Confidently Wrong: Linguistic Authority Signals as Predictors of LLM Incorrectness

Irene Lin, Caitlyn Holmstrom

Department of Computer Science, Stanford University

Stanford Computer Science

Project Overview

- Large language models (LLMs) are increasingly used in high-stakes domains such as healthcare, law, and education.
- Users often rely on authoritative-sounding responses without verifying their accuracy.
- Incorrect responses may appear trustworthy through linguistic signals such as certainty, expertise, and detailed explanations.
- We investigate whether linguistic features can predict LLM incorrectness across multiple models and domains.
- Using TruthfulQA, we compare GPT-4o, Claude, and Llama responses across Health, Law, and Misconceptions categories.
- Our goal is to understand when LLM confidence does not align with accuracy.

Research Questions

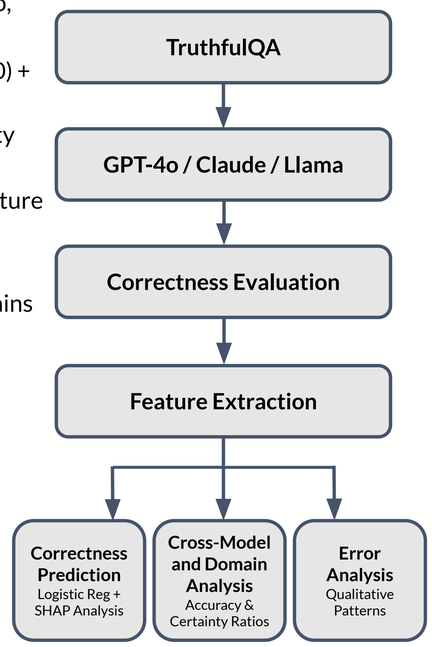
1. Can linguistic signals predict when an LLM response is incorrect?
2. How do relationships between confidence and correctness vary across models and domains?
3. Which linguistic features contribute most to correctness prediction?

Datasets & Metrics

- Dataset: TruthfulQA. 817 questions designed to elicit false or misleading LLM responses across 38 categories (Lin et al., 2022)
- Subset: 198 questions across Health (n=67), Law (n=64), and Misconceptions (n=67), yielding 594 total responses (3 models × 198)
- Correctness: binary labels (correct/not fully correct) via GPT-4o judge + manual audit
- Linguistic features extracted per response:
  - Certainty density: confident words ("definitely," "proven") per 100 words
  - Hedge density: uncertainty words ("might," "suggests") per 100 words
  - Authority cue density: phrases invoking expertise ("research shows") per 100 words
  - Certainty ratio: certainty count / (hedge count + 1)
  - Verbosity: word count

Methods & Experiments

- Generated responses to TruthfulQA questions using GPT-4o, Claude Haiku, and Llama 3.3 70B.
- Correctness scored by GPT-4o semantic judge (temperature=0) + manual audit; refusal responses excluded.
- Extracted linguistic features like certainty, hedging, authority cues, and verbosity from each response.
- Trained logistic regression models and applied SHAP for feature attribution.
- *Experiment 1:* Correctness Prediction
  - Evaluated whether linguistic features, models, and domains predict correctness
  - Used logistic regression and SHAP to identify the most influential predictors
- *Experiment 2:* Cross-Model Confidence Analysis
  - Compared certainty ratios and accuracy across various models (GPT-4o, Claude, and Llama) and domains (Health, Law, Misconceptions)
- *Experiment 3:* Error Analysis
  - Investigated incorrect responses with high certainty ratios and identified recurring patterns



Discussions & Future Research

Discussions:

- Miscalibration is model-specific, not universal. GPT-4o overconfident when wrong. Claude and Llama under-confident when wrong. No single linguistic threshold generalizes across all three models.
- Model and domain identity outweigh raw linguistic features (AUC 0.556 → 0.638 when model/domain added). So, who is speaking and what domain they're in matters more than how they phrase it.
- Law is the highest-risk deployment context: worst accuracy (51.6% for Llama) combined with suppressed certainty language.

Future Research:

- Train model-specific, domain-conditioned classifiers to flag responses
- Investigate whether RLHF training reward signals drive the direction of miscalibration (overconfident vs. underconfident when wrong).
- Extend to additional high-stakes domains (finance, mental health) and more recent model versions.

References

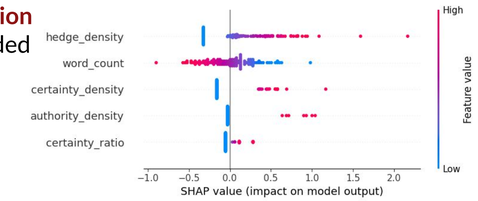
[1] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of ACL 2022*.

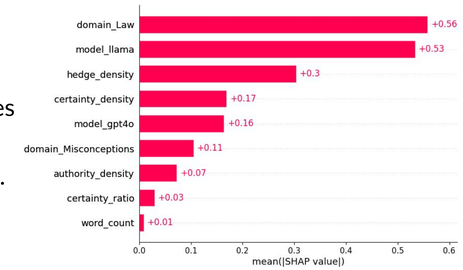
[2] Mielke, S. J., Szlam, A., Dinan, E., & Boureau, Y. (2022). Reducing conversational agents' overconfidence through linguistic calibration. *TACL 10, 857–872*.

Results

Experiment 1: Correctness Prediction

- Linguistic features alone provided limited predictive power **(AUC = 0.556)**.
  - 594 total responses
- Adding model and domain information improved prediction performance **(AUC = 0.638)**.
  - Strongest predictors: Law questions and Llama responses
- Hedge density emerged as the most influential linguistic feature.
- Model and domain effects outweighed linguistic signals.





Experiment 2: Cross-Model Confidence Prediction

Model	Accuracy	Correct CR	Incorrect CR	P
GPT-4o	87.9%	0.138	0.181	0.573 (ns)
Claude Haiku	83.3%	0.139	0.030	0.008**
Llama 3.3 70B	68.2%	0.175	0.061	0.005**

- **Law** domain suppresses certainty across all models **(mean CR = 0.076 vs. 0.192 Health; F(2,591) = 6.13, p < 0.001),** but is still the worst-accuracy domain
- Any flagging system must be model-conditioned

Experiment 3: Error Analysis

High-confidence incorrect responses often **combined misinformation with authoritative language and detailed explanations**, making them appear more credible than they really were.