How you can convince ChatGPT the world is flat

Stanford CS224N Custom study

Julian Cheng
Department of Statistics
Stanford University
jcheng2k@stanford.edu

Abstract

Question Answering (QA) systems are capable of generating answers based on training data and user prompting. It stands that the accuracy and reliability of these systems are paramount. Often taking the form of Large Language Models (LLMs), they are tasked with generating responses from a diverse array of sources. However, the presence of conflicting information within these sources presents a significant challenge, potentially undermining the credibility and effectiveness of the answers provided. This study aimed to critically assess how LLM QA retrieval models prioritize conflicting data. Through a series of linguistic characteristics, and prompt variations, we investigated the decision-making criteria these models employ when presented with conflicting information.

1 Key Information

- Mentor: Nelson Liu

2 Introduction

To answer a subjective question by choosing between two opposing perspectives, humans engage in critical thinking, attempting to evaluate the credibility and relevance of their knowledge. This process is subjective, influenced by personal experiences, values, and the perceived trustworthiness of the information sources. There may also be an emotional response, as feelings and intuitions can sway judgments and lead to preference for one perspective over another based on emotional connection in place of empirical evidence. Social influences such as societal norms and cultural values further complicate the decision and can pressure individuals to align with particular viewpoints, sometimes at the expense of personal beliefs or objective facts.

For LLMs, this process is both simpler and difficulter. Let's ask ChatGPT for example, whether "difficulter" is actually an English word, and using no information aside from what is provided with two opposing perspectives,"difficulter" is a word, and "difficulter" isn't a word. The model has no social or emotional biases, history, or external information to draw upon, and bases its perspective entirely from the contents of the two opposing documents. This study benchmark characteristics of opposing documents, such as unique tokens, text perplexity, ease of understanding, and n-gram count, in order to begin to understanding the decision-making process used by LLMS.

In order to assess a variety of model responses, various prompting techniques and applying minor text perturbations are used. Variations in prompting styles and the examples provided likely influence the outcome of the prompt, and are necessary in order to assess a holistic representation of the decision-making process.

3 Related Work

Previous research regarding the resolution of conflicting data in LLMs has highlighted particular linguistic characteristics that affects the response.

Previously developed, the ConflictingQA, a dataset consisting of questions and real web documents that lead to conflicting answers. While pitting documents of opposing perspectives against one another, the study performs sensitivity and counterfactual analyses to find features that correlate with document convincingness Wan et al. (2024). It considers a combination of features that describe both linguistic properties of a document and the document relevance to the question. Many of these were inspired by results from studies of human credibility. One example of this would be the use of including academic or scientific documents, appealing to the authority of the source in order to influence humans but also possibly the QA model. Previous studies benchmarked linguistic characteristics of each document, but found that adding text perturbations to the documents was most effective in changing the likelihood of a document being agreed with Hu et al. (2024).

Building off of the aforementioned foundations, this study continues to explore the effects of perturbations, especially negative perturbations, which were not explored, in prompting that may have an effect on the outcomes of LLM decision-making between contentious perspectives, combining changes in prompting styles with new linguistic characterizations.

4 Approach

Inspired by other wors, this study compiles and organizes documents from the ConflictingQA dataset, which will be detailed in the next section. After preprocessing of the documents, each document is categorized under the query that it provides a response for as well as whether it has supports a "yes" or "no" stance when answering the query. To generate prompts with conflicting perspectives, within each contentious query, every possible combination of an affirmative document and negative document is given to the LLM as an input. The model is asked to respond with a definitive one word answer: "yes" or "no", using only the information in the paragraphs. A zero-shot prompt, with no prior examples, is considered as the baseline prompt model.

