Introduction and Data

Large language models(LLMs) have rapidly become integrated into everyday decision-making, from answering factual questions to assisting with health, legal, and financial reasoning. Despite their impressive fluency and versatility, LLMs are known to generate statements that sound confident yet are factually incorrect. This often raises significant concerns about how users interpret and rely on model outputs. Prior research shows that people frequently overestimate the accuracy of LLM responses, especially when answers are written with high certainty, contain longer explanations, or resemble expert-like reasoning. As a result, users may form misplaced trust in incorrect model outputs, particularly in high-stakes situations such as medicine or public policy.

Our project is motivated by this broader problem: humans often judge credibility based on surface-level linguistic cues, while LLMs can generate persuasive but false answers. To investigate this dynamic, we analyze the TruthfulQA dataset, created by Stephanie Lin, Jacob Hilton, and Owain Evans, which explicitly tests models on questions based on common misconceptions or misinformation. Because each question contains both correct and incorrect answers, the dataset allows us to examine how truthful and untruthful responses differ in their language patterns.

We focus on evaluating whether simple linguistic features—such as answer length, hedging words[1], certainty markers[2]—are systematically associated with truthfulness. These are features that humans commonly rely on when interpreting reliability, even though they may have little connection to factual accuracy. If these features fail to predict correctness, this would reinforce the growing argument that human intuition about "trustworthy language" is poorly aligned with actual truthfulness in LLM-generated content.

Methods

Analytical Framework

We employed a mixed-methods computational approach using natural language processing and statistical visualization to examine relationships between linguistic features and factual accuracy in AI responses. Our analysis utilized two complementary data structures: (1) a question-level dataset for aggregate pattern analysis, and (2) an answer-level dataset for granular linguistic examination. This dual-structure design

[1] Words of uncertainty, for example: maybe, possibly, tend, etc.
[2] Words of certainty, for example: absolutely, definitely, clearly, etc.

enabled both macro-level performance assessment and micro-level response characterization.

Data Processing and Cleaning

We used the TruthfulQA generation dataset (817 validation questions) loaded via Hugging Face's datasets library. Initial inspection revealed that correct_answers and incorrect_answers were stored as string literals rather than Python lists. We converted these to native lists using ast.literal_eval with a regex fallback, enabling accurate list operations and membership testing.

No rows or columns were removed during cleaning, as all observations contained valid best_answer entries and non-empty answer lists. This complete dataset preservation ensured maximum statistical power while maintaining data integrity. The cleaned question-level dataset served as the foundation for all subsequent analyses.

Data Restructuring

To enable both aggregate and granular analyses, we transformed the data from question-level to answer-level format. The original structure (one row per question with nested answer lists) was programmatically expanded to create a long-format dataset where each row represents a single answer with a binary correctness variable (1 = correct, 0 = incorrect). This restructuring was essential for directly comparing linguistic features between correct and incorrect responses without aggregation bias.

Feature Engineering

We initially constructed a question-level correctness variable (ai_correct), but this measure proved unsuitable because the model's best answers matched the dataset's correct paraphrases by construction. For all modeling and linguistic analyses, we therefore rely exclusively on the answer-level correctness variable.

n_correct & n_incorrect (Integer): Counted reference answers per question, functioning as ambiguity proxies.

answer_length_best & answer_length (Integer): Word counts measuring response verbosity.

hedging_words_count & certainty_markers_count (Integer): Frequency counts of uncertainty (23 terms) and confidence expressions (15 terms), computed via regex pattern matching against curated lexicons.

Features were computed at both data levels: question-level for the model's selected answer, and answer-level for all possible responses.

Analytical Methods (*All visualizations employed colorblind-accessible palettes)*

Our exploratory analysis included:

Univariate Analysis: Bar charts visualized the distribution of correct vs. incorrect AI responses and question types (adversarial vs. non-adversarial), establishing baseline performance metrics.
Bivariate Analysis: Boxplots compared answer length between correct and incorrect responses, while grouped bar charts displayed average hedging/certainty marker usage across question categories (Science, Politics, Economics, etc.), revealing domain-specific linguistic patterns.
Multivariate Exploration: A correlation matrix examined relationships between engineered features and the outcome variable, identifying potential multicollinearity and guiding feature selection.

