Evaluating LLMs for Translations

Miki Fukushima mfukushima@g.ucla.edu
Esdras Aleman ismael98@ucla.edu
Jason Tay jtay20@g.ucla.edu
Trisha Agrawal trishaagrawal@g.ucla.edu

Abstract

In recent years, commercial large language model (LLM) chatbots have become prevalent, handling various tasks from text summarization to query responses. However, language translation remains a challenging task due to potential discrepancies introduced by ambiguity, context limitations, and nuanced meaning comprehension. This study evaluates the translation capabilities of ChatGPT and Google Gemini using Bible passages as a benchmark, assessing accuracy, fluency, and sentiment. We employed BLEU, METEOR, and cosine similarity with BERT embeddings for accuracy, human annotation for fluency, and VADER and TextBlob for sentiment analysis. Our publicly available dataset of Bible excerpts in English, Spanish, and German reveals that both LLMs perform better with Spanish translations and exhibit some sentiment alterations.

1 Introduction to Problem

1.1 Motivation

The widespread adoption of commercial large language model (LLM) chatbots has enabled them to perform a variety of tasks, from text summarization to answering queries. However, translating languages remains a complex and nuanced challenge. LLM translations can be incorrect or significantly alter the original meaning due to factors like source language ambiguity, lack of context, or the models' limitations in grasping subtle nuances. Additionally, different LLMs may handle translation tasks with varying levels of success, and some models might perform better with specific languages. This research seeks to investigate these performance discrepancies by evaluating the translation capabilities of popular LLMs: ChatGPT and Gemini Advanced. In particular, we seek to evaluate whether these LLMs are able to take an input text from one language and translate it into another language while preserving semantic meaning and syntactic structure. Before we can make this evaluation, we needed to create a database of phrases in various languages. In particular, our dataset contains phrases from the Bible in English, Spanish, and German, and is publicly available on our GitHub repository as listed in Section 4.3. Based on the analysis of the evaluation metrics for accuracy, we were able to conclude that ChatGPT and Gemini perform better when translating Spanish statements to English. We also conducted further analysis on how well the LLMs are able to preserve the sentence's sentiment. We found that both LLMs frequently change the overall sentiment of a translated phase, regardless of the language inputted.

1.2 Related Work

1.2.1 Prior Research on LLM Translations

In "Translation errors significantly impact low resource languages in cross-lingual learning"[1], the authors propose evaluating the disparity in performance between zero-shot evaluations on human and machine-translated texts across multiple languages using established benchmarks like XNLI. Their study underscores the presence of translation errors, particularly in low-resource languages like Hindi and Urdu, emphasizing the importance of high-quality translations for cross-lingual transfer evaluation.

In "Quantifying multilingual performance of large language models across languages"[6], the authors explore limitations issues with LLMs in regards to the extensive text corpora required for training, which therefore makes the LLMs tend to exhibit lower proficiency for low-resource languages. As there is currently no systematic method to perform quantitative evaluations against those languages, authors introduce the Language Ranker framework which assess and rank languages based on LLM performance. The study reveals discoveries which underscore the utility of the Language Ranker as a methodology for appropriately evalu-

ating LLM performance across multiple different languages and states it may provide further insights and guidance to future research in language model development.

The author of "A comparative analysis of translation brief and persona prompts"[4] examines the translation capabilities of commercial LLM chatbots, notably ChatGPT, under different prompts and personas. They find suboptimal performance in translation tasks and explore the impact of prompt engineering on translation quality, utilizing BLEU and COMET-22 scores for evaluation. Their results indicate that a simple translation prompt yields better performance than prompts including brief content summaries, although assigning the role of "Translator" to ChatGPT improves performance. These findings underscore the potential of prompt engineering to enhance ChatGPT's translation capabilities, aligning with the need for improved translation accuracy and fluency in various tasks, as highlighted in the abstract.

