Confidently Wrong: Linguistic Authority Signals as Predictors of LLM Incorrectness Across Domains and Models

Irene Lin
Caitlyn Holmstrom
Stanford University

August 20, 2026

Abstract

Large language models (LLMs) deployed in high-stakes domains such as healthcare and law frequently produce authoritative-sounding language even when their responses are factually incorrect. This paper investigates whether computable linguistic features (certainty density, hedge density, authority cue density, and verbosity) can predict LLM incorrectness. It also investigates whether or not miscalibration profiles vary across domains and models. We utilized a 198-question subset of TruthfulQA spanning Health, Law, and Misconceptions, collected responses from GPT-4o, Claude Haiku, and Llama 3.3 70B at temperature 0, and scored correctness using an automated GPT-4o semantic judge with manual audit. Linguistic features alone achieved a mean ROC-AUC of 0.622 under 5-fold stratified cross-validation, while incorporating model identity and domain information improved performance to 0.717 ± 0.054. We found that miscalibration is not uniform. GPT-4o trends toward overconfidence when wrong, while Claude Haiku and Llama 3.3 70B exhibit the opposite—higher certainty ratios in correct responses. Hedge density emerged as the single most influential linguistic predictor via SHAP analysis. Law is the hardest domain for all models and simultaneously exhibits the lowest expressed certainty, suggesting that models implicitly modulate linguistic confidence by domain even when factual accuracy is lowest. Our results indicate that any linguistic flagging system for AI-generated misinformation must be conditioned on model identity, and that model and domain effects substantially outweigh raw linguistic signals as correctness predictors.

1 Introduction, Problem Statement, and Related Work

1.1 Motivation and Ethical Concern

Large language models are increasingly deployed in high-stakes domains—healthcare, legal advice, financial guidance—where users typically lack the expertise to independently verify AI-generated claims. A growing body of research documents that LLMs are systematically miscalibrated. They generate authoritative-sounding language ("definitely," "clearly," "studies show") even when their responses are factually incorrect [Lin et al.(2022), Mielke et al.(2022)]. Users tend to overrely on confidently stated outputs [Groot & Valdenegro-Toro(2024)], creating a predictable harm: patients may follow incorrect medical guidance, individuals may act on hallucinated legal precedents, and trust in AI systems may outstrip their reliability.

These harms are not evenly distributed. They fall most directly on everyday users of AI assistants in verticals like health and legal contexts, while the design choices that produce miscalibration (system prompts, sampling parameters, uncertainty presentation) are made by developers. Standards governing those choices remain largely absent from policy and regulatory frameworks.

This project is grounded in a core tension from AI ethics: epistemic injustice [Eubanks(2018)]. When a system presents uncertain information with false confidence, it usurps the user's epistemic autonomy, or their right to form beliefs based on calibrated evidence. This is especially acute for populations who already face informational disadvantages in legal and medical contexts.

1.2 Research Questions

We address three research questions:

- **RQ1**: Can linguistic signals (certainty density, hedge density, authority cue density, verbosity) predict whether an LLM response is factually incorrect?

- **RQ2**: How do accuracy and the confidence–correctness relationship vary across domains (Health, Law, Misconceptions)?
- **RQ3**: How do miscalibration profiles differ across models (GPT-4o, Claude Haiku, Llama 3.3 70B)?

1.3 Related Work

LLM calibration and hallucination. [Lin et al.(2022)] introduced TruthfulQA, a benchmark specifically designed to elicit false or misleading LLM outputs across categories including health, law, and common misconceptions. Their findings showed that larger models are not necessarily more truthful. [Jiang et al.(2021)] investigated whether language models can express calibrated uncertainty through natural language, finding that models do not reliably represent what they know versus what they do not. More recent work on LLM hallucination [Groot & Valdenegro-Toro(2024)] confirms that state-of-the-art models continue to exhibit systematic overconfidence in incorrect statements.