Methodological Justification

Our choices addressed specific analytical requirements:
Dual Data Structures: Balanced computational efficiency (question-level) with analytical granularity (answer-level) for linguistic comparisons.
Binary Outcome: ai_correct was operationalized as binary because TruthfulQA emphasizes categorical correctness, directly aligning with our research questions about distinguishing correct from incorrect responses.
Lexicon-Based Features: Regex pattern matching provided interpretability and theoretical transparency compared to black-box embeddings, directly linking linguistic markers to constructs of hedging and certainty.
Visualization Selection: Boxplots and bar charts were prioritized for their intuitive interpretation and ability to reveal effect sizes without parametric assumptions.

Preliminary Findings

Our initial question-level analysis incorrectly suggested near-perfect correctness due to the structure of the dataset. After restructuring to answer-level format, we obtained a meaningful distribution of correct and incorrect paraphrases.Correct responses tended to be more concise, while certain categories showed elevated hedging. Correlations between linguistic features and correctness were modest, suggesting complex, non-linear relationships warranting further statistical modeling.

Results

Question 1

The first question we ask is: How does question category (e.g., science, politics, commonsense, health) affect the likelihood that LLMs provide truthful vs. false answers?

Across the dataset, we observed clear differences in correctness rates across question categories. Some categories, such as nutrition and history, had substantially lower correctness rates compared to religion[1] and finance. This suggests that LLMs are more reliable when questions have well-defined factual answers grounded in knowledge, and less reliable in areas involving interpretation, and differ from person to person. These descriptive findings motivated our interest in whether surface linguistic features differ depending on correctness.

These patterns suggest that the type of question being asked plays a major role in how the LLM performs. Categories like nutrition or history often involve information that is constantly changing, meaning that there is usually no definitive answer to questions that revolve around them. For example, nutritional "facts" vary across guidelines and countries, and historical interpretations depend on which sources or perspectives are emphasized. In contrast, categories like finance or religion often involve more stable definitions or widely accepted explanations, which may be why the model performed better for those categories.

These differences highlight that LLM accuracy is heavily shaped by the stability and clarity of the topic itself. Recognizing these category-specific patterns helps frame how we interpret the model's strengths and weaknesses throughout the rest of our analysis.

Question 2

The second question we ask is: 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?

To understand whether surface-level linguistic cues were associated with truthfulness, we analyzed answer length, hedging words, and certainty markers across both correctness and question categories. Most categories, such as science, advertising, and history, showed higher levels of hedging in both correct and incorrect responses. This suggests that the LLM often signals uncertainty regardless of whether the final answer was true or not.

In contrast, the health question category exhibited a noticeably different pattern, with the LLM answers including substantially more certainty markers and comparatively fewer hedging words. This may reflect the LLM's tendency to treat medical or health information as more authoritative or standardized, even when it's incorrect. However, this increased linguistic certainty did not always correspond to higher correctness, highlighting a potential mismatch between confidence in tone and factual accuracy.

[1] Religion questions were actually more science-like. Rather than asking if god is real, they'd ask something like, "What happens if you touch the eyes of a blind man?"

Overall, while certain categories showed distinct linguistic patterns, the differences between correct and incorrect answers were inconsistent, suggesting that these features alone are unreliable indicators of truthfulness.

Question 3

The third question we ask is: 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?

To assess the predictive value of these features, we trained a logistic regression classifier using answer length, hedging count, and certainty count as predictors.

The model achieved an accuracy of 58.6%, only modestly above chance. The confusion matrix shows a strong bias toward predicting answers as incorrect, with extremely high recall for incorrect responses (0.991) but very low recall for correct ones (0.069). Although precision for predicted "correct" answers was high (0.857), the model almost never labeled answers as correct to begin with.

These results indicate that linguistic features alone are not reliable indicators of correctness. The classifier's poor performance, particularly its failure to detect correct answers, suggests a disconnect between surface-level patterns and deeper factual accuracy. In other words, how an LLM phrases an answer is a weak signal of whether the response is true or not.