1.2.2 Metrics for evaluating translations

"Bleu: a method for automatic evaluation of machine translation"[7] introduces the BLEU metric, which has become a standard method for evaluating the quality of machine translations. BLEU calculates the precision of n-grams (contiguous sequences of n words) in the generated translation compared to one or more reference translations. It rewards translations that contain similar n-grams to the references, accounting for variations in the length of the translations. BLEU offers the benefits of being quick, inexpensive to calculate, language-independent, and being highly correlated with evaluations by human translators

Lavie and Agarwal[5] proposes the METEOR metric as an alternative to BLEU for evaluating machine translations. METEOR computes a score based on the harmonic mean of precision and recall, considering unigram matches, stemming variants, and synonym matches. It incorporates an explicit concept of semantic similarity through synonym matching and word stemming, offering a more flexible approach to word choice and order.

Devlin et al.[3] introduce the BERT model, revolutionizing natural language processing tasks by pre-training a deep bidirectional transformer architecture on large text corpora. BERT's pre-training involves two tasks: masked language modeling (MLM), where random words are masked and the model is trained to predict them, and next sentence prediction (NSP), where the model learns to predict whether two sentences are consecutive in the input text. This pre-training approach enables BERT to capture bidirectional contextual representations of words, leading to state-of-the-art performance on a wide range of NLP tasks.

2 Method

As shown in Figure 1, the methodology followed in this research can be broken up into four sections; dataset creation, LLM Translation Generations, Evaluation Metrics, and Analysis.

The dataset creation process is explained in depth in Section 3.2 of this paper. The LLM Translation Generations section of this research is closely related to the dataset creation process because the translated sentences were manually extracted from outputs produced by ChatGPT and Gemini. The output from both LLMs were then manually inserted into the final dataset, titled translations_output.json, which can be found in the GitHub repository linked in Section 4.3.

2.1 Metrics

The main part of this research consists of the various ways the outputs produced by the two LLMs were analyzed based on accuracy, sentiment, and fluency. As outlined in Figure 1, we used BLEU, METEOR, and BERTScore for the accuracy evaluation metric portion of this research.

We used BLEU, METEOR, and computed the cosine similarity for BERT embeddings to measure the accuracy of the translation outputs from ChatGPT and Gemini. BLEU, which emphasizes lexical overlap through n-gram matching, offers precision-based measures but may struggle with semantic nuances and variations in wording. METEOR analyzes semantic similarity through synonym matching and stemming, providing flexibility in word choice and order, yet it may not fully capture higher-level semantic relationships. Additionally, the cosine similarity of BERT embeddings captures both structural and semantic similarities between sentences, leveraging contextual representations for assessment.

We used VADER and TextBlob for the overall sentiment analysis portion of this research. VADER is a rule-based sentiment analysis tool, which is fully open source under the MIT license. We have performed the evaluations using the Python language utilizing the vaderSentiment

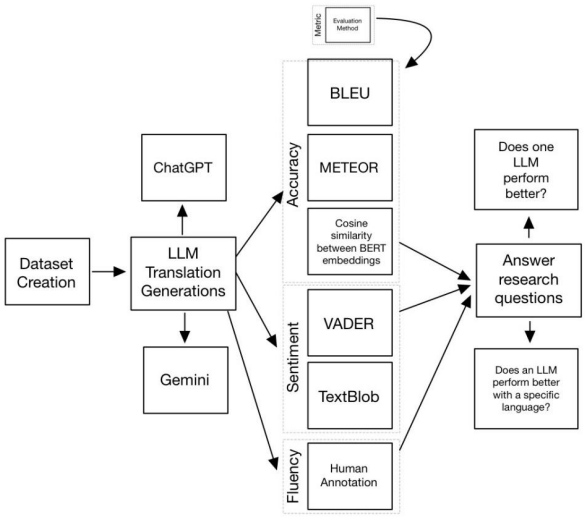

Figure 1: Methodology Figure

3.3.2 library. From this library, we called a function SentimentIntensityAnalyzer which enabled us to create the analyzer object and further invoke a method of polarity scores. By providing each of the passages as an input, the method returns the sentiment dictionary which contains pos, neg, and neu compound scores. We can then determine and map the resulting values to Positive, Negative, or Neutral depending whether the score is larger or equal to 0.05 or less than or equal to -0.05. Any values that do not align within these ranges are categorized as neutral. Similarly, for the TextBlob evaluations, we imported the TextBlob Python library to conduct the analysis. TextBlob is also an open source library and it is primarily used for processing textual data. First, we created a blob object by providing each of the passages as an input, then invoked additional methods such as sentiment polarity and subjectivity. Polarity scores can range from -1.0 to 1.0 and subjectivity scores can range between 0.0 to 1.0. 0 indicates that the sentence is very objective, whereas 1 indicates that it is very subjective.

