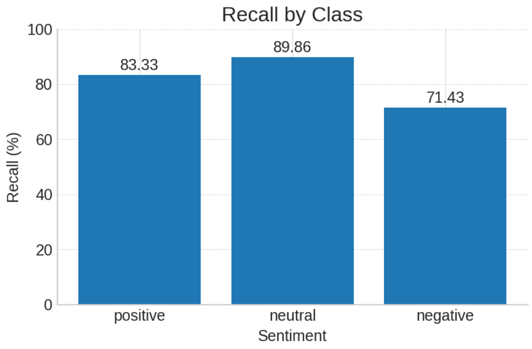

Figure 5: Recall

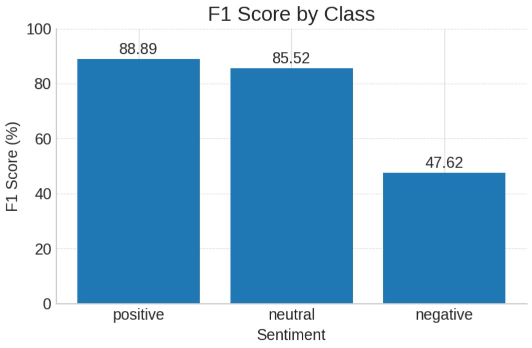

Figure 6: F1

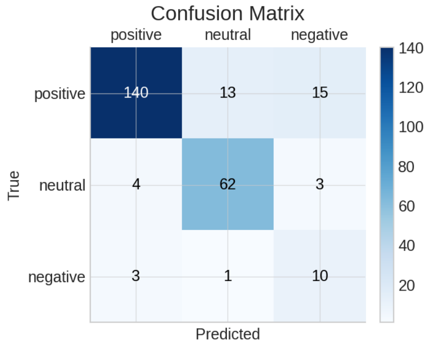

Figure 7: Confusion Matrix

When discussing the qualitative results of the experiment, the reason category is useful to understand why the model picked the sentiment that it did. The model's predictions mainly depend on obvious lexical clues in the text. If it sees swear words or angry/sad emojis (like "fuck," "odio), it almost always calls the post negative. When it detects laughter ("jajaja," "lol"), heart emojis, or excited exclamations ("OMG," "VAMOOOOOO"), it marks it positive. When none of these strong signals appear and it sees just links, hashtags, or plain statements, it defaults to neutral. Most of its mistakes happen when people use sarcasm or mild humor without clear negative words, or when casual laughter makes a neutral post look positive. The model works very well when experiencing direct statements, however, fails to catch things like sarcasm or hidden meaning that a human would understand.

7 Conclusions

In this work, we presented a systematic evaluation of a 3B-parameter LLaMA model, enhanced via 4-bit quantization and LoRA fine-tuning, on the task of Spanglish sentiment classification. Using the SemEval-2020 SentiMix Spanglish corpus, we demonstrated that our adapted model achieves an overall accuracy of 84.5%, with high precision and recall on positive and neutral examples, but exhibits substantially lower precision on negative instances. Through both quantitative metrics and qualitative analysis of the model's generated rationales, we showed that its decisions are driven primarily by explicit sentiment markers—profanity, emojis, and exclamations—while more subtle or context-dependent expressions remain challenging. These findings confirm that lightweight adapter fine-tuning can significantly improve code-mixed language understanding without retraining the entire model, yet also highlight persistent weaknesses in handling nuanced negative sentiment. Overall, our study underscores the promise of parameter-efficient approaches for multilingual and code-mixed NLP tasks and provides a clear diagnostic of where further model and data innovations are needed to achieve robust, equitable performance across all sentiment categories.

8 Future Works

There are many avenues of continued research in this domain. While our current study focuses on sentiment classification, the same Spanglish corpus could be used to evaluate model performance on other sequence-labeling tasks— for example, cross-language part-of-speech tagging or named-entity recognition in code-mixed text. Second, by sourcing or annotating a larger Spanglish dataset with explicit geographic labels (e.g. Texas, California, New York), one could perform a regional comparison to determine whether state-of-the-art LLMs exhibit varying accuracy when predicting sentiment (or other linguistic phenomena) across different bilingual communities, highlighting regional biases. Third, our experiments relied on a single GPT-style model from Hugging Face. Further research could extend this evaluation to include additional LLMs to discern if the results we observe are model-specific or generalized across architectures and training regimes. Together, these extensions would deepen our understanding of how large language models handle code-mixed language. Moreover, it could be used to guide more robust, regionally sensitive LLMs, and provide a basis for expanding this understanding.

9 References

Marina E. Franco. Gen Z Latinos embrace ”Spanglish”. Axios, 2023. https://www.axios.com/2023/06/20/latino-gen-z-language-spanish-english-spanglish

Patwa, P., Aguilar, G., Kar, S., Pandey, S., Solorio, T., & Das, A. SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020. https://aclanthology.org/2020.semeval-1.100.pdf

Pew Research Center. Latinos' Views of and Experiences With the Spanish Language. Pew Research Center, 2023. https://www.pewresearch.org/race-and-ethnicity/2023/09/20/latinos-views-of-and-experiences-with-the-spanish-language/

Syamkumar, A., Tseng, N., Barron, K., Yang, S., Karumbaiah, S., Uppal, R., & Hu, J. Improving Bilingual Capabilities of Language Models to Support Diverse Linguistic Practices in Education. arXiv preprint arXiv:2411.04308, 2024. https://arxiv.org/abs/2411.04308

Wenbo Zhang, Aditya Majumdar, and Amulya Yadav. CHAI for LLMs: Improving Code-Mixed Translation in Large Language Models through Reinforcement Learning with AI Feedback. arXiv preprint arXiv:2411.09073, 2024. https://arxiv.org/abs/2411.09073

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. arXiv preprint arXiv:2304.04675, 2023. https://arxiv.org/abs/2304.04675

Yuemei Xu, et al. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. arXiv preprint arXiv:2404.00929, 2024. https://arxiv.org/abs/2404.00929

10 Task Breakdown

Table 2: Project Work Breakdown

Task	**Person**
Research Paper Summary	Kailen McCauley
Project Proposal Outline	Scott Lenney
Prompt Engineering and Design	Kailen McCauley
Data Collection	Kailen McCauley
Baseline Evaluation	Scott Lenney
Evaluation of Data	Scott Lenney
Write Up Report	Scott Lenney and Kailen McCauley