> Using only the information from the perspectives and no outside information, answer yes or no in a single word, to the question.
> **Question:** Can testosterone increase the risk of prostate cancer?
> **Document 1:** This worldwide database collates information from all prospective studies of hormonal factors and prostate cancer risk, and contains over 17,000 prostate cancer cases with measured hormone levels (including 2,300 aggressive cases) and 37,000 controls...
> **Document 2:** New research presented this weekend at the National Cancer Research Institute (NCRI) Cancer Conference in Liverpool has concluded that men with naturally low levels of the male sex hormone testosterone are less likely to develop prostate cancer than those with higher blood levels of the hormone...
> **Answer:**

Figure 1: Example baseline zero-shot prompt composition

Few-shot prompts are provided examples of queries and perspectives, drawn from the same dataset, and single-word answers. In the experimental cases using perturbations, a brief phrase is added to the start of a paragraph, stating the following: "This document answers {yes/no} to the question". The purpose of this perturbation is to avoid any changes to the contents of each perspective, but instead to explore the effects of subtle manipulations that affect the outcome of the LLM output.

Within the results of each prompting style, LLM responses are then categorized based on response and several linguistic characteristics such as lexical richness, and readability are compared between the perspectives that either won (the LLM agreed with its stance), or lost. This study measures a document's likelihood to be selected within a query against opposing documents, known as its Winrate Wan et al. (2024).

While some approaches used in this study were inspired by previou work, unless explicitly stated, approaches are original. This study uses the RAG-convincinness dataset, but all written code for preprocessing, LLM testing, evaluation, and analysis is my own.

5 Experiments

5.1 Data

ConflictingQA is a dataset consisting of questions and documents that lead to conflicting answers. Question categories are generated across a range of subjects, including Neuroscience, Religion, Etiquette, Metaphysics, and Real Estate. Questions themselves are unambiguous, answerable with binary "yes/no", but facilitate conflicting responses from a variety of sources (i.e. "Are knee braces effective in preventing knee injuries?"). Evidence is webscraped in the form of Google Search results through the Google Search API, obtained by querying either the affirmative or negative format of the question. From each of the top suggested documents, raw text is extracted and further adjusted by choosing the 512 token window with the greatest token similarity to the query.

5.2 Evaluation method

In order to characterize linguistic features of documents, previous research analyzed the number of unique tokens, n-gram overlap with the question, or question embedding similarity. In addition to these, this study analyzes the perplexity, Flesch-Readability score and rating, and lexical richness. The implementation of each is as follows:

- **Number of Unique Tokens:** The number of unique tokens measures the complexity and variety of the document's vocabulary. The token count is calculated using the tiktoken library Muñoz-Ortiz et al. (2023) .
- **Question N-gram Overlap:** Significant overlap of n-grams between the question and the response often indicates that the response is closely related to the question which can lead to the model choosing that perspective. Ngram similarity is calculated from n ranging from one to four Jurafsky and Martin (2023).
- **Perplexity:** The perplexity of a document is how surprised the model is to see new data, a test of the quality of the training process. Perplexity is calculated as the exponent of mean of log likelihood of all the words in an input document, with a lower value indicating greater familiarity, using CausalLM from GPT2 McFarlane et al. (2009).
- **Flesch-Readability Characteristics:** Use the py-Readability Library allows implementation the Flesch-Kincaid Reading Ease test as a standard for readability of a document. The U.S. Department of Defense uses the Reading Ease test as the standard test of readability Solnyshkina et al. (2017).
- **Lexical Richness:** Using the lexicalrichness library, this characteristic measures textual lexical diversity, computed as the mean length of sequential words in a text that maintains a minimum threshold type-token ratio (TTR) score Kojima and Yamashita (2014).

The Winrate of each document is also measured under the conditions of each prompt. Winrate is measured as the empirical probability of the model's prediction aligning with its stance when paired with a set of conflicting (opposite stance) documents.

$$WR(p_{\text{yes}}, q) = \mathbb{E}_{P_p, P_{q,no}} \left[\mathbb{1} \left[f(p_{\text{yes}}, p, q) = \text{yes} \right] \right]$$

Figure 2: Winrate formula

5.3 Experimental details

The LLM in this study is the GPT-3.5-turbo model run with the OpenAI API, with prompts as described before. No changes were made to the model when prompting, and no prior information from previous runs was retained for each new document comparison or prompt style.

