Comparative Analysis of Lexical Diversity Across Large Language Models

Henrik Gombos

August 19, 2024

Abstract

Over the past few years, generative AI models like GPT-4, LLaMa, and Claude have rapidly grown in popularity. This study compares the linguistic capacity of these models by analyzing the lexical diversity of their outputs. It answers the following questions: 1) Does translating between languages in generative AI models decrease lexical diversity? 2) Which factor, prompt quality or model quality, is more impactful to the output of the LLM? This paper runs extensive experiments on LLM outputs, using canonical lexical diversity measures and state-of-the-art computational linguistic techniques, to answer these questions.

1 Introduction

Much of the conversation around Large Language Models has focused on intelligence – how capable are these models of "thinking" like humans? However, notably less attention has been given to their linguistic capacity. The utility of these models depends directly on how well they communicate through language. This paper seeks to quantify the linguistic capacity of large language models by analyzing the lexical diversity of their responses to various prompts. The most important takeaways in the paper are the following.

1. Prompt quality generally matters more than model quality. That is, given the same prompt, the lexical diversity of response is relatively similar across models. (There are some notable exceptions – see Section 4.3).

2. Specific prompts have a higher impact on model output than generic prompts. That is, the lexical diversity of the models' outputs is more responsive to detailed prompts than vague prompts.

3. Across languages, lexical diversity scores remain relatively consistent. (This observation doesn't account for natural lexical diversity variance across languages.). That is, for all models, the lexical diversities of a text sample and its translation in another language are relatively similar.

2 Background

Prompting as a distinct aspect of interacting with LLMs and its simplicity with no need to fine-tune the model, has evolved into a nuanced field of study, highlighting the intricate relationship between user inputs and LLM responses ([4], [5]). Early explorations, such as those by, delved into how varying prompt designs could dramatically influence the performance and outputs of language models, marking the birth of prompt engineering. This area rapidly expanded, uncovering the critical role of prompts in few-shot and zero-shot learning scenarios, exemplified by work with GPT-3.5/GPT-4, where strategically crafted prompts enabled the model to perform tasks with minimal prior examples. Beyond mere task instruction, recent studies have shifted towards understanding the semantic and contextual nuances in prompts, examining how subtle changes can lead to significantly different responses from the LLM ([6], [7]).

3 Methods

This study investigates the lexical diversity of Large Language Models (LLMs), a notable aspect of language quality. The Python library Lexical Richness was used to calculate the lexical diversity of the LLM outputs.

3.1 Lexical Diversity

This project uses three standard measures for calculating lexical diversity: TTR, MTLD, and Yule's I. These three measures vary in their modes of calculation, encode different linguistic characteristics, and respond differently to outliers. In this, they lend robustness to the experimental observations.

The lexical diversity measures depend on the following quantities.

- N: total number of words in the text
- V: number of different word types

3.1.1 TTR Metric

TTR measures lexical diversity by dividing the number of word types, V, in the text sample by the total number of words, N, in the sample.

$$TTR = \frac{V}{N}$$

3.1.2 MTLD Metric

At a high level, MTLD is the average number of consecutive words that maintain a specified TTR value. The algorithm for generating an MTLD score calculates the TTR for subsets of the text. When the TTR value drops below the specified value (0.72 in our experiments), the algorithm's parameters for calculating TTR values are updated.[1]

3.1.3 Yule's I Metric

Yule's I gives a higher value for texts with lower lexical diversity and a lower value for texts with higher lexical diversity. It considers the frequency of each word in the text, giving more weight to words that appear less frequently.[2]

$$I = \frac{V^2}{(M-V)},$$

where

- $f_v(i,\ N)$ is the number of types occurring i times in a text sample of length N, and
- $M = \sum_{V}^{i=1} (i^2)(f_v(i, N))$.

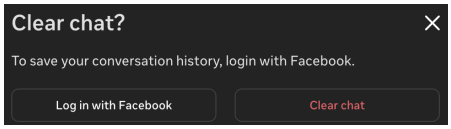
3.2 Large Language Models

OpenAI's GPT-4 (~1.76T parameters), Meta's LLaMa 3.0 (~400B parameters), Anthropic's Claude 3.5 Sonnet (~70B parameters), and Mistral AI's Large 2 (~7B parameters) are the four LLMs tested in the experiments. Since model size is the main driver for how well an LLM is perceived to perform nowadays, having a range of different-sized LLMs is crucial to testing whether the impact of prompting vs. model quality on lexical richness. Mistral AI's Large 2 is also a French-built model, providing variance from the other three models. It's expected that GPT-4 and LLaMa perform better in these linguistic experiments, given that the size of the models outpaces the other two models significantly.

The test material contains a baseline prompt, followed by a collection of eight different prompts to encourage higher or lower lexical diversity. For example, the first baseline prompt tested is "*Write a poem about apples.*" Baseline prompts do not prompt the LLM for a specified range of lexical diversity. A few examples of lexical prompting include "*Write a poem about apples* ***in the***

[1] See McCarthy (2005).
[2] See Yule, G. U. (1944)

style of a children's book", prompting for low lexical richness, or "*Write a poem about apples* ***using advanced vocabulary***" to prompt for high lexical diversity.