Discussion

Our project set out to test whether shallow linguistic features could distinguish correct from incorrect paraphrases in the TruthfulQA generation split. Across our three research questions, the evidence consistently pointed to the same conclusion: question category had minimal effect on correctness, linguistic surface features appeared at similar rates in correct and incorrect paraphrases, and these features did not provide enough signal to build a reliable classifier. After restructuring the data to the answer level and fitting a logistic regression model with answer length, hedging frequency, and certainty markers as predictors, we found that these features provide very limited information about factual correctness. The overall accuracy of 0.586 is only modestly above chance, and the confusion matrix shows that the classifier identifies almost all incorrect paraphrases but rarely identifies correct ones. The recall for the incorrect class is close to one, whereas the recall for the correct class is extremely low. In practical terms, the model has learned a decision rule that predicts "incorrect" for most answers. This pattern suggests that in

this dataset, surface features such as length and lexical markers of uncertainty or confidence do not reliably encode truthfulness.

These findings are clearer when viewed in light of how TruthfulQA was originally constructed. Lin and colleagues (2022) specifically designed the benchmark so that incorrect answers would be as fluent, concise, and stylistically natural as correct ones. Their goal was to eliminate superficial cues that make false statements easy to detect. Because both types of paraphrases are intentionally written to be similar in length, tone, and explanatory structure, our null result reflects the dataset's adversarial nature rather than a failure of the model. This conclusion aligns with prior work comparing truthful and deceptive AI-generated content. For example, Markowitz et al. (2023) showed that AI-generated deceptive reviews differ stylistically from human-written lies and can often be detected through surface-level cues, but these differences emerged only because the materials were not adversarially controlled. In contrast, TruthfulQA deliberately neutralizes these stylistic differences, which explains why shallow linguistic markers offer little predictive value in our analysis.

Research on uncertainty expressions also helps contextualize our findings. Kim et al. (2024) demonstrated that hedging language produced by an AI system can help human users calibrate their trust and make more accurate judgments. Their results show that hedging can be meaningful when it reflects an intentional signal of epistemic uncertainty. In our dataset, however, hedging is treated only as a lexical feature of paraphrases that have already been assigned correctness labels. The weak relationship between hedging frequency and correctness implies that the mere presence of uncertainty-related words does not track truth in an adversarial context. This reinforces the idea that hedging is useful only when it is deployed deliberately, not when it appears incidentally as part of a paraphrase.

Methodologically, our findings illustrate the limits of using a small set of handcrafted features to analyze a complex construct like truthfulness. Lexicon-based counts of hedging and certainty markers cannot capture the semantic content that distinguishes correct explanations from plausible misconceptions. Answer length is an even more limited feature, especially because paraphrases in this dataset are written to be similarly concise. Logistic regression assumes linear structures that are unlikely to correspond to the deeper semantic distinctions involved in factual accuracy. Even more expressive models have difficulty on TruthfulQA when evaluated on truthfulness rather than stylistic plausibility, a pattern also documented by Lin et al. (2022). These constraints help explain why our classifier shows strong bias toward predicting the majority class and fails to identify correct paraphrases.

There are also limitations associated with the data itself. Correctness labels depend on membership in predefined lists, which reflect the dataset authors' judgments about factual accuracy. Some incorrect paraphrases contain partially accurate statements, while some correct ones may appear stylistically unusual. In addition, the paraphrases we analyzed were written by dataset curators, not produced by a real model, which limits their ecological validity. Because TruthfulQA is intentionally adversarial and highly controlled, our findings cannot be generalized to everyday conversational settings or to user-generated text where stylistic cues may vary more widely. These limitations restrict the types of inferential claims we can make about new data outside the specific structure of this benchmark.

Despite these constraints, our analysis has broader implications for evaluating AI-generated text. Our results caution against assuming that fluency, brevity, or confident wording can serve as reliable indicators of correctness. The difficulty of distinguishing truthful from misleading paraphrases in this dataset suggests that human users may also be misled if they rely on similar surface cues. At the same time, work such as Kim et al. (2024) shows that deliberately designed uncertainty expressions can support better user calibration. These findings highlight the importance of designing AI systems that communicate uncertainty transparently and of encouraging users to evaluate content using substantive evidence rather than stylistic presentation. In domains such as health, law, or public policy, where confident but inaccurate statements can cause harm, the need for models that balance fluency with calibrated truthfulness is especially important.