There are five variations of prompts used in this study:

- **Zero-shot:** The baseline prompt as shown above. This prompt is intended to assess a preliminary unperturbed comparison between two opposing document stances.
- **Zero-shot ’Yes’ Perturbation:** "This document answers **yes** to the question." is added before each document that has the ’yes’ stance. The purpose of this perturbation is to avoid any changes to the contents perspectives, but elicit possible changes to the decision-making process of the LLM.
- **Zero-shot ’No’ Perturbation:** "This document answers **no** to the question." is added before each document that has the ’no’ stance. Just as in the other perturbation, this prompt has an identical intended purpose.
- **One-shot:** In the One-shot prompt, the LLM is provided an example queries and two associated opposing perspectives, drawn from the same dataset, and a single-word answer before the query and perspectives the LLM is intended to answer.
- **Two-shot:** Similar to the One-shot prompt style, but there are two example queries with accompanying perspectives and answers.

5.4 Results

GPT-3.5-Turbo LLM Perspective Alignment			
Prompt Style	Yes	No	Unsure
Zero-shot	4994	3097	84
Zero-shot ’Yes’ Perturbation	3905	4261	9
Zero-shot ’No’ Perturbation	1691	6479	5
One-shot	2949	5217	9
Two-shot	4318	3842	15

Figure 3: LLM perspective alignment frequency for all prompt variations

Regarding the zero-shot prompt as the baseline, it is evident that each variation of the prompt affects the LLM’s decision-making process. In both cases of minor perturbations, the perturbation changes work to the favor of increasing the LLM’s alignment with the "no" perspective, which is a surprising result given that one prompt variation specifically only makes changes to the "yes" perspective. Under the conditions of the zero-shot ’No’ Perturbation, nearly 80% of responses being "no" when the contents of the prompt (the query and perspectives) are otherwise unchanged. The outcomes under few-shot prompting are inconclusive, as the one-shot prompt style heavily increases the amount of "no"-aligned decisions. It is observed that is a lower amount of "Unsure" decisions in variations of the prompt styles.

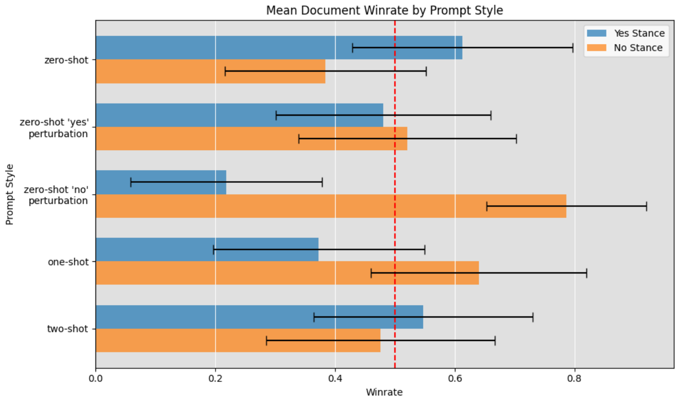


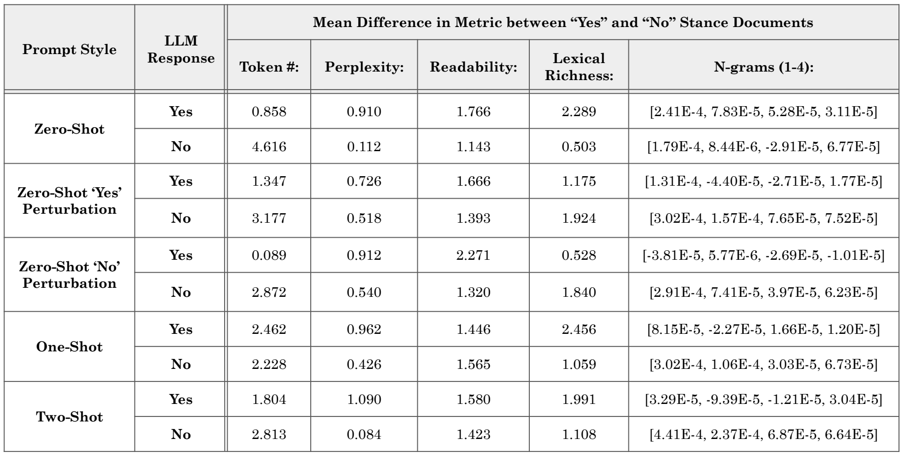
Figure 4: Average document Winrate across prompt variations, grouped by LLM response

The average Winrate, grouped by prompt styles and LLM response, highlights the individual variation under each prompt. The mean winrate is reflective of the overall frequencies reported in Figure 3.

