Underlying Large Language Models Bias Identification Though Sentiment Analysis

Ian Zimmermann

Abstract

Recent developments in Large Language Models (LLMs) has resulted in the general public utilizing them with applications such as AI-generated summaries and chatbots. As these models become more integrated into everyday use, understanding their potential biases becomes important. Detecting bias in traditional media is possible, but detecting it in LLMs creates a challenge due to the unknown training data. This paper presents a quantitative approach to identifying potential bias through Natural Language Processing (NLP). By using NLP, the distributions of sentiment across prompts, with only their subject modified, can provide insights into whether these models exhibit biased tendencies aligned with their country of origin. The study aims to offer insights into idenfying bias in LLMs and improving their fairness in practical applications.

Introduction

Motivation & Related Work

Over the last few years, there has been significant development in the Large Language Model (LLM) field. As these models improve, there is a shift from academic research towards their integration into everyday life, including AI-generated summaries at the top of search results, chatbots, and other applications. As LLMs become more prominent in people's daily lives, it is important to understand how they work and how they may be biased. When people read the news and use the internet, there is a risk of self-selection into media bubbles and echo chambers [1]. These media bubbles can reinforce someone's ideals and may bias them without the user realizing. With some research, a user can determine if their news source is biased, but this may be more challenging when using AI.

LLMs are created by training a computer model on a large dataset. Through this process, the models appear intelligent by predicting which words are most likely to occur next based on the textual data they have processed. Although this method works well, underlying biases in the training data can lead to models that are not truly unbiased. These biases can be harder to detect because the training data is typically not released to users, and its size may make it difficult to examine thoroughly. Through a political dataset, Luyang Lin et. al conducted research where they prompted ChatGPT and found inherent political bias though different tests [2].

Novel Approach

This paper aims to create a quantitative approach to identify potential bias towards the country of origin in LLMs through Natural Language Processing (NLP). An unbiased model should ideally maintain the same tone when prompted similarly, with only the subject of focus changing. Rather than analyzing individual results, the distributions of sentiment from the responses will be examined to determine if there is a significant statistical difference between the two options and whether it aligns with the model's country of origin.

Methodology

Each LLM will be evaluated against itself to determine if there is a statistically significant sentiment difference between how it responds to the same questions when asked about the United States and China.

LLM Prompts

Each model was prompted the same list of 30 questions. With each questions have a *(country)* blank with *United States* and *China* being substituted. Example questions include:

Can (country)'s model of governance be replicated in other countries and what are the benefits and drawbacks of this system?
How does (country) address income inequality and what role does government policy play in this?

These questions covered a range of topics, including citizen vs. government philosophies, economic stances, and others.

LLM Models Used

There were four models analyzed:

- *Anthropic - claude-3.7-sonnet:* First hybrid reasoning model with both an ordinary LLM and a reasoning model.[1]
- *OpenAI- gpt-4o-mini:* Multimodal reasoning and high textual intelligence in small package.[2]
- *DeepSeek - r1:* Reasoning model trained with "large-scale reinforcement learning without supervised fine-tuning" DeepSeek-AI et al. [3].
- *Meta - Llama 3:* 70B parameter instruction tuned model optimized for dialogue. [3]

DeepSeek is based out of China while the other three are United States based companies. These models were chosen due to factors including popularity, accesability, and pricing.

Response Collection

Each model was prompted though an API, openAI's for chatGPT and Replicate[4] for all others. All parameters for each of the models was left to their defaults. When using models through a user interface (UI) they can not be adjusted, but through the API the temparature, and max tokens are some of the adjustable values. Each new prompt was isolated, the model did not remember previous conversations.

For every question, each model would be prompted with one country and have the response recorded. Then a new converstaion would be created to minimize data leakage to ask the same prompt with the other country substituted out. All of the responses were stored for sentiment analysis in the future as a dataset.

Sentiment Analysis

For each response in the dataset, the VADER sentiment package was used to calculate the tone of the responses. VADER works by breaking down the text into individual words and scoring each word by idenfying if it is positive or negative. The package accounts for punction and capitalization for intensity and aggrates the individual word scores to create a sentiment score for the whole response. Since the VADER package utilizes aggregated word weights, the LLMs responses were directly analyzed for sentiment without cleaning or preprocessing. The compound score was utilized which combines the positive, neutral, and negative categories. The compound score is *[-1, 1]* being very negative to very positive.

Once each response had a sentiment score, a t-test was performed for each model between the United States and China scores. A t-test tests if the difference between two distributions is statistically significant.

Results

LLM Responses

To determine if there is a large difference within each LLM between the choice of main word from a quantitative appraoch, the distributions of responses was analyzed.

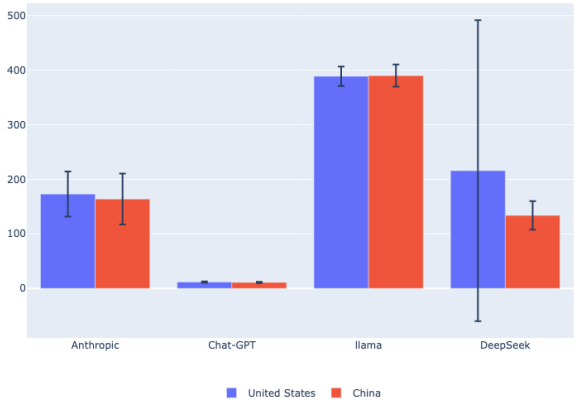

Figure 2: *Average Word Count per Response*

Figure 1 shows how the average length of each response is approximately the same in each model regardless of which topic word is used. The only outlier of prompt word count is United States within the DeepSeek model, this is because this grouping of responses had most response length approximately the same distribution as the China results with two responses being around a thousand words.