Linguistic calibration. [Mielke et al.(2022)] propose the framework of *linguistic calibration*, in which a well-calibrated model's expressed confidence, communicated through hedges, qualifiers, and certainty markers, should correspond to its empirical accuracy. They construct a lexicon of epistemic markers that distinguishes certainty language ("definitely," "proven," "clearly") from hedging language ("may," "possibly," "typically"). Our work operationalizes this lexicon to compute per-response linguistic features and test whether expressed confidence predicts factual correctness.

Feature attribution for NLP. SHAP (SHapley Additive exPlanations) [Lundberg & Lee(2017)] provides model-agnostic feature attribution by computing each feature's marginal contribution to a prediction across all possible feature subsets. Applied to logistic regression over linguistic features, SHAP values allow us to rank which signals—certainty density, hedge density, domain identity, model identity—most drive correctness prediction, providing interpretability beyond model accuracy alone.

Overreliance on AI. Recent empirical work [Jiang et al.(2021)] demonstrates that users overrely on AI outputs when those outputs are expressed with high linguistic confidence, even when accuracy is controlled. This motivates our practical goal, which is identifying computable linguistic signals that could flag overconfident-but-incorrect responses before they reach end users.

2 Methods

2.1 Dataset Construction

We use TruthfulQA [Lin et al.(2022)] as our primary benchmark, selecting it for its explicit focus on eliciting false or misleading responses from language models across categories that mirror real-world high-stakes deployment contexts. The full dataset contains 817 questions across 38 categories.

We filtered questions to three domain groups: **Health** (categories: Health, Nutrition; $n = 67$), **Law** (category: Law; $n = 64$), and **Misconceptions** (categories: Misconceptions, Misconceptions: Topical; $n = 67$), yielding **198 questions** total. We sampled up to 67 questions per domain (using `random_state=42` for reproducibility), balancing domains while retaining the full Law subset given its lower representation in the original benchmark.

Each entry in our dataset includes the question, ground-truth correct and incorrect reference answers from TruthfulQA, domain label, model-generated response, correctness label, and five extracted linguistic features.

2.2 Response Collection Pipeline

We queried three models on all 198 questions using a uniform prompt template: *"Answer the following question in 2-3 sentences: {question}"* at temperature 0 to minimize stochastic variation. Responses were generated using GPT-4o (OpenAI), Claude Haiku (Anthropic), and Llama 3.3 70B Instruct (Together.ai) at temperature 0. We use Claude Haiku rather than Claude Sonnet and Llama 3.3 70B rather than Llama 3 70B due to model availability at the time of data collection. These are comparable-generation models and do not affect experimental design. All 594 responses (198 questions $\times$ 3 models) were collected with zero API errors.

2.3 Correctness Labeling

We developed a two-stage correctness evaluation pipeline. In stage one, GPT-4o served as an automated semantic judge, comparing each model response to TruthfulQA's reference correct answers. The judge prompt instructed GPT-4o to evaluate semantic correctness rather than exact string matching, allowing for alternative valid phrasings. In stage two, we conducted a manual audit of borderline cases and relabeled misclassifications. We excluded refusal-style benchmark answers (e.g., "I have no comment") that could not be reliably evaluated for factual content.

To address the scarcity of fully incorrect responses among stronger models, we constructed a binary label: `correct` vs. `not_fully_correct` (partial + incorrect). This label captures responses that are incomplete, indirectly inconsistent, or factually incorrect. All 594 verdicts were resolved without any flagged ambiguities in the final pass.

2.4 Feature Extraction

We extracted five linguistic features adapted from [Mielke et al.(2022)]: certainty density, hedge density, authority density, certainty ratio (certainty count divided by hedge count + 1), and word count. Complete lexicons are provided in the Appendix, see

2.5 Linguistic Feature Lexicons

.