Finally, we used human annotation to determine the fluency of the translations from the LLMs. Esdras Aleman, a native Spanish speaker, conducted the human annotation on the dataset based on three metrics; fluency, accuracy, and punctuation. Each translation was evaluated based on these three metrics and given a score from 1 to 5.

3 Experiments

3.1 Baseline Model Selection

Google's Gemini Advanced (Gemini 1.0 Ultra) and OpenAI's ChatGPT (GPT 3.5) were selected as the baseline models for this research.

In a related work, Sui He[4] found that using simple prompts with an LLM provided better results compared to more complex prompts. Based on this discovery, we chose to use a simple and straightforward prompt when collecting translations from LLMs. The prompt that was consistently used for this research was: "Translate this to English: ", with the Spanish or German statement directly after.

3.2 Dataset Creation

The dataset used to conduct this research was created manually by extracting datasets from a website[2] that consisted of the Bible in multiple languages. 53 sentences from the New Testament, Book of Matthew, were used for this research. Each sentence in English, Spanish, and German were taken from this section of the Bible.

Our final dataset, translations_output.json, consists of 106 objects, 53 objects containing Spanish to English translations and another 53 objects containing German to English translations. Each object consists of four values; "English", "Spanish" or "German", "gpt_english" for output from ChatGPT, and "google_english" for output from Google Gemini.

3.3 Evaluation Metrics

3.3.1 Accuracy

We have selected three evaluation methods for analyzing the accuracy of translations generated by LLMS; BLEU, METEOR, and BERTScore. To implement these, we utilized several libraries, including NLTK, PyTorch, and the transformers library from Hugging Face. First, we downloaded necessary NLTK data and initialized a pre-trained BERT model and tokenizer (bert-base-uncased). For sentence preprocessing, we tokenized and converted each sentence to lowercase. For BLEU and METEOR scores, we defined a function to calculate both metrics for a given reference and candidate sentence. BLEU score was computed using NLTK's sentence_bleu with a smoothing function to handle short sentences, ensuring more robust and reliable scores. METEOR score was calculated using NLTK's meteor_score, which considers synonyms and paraphrasing, providing a more comprehensive evaluation of semantic similarity. To compute cosine similarity, we leveraged BERT embeddings. We created a function to tokenize sentences, add special tokens, and extract embeddings from BERT. Specifically, we used the embedding of the [CLS] token, which represents the entire sentence. Another function was used to calculate the cosine similarity between two such embeddings, utilizing cosine_similarity from sklearn. This multi-faceted approach ensured a thorough evaluation of our model's textual accuracy.

3.3.2 Fluency

Fluency of the translations from Spanish to English was evaluated by Esdras Aleman, who is a native Spanish Speaker. The Spanish to English translations generated using Google LLM and ChatGPT were evaluated using three metrics: Translation Fluency, Translation Accuracy, and Translation Punctuation. For each metric, a translation was given a score from 1 to 5, where one indicates the lowest possible value given the metric, and 5 indicates the highest possible value.

Translation Fluency was evaluated by determining how well the translated English text flowed in English. For example, if the translated text was not fluid in English, the text would receive a lower score, and vice versa.

Translation Accuracy was evaluated by determining how accurate the translated text was to the original text. If the original text used certain language, then to get a higher score, the translated language must also use similar language. If the translated text used certain descriptive language, then the translated text must use similar language. For example, if the original text described a lake shore, then the translated text could not describe the same location as the shore of sea. Translated texts that used inaccurate translations for phrases or words received a lower score for this metric.