Variance remains similar across prompt styles and responses, likely owning to unchanged parameters of the GPT-3.5-turbo model.

6 Analysis

In order to understand the differences across prompt variations and responses, it is worthwhile to capture empirical features of documents. Under each prompting style, the "yes" and "no" categories of LLM responses are grouped, and the linguistic characteristics are calculated for each document. To compare opposing perspectives under a query, this study calculates the difference for each linguistic metric for every combination of opposing documents.



Prompt Style	LLM Response	Mean Difference in Metric between "Yes" and "No" Stance Documents				
		Token #:	Perplexity:	Readability:	Lexical Richness:	N-grams (1-4):
Zero-Shot	Yes	0.858	0.910	1.766	2.289	[2.41E-4, 7.83E-5, 5.28E-5, 3.11E-5]
	No	4.616	0.112	1.143	0.503	[1.79E-4, 8.44E-6, -2.91E-5, 6.77E-5]
Zero-Shot 'Yes' Perturbation	Yes	1.347	0.726	1.666	1.175	[1.31E-4, -4.40E-5, -2.71E-5, 1.77E-5]
	No	3.177	0.518	1.393	1.924	[3.02E-4, 1.57E-4, 7.65E-5, 7.52E-5]
Zero-Shot 'No' Perturbation	Yes	0.089	0.912	2.271	0.528	[-3.81E-5, 5.77E-6, -2.69E-5, -1.01E-5]
	No	2.872	0.540	1.320	1.840	[2.91E-4, 7.41E-5, 3.97E-5, 6.23E-5]
One-Shot	Yes	2.462	0.962	1.446	2.456	[8.15E-5, -2.27E-5, 1.66E-5, 1.20E-5]
	No	2.228	0.426	1.565	1.059	[3.02E-4, 1.06E-4, 3.03E-5, 6.73E-5]
Two-Shot	Yes	1.804	1.090	1.580	1.991	[3.29E-5, -9.39E-5, -1.21E-5, 3.04E-5]
	No	2.813	0.084	1.423	1.108	[4.41E-4, 2.37E-4, 6.87E-5, 6.64E-5]

Figure 5: Mean linguistic characteristic difference across all prompt variations grouped by LLM response

Across all prompt variations, GPT-3.5-Turbo would generally respond "No" for a relatively greater magnitude of difference in unique token count and perplexity between opposing documents. This observation could be explained as the model answering "no" when the "yes" stance document has notably more tokens than its counterpart. The other case to consider is that the model selects "yes" when the perplexity difference is greater (the model is less familiar with the "yes" stance document than the "no" stance), which could be attributed to a model making this decision when it has lower familiarity with the "yes" stance document.

There is less of a consistent observable difference in the Flesch-Kincaid readability and Lexical Richness metrics. The mean errors of the n-gram overlap with the query were too small to be appreciably different between "yes" and "no" LLM responses.

Comparing the differences for unique token count and readability, between different prompting styles, changing from zero-shot to few-shot prompts reduces the magnitude of differences between LLM response categories. Noticeably, the prompts using perturbations causes the readability difference when the LLM takes a "yes" stance to decrease drastically from the unperturbed zero-shot prompt, while the opposite could similarly be said for the case of the "no" stance. While this could provide explanation for the higher winrate and frequency of the latter stance in the perturbed prompts, the one-shot prompt style has a large difference under the LLM's "yes" stance case and vice versa for the "no" stance.

There were 2072 instances out of 8175 prompts where the model responded with the same answer across all five prompt variations. The salient linguistic metrics are reported in Figure 6 for these consistent cases. Notably, as before, "no" responses from the LLM occur in instances where the the the "yes" stanced document has on average nearly 5 unique tokens than the "no" document, while the in cases of higher (above two) mean Flesch-Kincaid readability and Lexical Richness differences between documents, the LLM tends to consistently decide yes. It is important to note however, that these metric values are not substantially different from those observed above for each category, hinting at possibly yet another undiscovered feature that could explain the consistency of the LLM's decision-making when presented with pairs of certain prompts.