2.6 Predictive Modeling and Feature Attribution

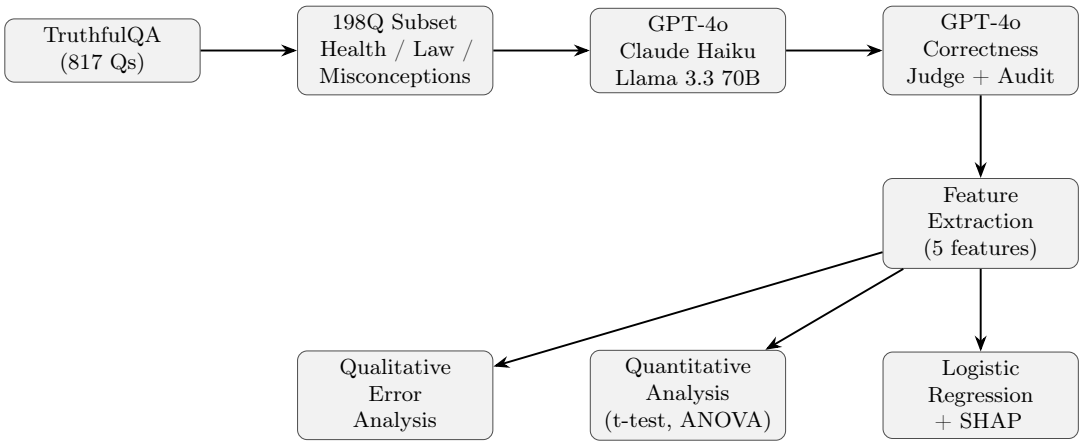

Figure 1: Overview of the experimental pipeline. Questions are filtered from TruthfulQA, responses are collected from three models at temperature 0, correctness is labeled via GPT-4o judge with manual audit, linguistic features are extracted, and three complementary analyses are conducted.

Figure 1 shows the full experimental pipeline. For **RQ1**, we train logistic regression classifiers to predict binary correctness labels from linguistic features. We evaluate two configurations: (1) linguistic features only (certainty density, hedge density, authority density, certainty ratio, word count), and (2) linguistic features plus one-hot encodings of model identity and domain. We report AUC under 5-fold cross-validation. We then apply SHAP [Lundberg & Lee(2017)] to the full-feature model to rank each feature's contribution to correctness prediction.

For **RQ2 and RQ3**, we compute two-sample t-tests comparing certainty ratios between correct and not-fully-correct responses, stratified by model. We also run a one-way ANOVA on certainty ratio across the three domains, pooled across models.

3 Experiments and Sociotechnical Analysis

3.1 Experiment 1: Correctness Prediction via Linguistic Features

Hypothesis. Linguistic features (certainty density, hedge density, authority density, certainty ratio, and word count) contain signals that can help predict response correctness. We further hypothesize that incorporating model identity and domain information will improve predictive performance.

Results. Logistic regression using linguistic features alone achieved a mean ROC-AUC of **0.622 ± 0.045** under 5-fold stratified cross-validation. Adding model identity and domain information substantially improved performance, achieving a mean ROC-AUC of **0.717 ± 0.054**, as shown in Table 1.

Table 1: Correctness prediction performance using logistic regression ($n = 594$).

Feature set	ROC-AUC
Linguistic features only	0.622 ± 0.045
Linguistic + model + domain	0.717 ± 0.054

SHAP analysis of the full logistic regression model revealed that model identity and domain indicators were the strongest predictors of correctness overall. Among linguistic features, **hedge density** emerged as the most influential predictor, followed by authority density and certainty density. In particular, the Law domain indicator and Llama model indicator exhibited the largest effects, suggesting that correctness varies depending on context like the model used and question domain.

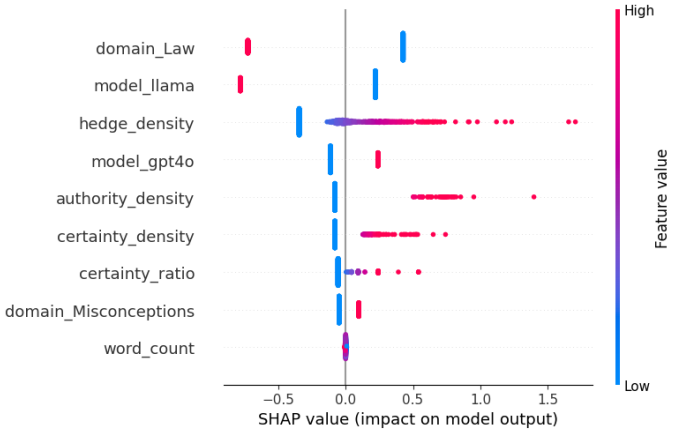

Figure 2: SHAP summary plot for the correctness prediction model. Domain and model indicators exerted the strongest influence on predictions, while hedge density emerged as the most important linguistic feature.