Translation Punctuation was evaluated by seeing how well the translated text preserved the punctuation of the original text. If the original text was split into several sentences, then the translated text must have similar number of sentences, and the sentences must have been punctuated well. In general, a translated text was evaluated by how well the translation kept the structure of the original text.

3.3.3 Sentiment

We have selected two methodologies to validate the change in sentiment of the original English and translated English passages. First, general sentiment was measured using VADER to identify if the meaning or nuances of the sentence changed when comparing the ChatGPT and Gemini translations with the base English translations. The resulting score was grouped into three separate categories: positive if it is greater than 0.5, negative if it is less than 0.5; otherwise it is neutral. To enable visualization, we then assign 1, -1, 0 to the positive, negative, neutral category, respectively. Table 1 and Table 2 show the VADER scores for the LLM translations, grouped by Positive, Neutral and Negative.

Second, TextBlob was used in addition to evaluate the polarity and subjectivity. For each sentence, we generated two scores which ranged from -1.0 to 1.0 for polarity, 0 to 1 for subjectivity. For subjectivity, 0 indicates that the sentence is very objective, and 1 indicates that the sentence has a strong sentiment that it is highly subjective.

3.4 Results

3.4.1 Quantitative Analysis

By looking at the BLEU, METEOR and BERT scores outlined in Table 3, Table 4, and Table 5, respectively, we can see that ChatGPT generally performed better than Gemini in terms of precision with the exception of the BLEU analysis conducted on translated English originated from Spanish. Across all evaluations, the BERT similarity

Dataset	Positive	Neutral	Negative
English	6	32	15
Spanish	0	47	6
GPT English	8	30	15
Google English	5	35	13

Table 1: VADER scores grouped by Positive, Neutral, Negative respectively for English, Spanish, and ChatGPT and Google translated English passages. This table is for the Spanish to English translations.

Dataset	Positive	Neutral	Negative
English	6	32	15
German	1	27	25
GPT English	4	36	13
Google English	6	33	14

Table 2: VADER scores grouped by Positive, Neutral, Negative respectively for English, German, and ChatGPT and Google translated English passages. This table is for the German to English translations.

score is generally higher than the BLEU and METEOR scores when average scores were compared against each result.

Table 1 and Table 2 outline the aggregated total number when the output was categorized as Positive, Negative, and Neutral respectively based on VADER analysis. When the untranslated English and translated English through GPT and Google LLM are compared, the number of positive sentiment increased by 0.33 for GPT and decreased by 0.16 for the Google LLM, decreased by 0.0625 for GPT and increased by 0.09375 for the Google LLM for the neutral sentiment, and lastly for the negative sentiment, the number remained consistent for the GPT and decreased by 0.133 for the Google LLM for the Spanish translations. For the German translations, the number of positive sentiment decreased by 0.33 for GPT and remained consistent for the Google LLM, the number of neutral sentiment increased by 0.125 for GPT and 0.03125 for the Google LLM, lastly the number of negative sentiment decreased by 0.13 for GPT and 0.066 for the Google LLM.

In addition to VADER, we have evaluated the polarity and subjectivity using TextBlob. For base English, approximately 85 percent of the passages scored between 0 and 0.5, indicating that they contain rather objective sentiment. For translated English originating from Spanish, 45 of the total 53 passages also scored between 0 and 0.5 exhibiting that they contain objective sentiment for the GPT. There is a slight increase for the Google LLM where 46 out of a total of 53 passages scored between 0 and 0.5. For the translated English originated from German, 92 percent of GPT translated English contains objective sentiment whereas 77 percent of Google translated English contains the objective tones.

Finally, Table 6 and Table 7 indicate the average fluency, accuracy, and punctuation values for Spanish to English translation by ChatGPT and Google Gemini. GPT scores for all three categories exceeded slightly when compared to the score generated based on the Google LLM.

Language	Chat GPT	Gemini
English	0.699	0.749
Spanish	0.473	0.347

Table 3: BLEU Scores for English and Spanish translations when comparing the reference text with ChatGPT and Gemini.