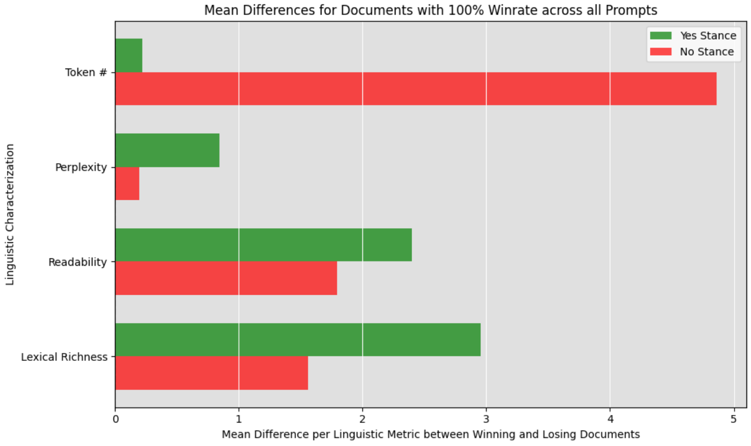


Figure 6: Linguistic characteristic differences for documents with 100% Winrate

As a broad trend within the ConflictingQA dataset, documents with the "no" stance scored higher in each metric than their "yes" stance counterparts, having a higher average number of unique tokens, greater perplexity (the model is less familiar with the document contents) yet greater Flesch-Kincaid readability, and lexical richness.

7 Conclusion

This study endeavored to benchmark the decision-making process of LLMs when faced with conflicting information. To do so, a number of approaches were leveraged, from evaluating linguistic features of documents and the difference in feature value against opposing documents, and manipulating the prompt styles and contents. As evident in prior research, language characteristics often provide inconclusive information on the predictability of an LLM's response, although token difference between documents is generally greater in magnitude when the model answers "no". Perturbations, on the other hand, demonstrated great efficacy in manipulating the response of the LLM, even when the perturbation itself did not change any of the contents of either document. Especially prevalent in the case of the zero-shot "no" perturbation, it was found that the winrate of a document taking the "no" stance would increase substantially in comparison to the baseline zero-shot and few-shot prompt approaches.

In this study, there were no repeat trials of any prompt style, owing to the computational costs of using the OpenAI API, which could have impacted the summary statistics of the decision-making process. After calculating the linguistic features of the models, it was found that the documents in the dataset ConflictingQA that took the "no" stance against the provided queries on average scored higher in every linguistic metric that was analyzed in this study. For future work, it would be worthwhile to find documents that are more uniform in these characteristics for both stances.

Additional studies could investigate the results of further perturbations of the opposing documents. Cases where both documents are given the same perturbation, or few-shot examples that leverage the aforementioned manipulations could likely have observable effects on the winrates of individual documents and overall trends of decision distributions.

References

Zhibo Hu, Chen Wang, Yanfeng Shu, Helen Paik, and Liming Zhu. 2024. Prompt perturbation in retrieval-augmented generation based large language models. arXiv.org.

Daniel Jurafsky and James Martin. 2023. N-gram language models. In *Speech and Language Processing*. Stanford University.

Masumi Kojima and Junko Yamashita. 2014. Reliability of lexical richness measures based on word lists in short second language productions. Elsevier.

Delano McFarlane, Noémie Elhadad, and Rita Kukafka. 2009. Perplexity analysis of obesity news coverage. In *AMIA Annual Symposium Proceedings Archive*. https://www.ncbi.nlm.nih.gov/.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated text. arXiv.org.

Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov. 2017. Evaluating text complexity and flesch-kincaid grade level. In *Journal of Social Studies Education Research*. www.jsser.org.

Alexander Wan, Eric Wallace, and Dan Klein. 2024. What evidence do language models find convincing? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. arXiv.org.