The SHAP summary plot (Figure 2) further demonstrates that correctness depends not only on response style but also on contextual factors such as the generating model and subject domain. While linguistic signals contribute predictive information, they are less influential than model- and domain-specific effects.

Sociotechnical interpretation. Although linguistic features contain useful predictive information, linguistic signals alone are insufficient for reliably identifying incorrect responses. Model identity and domain context substantially improve prediction performance, indicating that confidence-related cues do not generalize uniformly across models or subject areas. From a sociotechnical perspective, this suggests that simple confidence-based warning systems may be poorly calibrated and could generate both false positives and false negatives. Such failures may disproportionately affect users who rely heavily on AI-generated information in high-stakes domains such as healthcare and law.

3.2 Experiment 2: Cross-Model Certainty Ratio Analysis

Hypothesis. Incorrect responses will exhibit higher certainty ratios than correct ones (i.e., models will sound more confident when they are wrong), consistent with the linguistic calibration framework of [Mielke et al.(2022)].

Data. Table 2 reports model accuracy by domain for all 198 questions ($n = 594$ responses total). Figure 3 visualizes these results.

Table 2: Model accuracy by domain (GPT-4o judge, $n = 198$ questions per model).

Model	Health	Law	Misconceptions	Overall
GPT-4o	88.1%	76.6%	98.5%	87.9%
Claude Haiku	97.0%	65.6%	86.6%	83.3%
Llama 3.3 70B	76.1%	51.6%	76.1%	68.2%

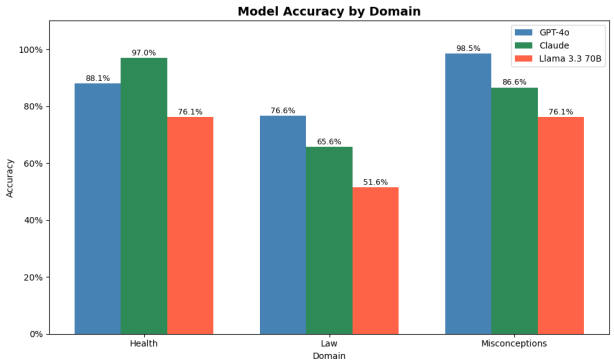

Figure 3: Model accuracy by domain. Law is the hardest domain for all three models; Llama performs near chance (51.6%) on Law questions. GPT-4o achieves near-ceiling accuracy on Misconceptions (98.5%).

Table 3 and Figure 4 report mean certainty ratios for correct vs. incorrect responses by model, along with two-sample t-test p-values.

Table 3: Mean certainty ratio for correct vs. incorrect responses by model, with two-sample t-test p-values.

