Introduction

Humans have been making decisions for as long as they can remember. With the invention of AI and increasingly powerful models, recommendation systems have become stronger, but they still replicate errors.Therefore, relying on AI to accurately assess and answer questions could produce harmful results. Prior research has shown that humans often overestimate the reliability of LLM responses, especially when answers are framed with high certainty or longer explanations. This miscalibration can cause harmful consequences when users rely on incorrect outputs in high-stakes domains such as health, law, or politics.(Steyvers et al., 2025)

Our study critically examines how LLMs reproduce false beliefs and how the format and framing of their answers affect truthfulness. By using the TruthfulQA dataset, which contains questions based on human misconceptions along with correct and incorrect answers, we can study systematic patterns of accuracy and error. We will also consider the broader context of user trust in AI, drawing from recent studies on trust and reliance on AI-generated explanations (e.g. Steyvers 2024). This situates our project within critiques of generative AI, particularly its tendency to reinforce persuasive but false information.In addition to descriptive analysis, we will train classifiers on linguistic features such as length, certainty markers, hedging words, and pronoun use. Even if these features do not reliably predict truthfulness, such findings would be valuable: they would demonstrate that the surface cues humans often rely on provide little evidence about factual accuracy.

Method

We will use the TruthfulQA dataset (split by generation and multiple_choice). Each entry contains a question, a “best” LLM-generated answer, a set of correct answers, and a set of incorrect answers. Many questions are adversarial or based on common misconceptions (e.g., “Do veins appear blue?”). This structure allows us to quantitatively analyze truthfulness rates across categories and compare language features between correct and incorrect answers.

We may also incorporate supporting insights from the Anthropic HH-RLHF dataset, which contains human-labeled judgments of helpful vs. misleading responses, to connect with the literature on human trust and preference in LLM explanations.

Research Questions

1. How does question category (e.g., science, politics, commonsense, health) affect the likelihood that LLMs provide truthful vs. false answers?
2. What linguistic or structural features (e.g., explanation length, hedging words, certainty markers) appear more frequently in correct vs. incorrect answers, and are these features systematically related to truthfulness?

3. Can these linguistic features be used in a classifier to predict correctness? If the classifier performs poorly, what does this reveal about the disconnect between surface-level cues and actual factual accuracy?

Hypothesis

Based on previous studies, we hypothesize that there may be systematic differences in how correct and incorrect information are presented, but we also acknowledge that such differences may not always emerge.Steyvers et al. (2025) found that people often overestimate the accuracy of model responses, especially when the answers are written with high confidence or are longer in form, even if this additional length does not improve correctness. This suggests that incorrect answers in our dataset coulduse persuasive structures such as confident phrasing or longer explanations, potentially making them appear more trustworthy even when they are factually incorrect. Other research on trust and reliance has shown that people tend to judge the reliability of responses using surface-level cues such as length or certainty rather than factual accuracy. Based on this evidence, we will test whether response length and certainty markers will correlate with correctness. Even if they do not, this result would still be important, showing that the cues people rely on provide little indication of factual accuracy. For multiple-choice questions, we expect that accuracy will decrease as the number of answer options increases, since larger choice sets create more opportunities for plausible but incorrect alternatives. To examine these expectations, we will combine with descriptive statistics of truthfulness across categories, with textual feature analysis (length, hedging, certainty language), and we will also train classifiers such as logistic regression or random forest models to test whether these features can predict correctness. Through this analysis, our study will build on existing findings about human miscalibration, either by identifying linguistic patterns associated with truthfulness or by showing that such patterns are unreliable indicators, both of which have implications for evaluating and designing reliable language models.

Reference

• Klingbeil, L., & Schuppler, B. (2024). Trust and reliance on AI: When and why people over-rely on machine-generated advice. Nature Machine Intelligence, 6, 707–716. https://doi.org/10.1038/s42256-024-00976-7
• Steyvers, M., & Lee, M. D. (2025). What large language models know and what people think they know. Nature Human Behaviour, 9, 15–27. https://doi.org/10.1038/s41562-024-01995-8
• Lin, S., Hilton, J., Evans, O., & Bowman, S. R. (2021). TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958. https://arxiv.org/abs/2109.07958

Group Contract and Work Agreement

We expect each other to work our hardest on the project, spending the time we have on it while making time to work on it in person with each other. As far as when we will work on what, the times will fluctuate based on what part of the project we are immersed in. For example if we're closer to the beginning of the project and it's just about creating a project proposal, we would meet at the library 2 to 3 times in order to get an understanding of what we want to accomplish while we're together and while we're separated, so that when we reconvene, we've made significant progress and can discuss all the changes that took place. By splitting up the work evenly (relatively), we can maximize efficiency while minimizing stress.