Language	Chat GPT	Gemini
English	0.859	0.849
Spanish	0.760	0.660

Table 4: METEOR Scores for English and Spanish translations when comparing the reference text with ChatGPT and Gemini

3.4.2 Qualitative Analysis

Based on the evaluations performed using BLEU, METEOR, BERT for the similarity analysis, our observations indicate that the both OpenAI ChatGPT and Google Gemini LLM have superior performance for Spanish translation over German. We also observed that the BLEU, METEOR, BERT exhibited the sensitivity differently where BLEU scores were consistently lower across all comparisons while BERT scored consistently and noticeably higher than BLEU or METEOR.

VADER analysis revealed that Spanish passages contained more neutral sentiment whereas German

Language	Chat GPT	Gemini
English	0.970	0.954
Spanish	0.925	0.895

Table 5: BERT Scores for English and Spanish translations when comparing the reference text with ChatGPT and Gemini

Fluency	Accuracy	Punctuation
4.62745098	4.254901961	4.725490196

Table 6: Average fluency, accuracy, and punctuation values for Spanish to English translations by Gemini.

Fluency	Accuracy	Punctuation
4.75	4.615384615	4.846153846

Table 7: Average fluency, accuracy, and punctuation values for Spanish to English translations by ChatGPT.

passages contained more negative sentiment when compared to the base English passages. Though once translated to English through ChatGPT and Gemini, the overall sentiment across all passages is in alignment with the original base English.

The subjectivity was analyzed using TexBlob, passages evaluated in this study generally resulted as they contain rather objective sentiment for all base English and translated English phrases. A very small percentage of the overall passage contained highly subjective sentiment such as "*But he did not consummate their marriage until she gave birth to a son. And he gave him the name Jesus*" and "*So there were in total fourteen generations from Abraham to David, fourteen from David to the deportation to Babylon, and fourteen from the deportation to the Christ.*"

4 Conclusion

Our research and analysis demonstrates that both OpenAI ChatGPT and Google Gemini LLM exhibit superior performance in Spanish translation compared to German. The BLEU, METEOR, and BERT scores indicate that ChatGPT generally outperforms Gemini, particularly in precision, except for the BLEU score in Spanish translations. Sentiment analysis using VADER showed that translated texts maintain sentiment alignment with the original English, with ChatGPT slightly enhancing positive sentiment and reducing negative sentiment in German translations. TextBlob analysis confirmed that the majority of translated passages remain objective, with ChatGPT maintaining higher objectivity in translations. Lastly, ChatGPT scored marginally higher in fluency, accuracy, and punctuation compared to Google Gemini LLM.

4.1 Future Work

4.1.1 Variety of Prompts

Future work for this research topic could include testing the LLMs with a variety of prompts. The research we conducted involved a very simple prompt that consisted of a command. In the future, prompts with in-context examples can be tested and analyzed to see if LLMs perform better when context is given in the task of translation. Another change that can be made to the prompts is providing more instructions or requirements for the translation, such as saying that the meaning of the sentence should not change.

4.1.2 Variety of Languages

Future work for this research topic could also include testing the LLMs translation capability with more languages. Spanish, German, and English were chosen for this research for no specific reason other than it was available on the website used to find translations of the Bible. Additional translations of the Bible are available on the source we used and on other sources. By testing the LLMs with additional languages, one could determine if an LLM performs better with a specific language.

4.1.3 Different Reference Text

The Bible was used as the reference text for this research because of how publicly available the text is in various languages. Future work could involve using a different reference text for translations, one that is not so publicly available, in order to see if the LLM would perform differently with a less known reference text.

4.1.4 Persona

Finally, giving an LLM a persona, such as a writer or translator, could affect the translations by making the translations more accurate.

4.2 Work Distribution

Jason - Accuracy metrics code (BLEU, METEOR, BERTScore), report, slides
Miki - Sentiment analysis metric code (VADER, TextBlob), report, slides
Esdras - Dataset creation (LLM generations), Spanish human annotation, report, slides
Trisha - Dataset creation (find Bible phrases in three languages, Python functions to convert to JSON, LLM generations), report, slides

4.3 Source Code

This repository contains all the code used for this research as well as the dataset.