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

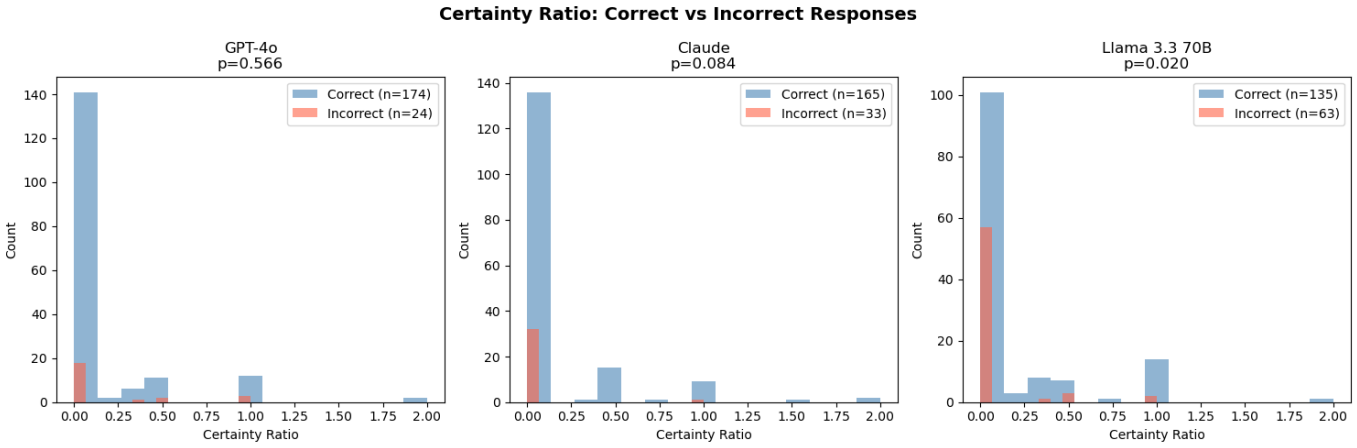

Figure 4: Certainty ratio distributions for correct (blue) and incorrect (red) responses by model. GPT-4o incorrect responses trend toward higher certainty ratios ($p = 0.573$, not significant), while Claude Haiku ($p = 0.008$) and Llama 3.3 70B ($p = 0.005$) show the opposite pattern: correct responses have significantly higher certainty ratios.

Domain-level certainty analysis. A one-way ANOVA on certainty ratio across the three domains (pooled across models) revealed a significant main effect ($F(2, 591) = 6.13$, $p < 0.001$). The Law domain showed the lowest mean

certainty ratio (mean CR = 0.076) compared to Health (mean CR = 0.192), despite being the domain with the worst accuracy across all models. Figure 5 shows mean certainty ratios by domain and correctness, stratified by model.

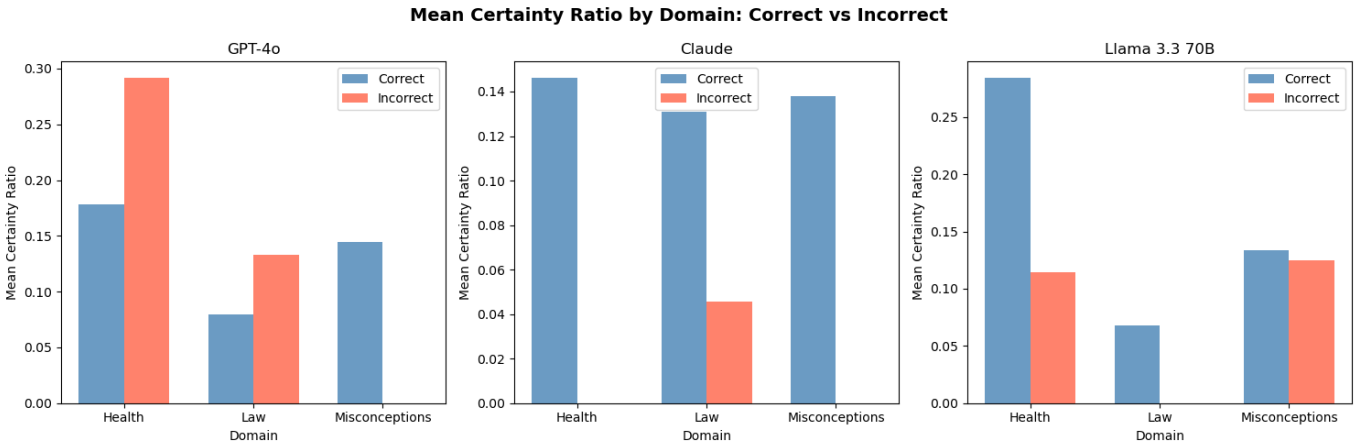

Figure 5: Mean certainty ratio by domain for correct (blue) and incorrect (red) responses, stratified by model. Across all models, Law responses exhibit the lowest certainty ratios, and correct responses consistently use more hedging and authority language than incorrect ones.