After each prompt, a new conversation was started in each model, to prevent output bias based on the previous input. Clearing chat history with an LLM removes its history, allowing for unbiased or altered outputs from the LLMs.

Additionally, each prompt was tested twice for every model. Because typing the same prompt into a model generates different results each time, inputting the same prompt twice into a model exposes outliers in the lexical diversity of the outputs. The two prompt's calculations were averaged in the data collected.

4 Experiments

4.1 Discussion

Based on Charts 1, 2, and 3, it is clear that what is written in the prompt to the LLM influences the lexical diversity of the output text. While the TTR value varies slightly, the Yule's I and MTLD quantities of lexical diversity vary greatly depending on whether the model was prompted for high or low lexical diversity. From this observation, a conclusion can be made that the prompt for the LLM matters more than the model itself. However, findings about the models can still be taken away from this data. In chart 2, Claude AI represented in the green bars has the highest Yule's I values from the data out of all the other models in the high prompting and baseline tests. Claude is also strong in some of the low LD prompts. Being a lesser-known AI model with around 150000 words in training data from August 2023[3], the statistics are impressive for Claude. Compared to GPT-4's 300 billion word training corpus[4], Claude's lexical diversity measures are impressive with a fraction of GPT-4's training data.

Based on Chart __: a conclusion can be drawn that lexical diversity doesn't change when an LLM translates text. Whether it's a poem or information article, the lexical diversity measures of

---

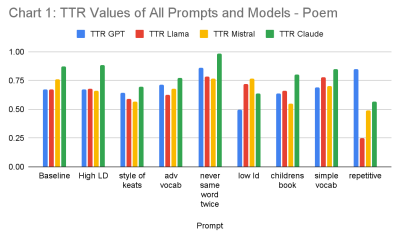
3 https://www.notta.ai/en/blog/claude-statistics#:~:text=Claude%20AI%20is%20trained%20on,150%2C000%20words%20or%20500%20pages.

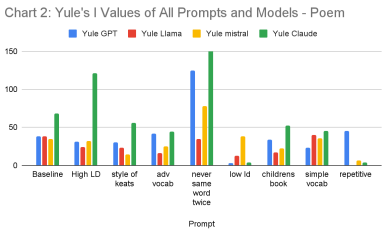
4 https://www.businessinsider.com/google-researchers-openai-GPT-4-to-reveal-its-training-data-study-2023-12#:~:text=The%20AI%20model%20powering%20GPT-4,or%20570%20GB%2C%20of%20data.

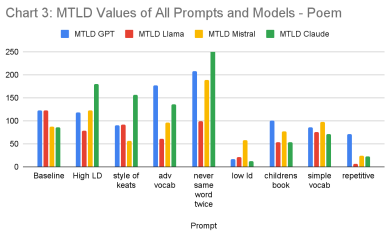
TTR, Yule's I, and MTLD don't vary greatly when translated across languages. This is a development from old machine learning models and a promising sign that generative AI models are more linguistically capable than old technology. Describe design experiments

4.2 Results

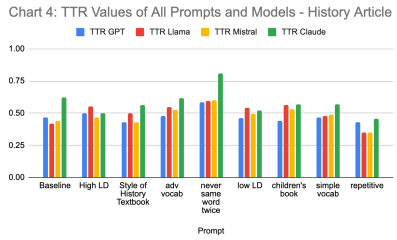
Prompt: *Write a Poem about Apples.*

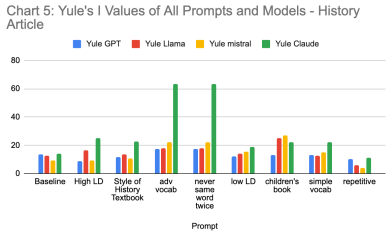


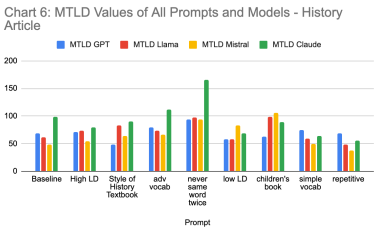




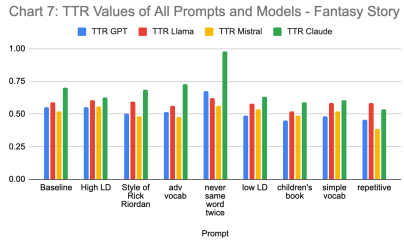
Prompt: *Write an informational article on Hungarian History.*