5 Appendix

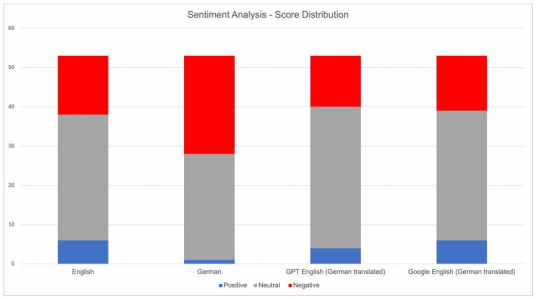

Figure 2: Sentiment score distribution for Positive, Negative, Neutral - German translated English

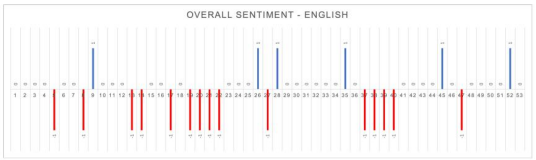

Figure 3: VADER sentiment analysis on base English passages using Positive (1), Negative (-1), Neutral (0) values

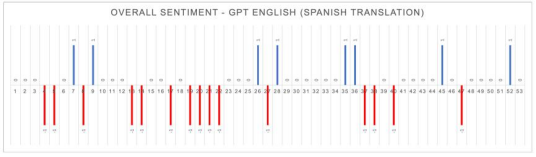

Figure 4: VADER sentiment analysis on Spanish translated (GPT) English passages using Positive (1), Negative (-1), Neutral (0) values

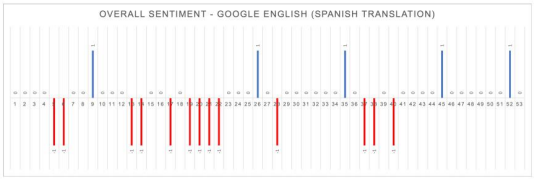

Figure 5: VADER sentiment analysis on Spanish translated (Google) English passages using Positive (1), Negative (-1), Neutral (0) values

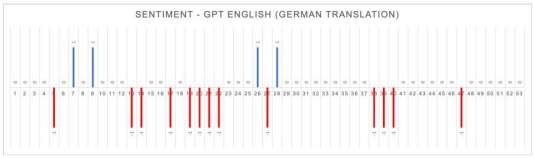

Figure 6: VADER sentiment analysis on German translated (GPT) English passages using Positive (1), Negative (-1), Neutral (0) values

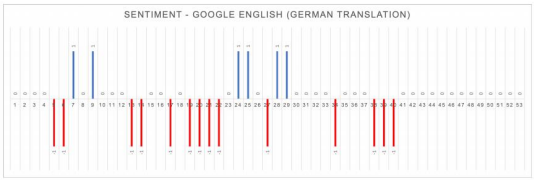

Figure 7: VADER sentiment analysis on German translated (Google) English passages using Positive (1), Negative (-1), Neutral (0) values

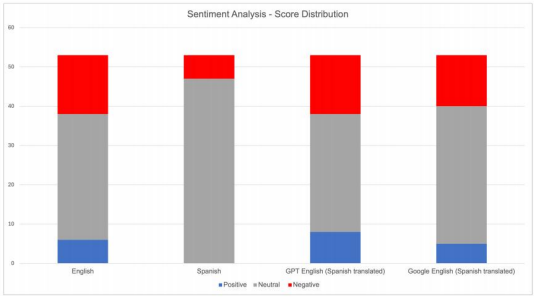

Figure 8: Sentiment score distribution for Positive, Negative, Neutral - Spanish translated English

References

[1] Ashish Sunil Agrawal, Barah Fazili, and Preethi Jyothi. Translation errors significantly impact low-resource languages in cross-lingual learning, 2024.

[2] Biblica. Matthew 1 – new international version (niv). Accessed: 2024-05-29.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[4] Sui He. Prompting chatgpt for translation: A comparative analysis of translation brief and persona prompts, 2024.

[5] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[6] Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. Quantifying multilingual performance of large language models across languages, 2024.

[7] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.