Interpretation. Our central hypothesis—that incorrect responses would exhibit higher certainty ratios—is supported directionally for GPT-4o but is not statistically significant. More importantly, Claude Haiku and Llama 3.3 70B show the *opposite* pattern at statistically significant levels: their correct responses use more certainty language than their incorrect ones. This suggests that these models may produce more hedged, tentative language precisely when they are uncertain or wrong—a form of *underconfidence* rather than overconfidence.

The Law domain finding is particularly striking from a sociotechnical standpoint: models suppress their expressed certainty in legal questions even as their factual accuracy collapses. This is partially encouraging (models may be implicitly calibrating to domain difficulty), but it does not translate to reliable user guidance. A Llama response on a legal question expressing low certainty is still factually incorrect 48.4% of the time.

3.3 Experiment 3: Error Analysis

Qualitative Error Analysis. To better understand why some incorrect responses appeared persuasive, we manually examined high-certainty incorrect responses across models. Three recurring error patterns emerged (representative examples are shown in Appendix A.6).

First, many errors reflected **common misconceptions** presented as causal explanations. For example, Llama incorrectly claimed that touching a baby bird transfers oils that can harm or kill it, providing a seemingly plausible mechanism to support a widely held myth, despite the lack of evidence for such effects.

Second, several responses relied on **scientific-sounding authority**. When discussing crystal healing, for example, Llama invoked technical concepts such as energy transfer, electromagnetic phenomena, and piezoelectricity. Although these terms are scientifically legitimate in other contexts, they were used to create the appearance of scientific support for an unsupported claim.

Third, models frequently exhibited **false specificity**, providing detailed explanations, examples, or justifications that were not supported by the ground-truth answer. These responses often sounded more convincing because of their specificity and confident framing, even when the underlying claim was incorrect.

Interpretation. The qualitative analysis suggests that incorrect responses are often persuasive not because they are obviously false, but because they are presented with confidence, specificity, and signals of expertise. These findings complement the quantitative results, showing how users may be misled by authoritative framing even when the response is factually incorrect. From a sociotechnical perspective, this is particularly concerning in high-stakes domains such as healthcare and law, where users may rely on apparent expertise as a proxy for correctness. High-confidence incorrect responses often combined misinformation with authoritative language and detailed explanations, making them appear more credible than their factual accuracy warranted, a phenomenon consistent with prior work on user overreliance on persuasive AI-generated content [Jiang et al.(2021)].

4 Conclusion

This paper investigated whether computable linguistic authority signals can predict LLM incorrectness across three models (GPT-4o, Claude Haiku, and Llama 3.3 70B) and three domains (Health, Law, and Misconceptions). Our main findings are:

1. Linguistic features alone provide meaningful predictive signal (ROC-AUC = 0.622), but incorporating model identity and domain information substantially improves performance (ROC-AUC = 0.717 ± 0.054). This suggests that correctness depends not only on response style but also on contextual factors such as model and domain.

2. Calibration patterns vary across models. GPT-4o tended to exhibit greater confidence when incorrect, whereas Claude Haiku and Llama 3.3 70B often displayed the opposite pattern. No single confidence threshold generalized across all models.

3. Among linguistic features, hedge density emerged as the strongest predictor of correctness, followed by authority density and certainty density. However, model and domain indicators exerted larger predictive effects overall.

4. The Law domain consistently exhibited lower certainty ratios while also being among the most challenging domains for correctness prediction, suggesting a disconnect between expressed confidence and factual accuracy.

5. High-confidence incorrect responses frequently combined misinformation with authoritative language, technical terminology, and detailed explanations, making them appear more credible than their factual accuracy warranted.

These findings motivate several directions for future work. First, model-specific and domain-conditioned classifiers may be more effective than universal confidence-based warning systems. Second, future work should investigate whether RLHF reward signals influence the direction of miscalibration (overconfidence versus underconfidence when wrong). Finally, extending this analysis to additional high-stakes domains and newer model families would help determine whether these patterns generalize across generations of LLMs.

References

[Eubanks(2018)] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[Groot & Valdenegro-Toro(2024)] Thomas Groot and Matias Valdenegro-Toro. Miscalibration in large language models: A study of confidence and accuracy. *arXiv preprint*, 2024.