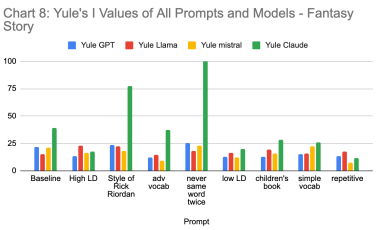


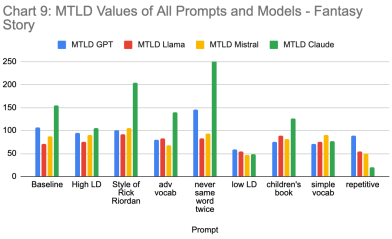




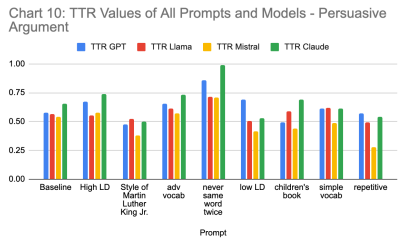
Prompt: *Write a fantasy story surrounding Greek Mythology.*

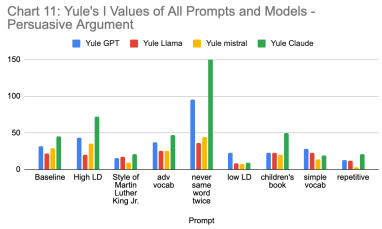


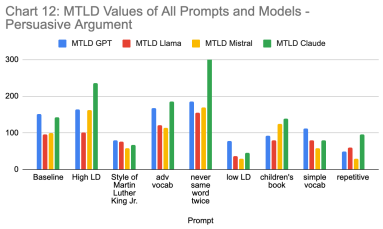




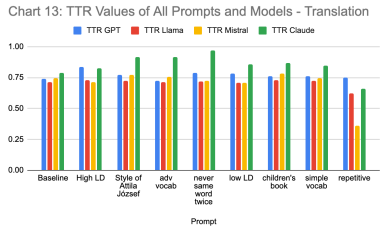
Prompt: *Write a persuasive article about eating healthy.*

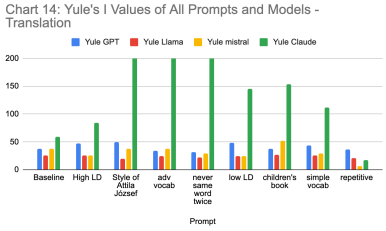


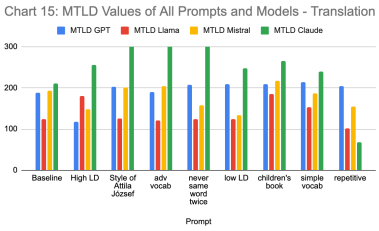




Translation Data “Translate this poem to Hungarian: [poem]”:







4.3 Outliers

The largest outlier produced was in Table 1 Claude’s generation of a Poem about Apples without using the same word twice. Yule’s I and MTLD values are inflated above 1000, whereas Yule’s I and MTLD generally struggle to retain values above 100. Claude’s generation with this prompt resulted in these inflated numbers because of MTLD’s formula and how Claude reacted to the prompt. MTLD is calculated by looking at the TTR value of a piece of text, and when it goes above or dips below a threshold (0.72, in the lexical richness Python library), the MTLD value gets calculated. Since Claude follows the rule never to repeat the same word, the TTR value never drops below the threshold of 0.72, inflating the MTLD value. One notable observation from the “Never Repeating a word twice” prompt is that all of the values in that row are significantly high. Since Claude’s values are the highest within this row, it can be inferred that Claude followed the rule more than the other LLMs.

5 Conclusion

Before generative AI models existed, traditional ML (Machine Learning) was used to translate text. Specifically, those neural networks weren’t effective in translating between languages. This study shows how generative AI models surpass the quality of traditional ML. The lexical diversity stays consistent across translations. Additionally, generative AI models produce high quality pieces of text when prompted, where the prompt matters more than the model used. Models are being produced at an alarming rate, and one challenge is deciding which model produces higher quality outputs. This study provides one perspective on how to differentiate models. For example, Claude takes prompts very directly, whereas ChatGPT and Meta take in the bigger picture, loosely following specific guidelines. Future work on this topic could be done to test more prompts, or text other means of linguistic quality beside lexical diversity. Models

can also be analyzed outside of Linguistics, through a mathematical standpoint or any quantifiable measure. As AI models continue to evolve at a rapid pace, this study lays out a foundation for further exploration comparing generative AI.

6 Acknowledgements

Thank you to Professor Janet Randall (j.randall@northeastern.edu), and Professor Tamás Biro (biro.tamas@btk.elte.hu) for their meaningful contributions to this paper.

7 References

[1] G. Udnv Yule. 1944. The Statistical Study of Literary Vocabulary. Cambridge University Press.

[2] Philip M McCarthy. 2005. An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD). In PhD Thesis, Dissertation Abstracts International, Volume 66:12. University of Memphis, Memphis, Tennessee, USA.

[3] Mildred C. Templin. 1975. Certain Language Skills in Children: Their Development and Interrelationships. Greenwood Press, Westport, Connecticut, USA.

[4] https://arxiv.org/html/2402.07927v1

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11036183/

[6] https://www.sciencedirect.com/science/article/pii/S2949719123000213

[7]https://arxiv.org/pdf/2305.00948