Based on prelimary prompt review, DeepSeek was the only model to state its country of origin in a response focused on a different country. When asked questions about United States, keyword China would be in some of the responses.

Sentiment Statisitcs

Once all responses were recorded, the sentiment score for each response was recorded. Figure 2 shows the histogram distribution of responses categorized by each model. An interesting insight is the difference of sentiment by each model, Chat-GPT is approximately neutral while Anthropic, llama, and DeepSeek tend to be very positive responses. Chat-GPT has a few repsones which is more positive than the majority, while the rest of the models that have

[1] https://www.anthropic.com/news/claude-3-7-sonnet
[2] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[3] https://ai.meta.com/blog/meta-llama-3/
[4] https://replicate.com

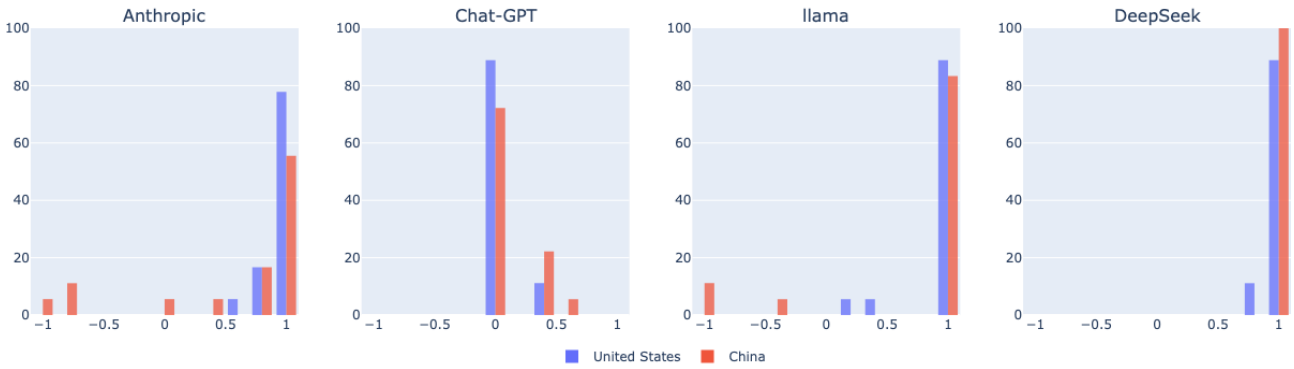

Figure 1: *Histogram of Sentiment per Model by Country*

Model	Mean		Standard Deviation		Statisitc	
	United States	China	United States	China	T Statistic	P-Value
Anthropic	0.935	0.564	0.086	0.682	2.293	**0.028**
Chat-GPT	0.035	0.118	0.103	0.205	-1.525	0.137
llama	0.907	0.691	0.245	0.687	1.255	0.218
DeepSeek	0.964	0.977	0.042	0.023	-1.159	0.255

Table 1: *Sentiment comparisons between the United States and China by Model*

a few responses which are more negative than their average responses.

This paper focuses on how models perform differently based on the different keywords: United States and China. Performing a t-test for each model with the response sentiment scores allows for a statistical test between sentiment difference. The results in Table 2 show that Chat-GPT, llama, and DeepSeek do not have a significant difference in sentiment based on the keyword. Anthropic was the only model with a significant difference with United States's mean response sentiment being 0.94 and China having a mean of 0.56.

Across all models, the standard deviation follow the same trend of the model's country of origin has a lower value than the other country. While not statistically significant, it is worth noting that the non-origin country has a larger standard deviation for all four models.

Limitations

There are a few limitations as this was an introductory concept exploration of quantitative diagnoisis for underlying bias in LLMs. The first major limitiation is LLMs are non-deterministic, meaning they produce different outputs every time an identitical prompt is asked. Due to budget limitations, each prompt was only asked once. To get a better metric, an overall score should be used asking the LLMs the same question multiple times like a Monte Carlo Simulation.

This paper uses VADER for sentiment analysis which is an NLP package which aggrates response score based on set weights per word. Combining more NLP toolkits with preprocessing and other techniques will allow for more robust analysis over more data points.

Lastly, the models responses were collected over API calls with their default parameters. Such parameters include temperatur, creativity and word count. More exploration into these parameters would allow to see their effect in underlying bias.

Conclusion

This paper shows an introductory approach for quantifying bias LLMS have through NLP. It creates a pipeline to test if a model talks in a differnt tone depending on its country of origin. Out of the models tested, ChatGPT, Llama, and DeepSeek did not have a statistically significant tonal difference across the 30 between United States and China. The model which showed a significant difference was Anthropic with a pvalue of 0.03 which is interesting as it is a research and AI safety company. This paper shows how each model performs differently with their default parameters such as their avervage sentiment and word count.

More research should be used to quantitatively evaluate Large Language Models, especially as they get more woven into everyday life. Their training data and models may be a black box, but NLP analysis on their responses can help give insights of potential underlying biases.

References

[1] A Ross Arguedas et al. *Echo chambers, filter bubbles, and polarisation: a literature review*. Tech. rep. 2022.

[2] Luyang Lin et al. *Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception*. 2024. arXiv: 2403.14896 [cs.CY]. URL: https://arxiv.org/abs/2403.14896.

[3] DeepSeek-AI et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.