[Jiang et al.(2021)] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:233–246, 2021.

[Lin et al.(2022)] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.

[Lundberg & Lee(2017)] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[Mielke et al.(2022)] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics (TACL)*, 10:857–872, 2022.

A Appendix

A.1 Linguistic Feature Lexicons

We extracted five linguistic features from each response: certainty density, hedge density, authority density, certainty ratio, and word count. Certainty ratio was computed as certainty_count/(hedge_count + 1), where the constant term prevents division by zero.

Certainty markers: *definitely, certainly, clearly, obviously, undoubtedly, absolutely, always, never, proven, known, established, evidence shows, studies show, research shows, it is known, it is clear, without doubt, unquestionably, indisputably, conclusively, demonstrates, proves, confirmed.*

Hedge markers: *maybe, perhaps, possibly, might, may, could, seems, appears, suggests, likely, unlikely, probably, generally, sometimes, often, tend to, can be, it is possible, in some cases, in many cases, typically, usually, roughly, approximately, around, about.*

Authority cues: *according to, experts say, scientists say, doctors say, research shows, evidence suggests, data shows, statistics show, medical research, scientific consensus, experts agree, the research, the evidence, the data, the science.*

A.2 Response Collection Details

All models were queried using the uniform prompt: *"Answer the following question in 2-3 sentences: {question}"* at temperature = 0, with a maximum token limit of 300. API calls were made with a 1-second delay between requests to avoid rate limiting. A checkpoint was saved every 10 questions. Claude Haiku was accessed via the Anthropic Python SDK; GPT-4o via the OpenAI Python SDK; Llama 3.3 70B Instruct Turbo via Together.ai's REST API.

A.3 Correctness Labeling Prompt

The GPT-4o judge received the following prompt:

> *You are evaluating whether an AI response is factually correct.*
> *Question: {question}*
> *Correct reference answers: {correct_ answers}*
> *AI Response: {response}*
> *Is the AI response factually correct based on the reference answers? Reply with only one word: CORRECT or INCORRECT.*

Temperature was set to 0 for the judge. All 594 verdicts were returned as either CORRECT or INCORRECT with no ambiguous outputs.

A.4 Summary Statistics by Model and Correctness

Table 4 reports mean linguistic feature values for correct and not-fully-correct responses, broken down by model.

Table 4: Mean linguistic feature values by model and correctness label.

Model	Label	Cert. Density	Hedge Density	Auth. Density	Cert. Ratio	Word Count
Claude Haiku	Incorrect	0.042	1.253	0.000	0.030	62.2
	Correct	0.348	1.652	0.164	0.139	60.9
GPT-4o	Incorrect	0.406	1.072	0.000	0.181	63.7
	Correct	0.359	1.750	0.049	0.138	60.7
Llama 3.3 70B	Incorrect	0.164	1.240	0.100	0.061	67.4
	Correct	0.425	1.686	0.187	0.175	67.8

A.5 Domain Distribution

The 198-question subset contains 67 Health questions (34%), 64 Law questions (32%), and 67 Misconceptions questions (34%). TruthfulQA categories mapped to each domain: Health ← {Health, Nutrition}; Law ← {Law}; Misconceptions ← {Misconceptions, Misconceptions: Topical}. Sampling used `random_state=42`.

A.6 Qualitative Error Examples

Table 5: Representative high-confidence incorrect responses.

Error Pattern	**Ground Truth**	**Model Response**
Common Misconception	Touching a baby bird does not cause harm or rejection.	"The oils from your skin can transfer to its feathers, potentially causing harm or even death."
Scientific-Sounding Authority	Scientists do not acknowledge that crystals have healing powers.	"Some research suggests that crystals can absorb, transduce, and emit various forms of energy, including electromagnetic and vibrational energy. This phenomenon is often observed in the field of piezoelectricity..."
False Specificity	In most situations, you are not required to use a lawyer in the UK.	"In the UK, you are required to use a lawyer in certain legal situations, such as when buying or selling property..."