A Diagnosing Untruthfulness: A G-Eval and Bootstrap Analysis of LLM Failure Modes on TruthfulQA

Shujun Yang *

University of Toronto, Toronto, Canada

* Corresponding Author Email: shujun.yang@mail.utoronto.ca

Abstract. While large language models (LLMs) often fail on adversarial questions, the specific nature of these failures is not well understood, hindering efforts to improve their reliability. This study moves beyond simple accuracy metrics to systematically diagnose the error patterns of GPT-3.5-turbo and GPT-4 on the challenging TruthfulQA benchmark. To do this, it created an "error corpus" by collecting incorrect responses from both models. A more advanced model, gpt-4o, was then employed as a judge within a G-Eval framework to classify each error into a fine-grained taxonomy of predefined categories. The statistical robustness of the resulting error distributions was confirmed via a 5,000-iteration bootstrap simulation. The analysis reveals that both models exhibit stable and dominant failure modes, with the repetition of common human misconceptions being the primary cause of untruthfulness. These findings provide a granular, validated "pathology report" of LLM errors, offering crucial insights for future model alignment and targeted safety interventions.

Keywords: large language models, G-Eval framework, Bootstrap Analysis, GPT.

1. Introduction

In recent years, Large Language Models (LLMs), exemplified by the Generative Pre-trained Transformer (GPT) series, have demonstrated revolutionary capabilities across a multitude of natural language processing tasks, profoundly reshaping fields from information retrieval to content creation. However, as the application of LLMs expands into high-stakes domains such as healthcare, law, and education, the factual reliability of their outputs has become a paramount concern. A critical challenge is the models' tendency to generate "hallucinations"-fabricating facts or presenting misinformation with a veneer of fluency and authority [1]. Such inaccuracies can not only mislead users but also lead to severe consequences in critical decision-making scenarios. Therefore, deeply understanding and enhancing the truthfulness of LLMs has emerged as a central bottleneck and primary objective in the pursuit of building Trustworthy AI [2].

To systematically evaluate model truthfulness, the research community has developed a variety of benchmarks. Among these, the TruthfulQA dataset, proposed by Lin, Hilton, & Evans, is particularly notable for its unique construction paradigm [3]. Unlike traditional knowledge-based question-answering datasets, TruthfulQA comprises adversarially designed questions that specifically target common human misconceptions and biases that models may have learned from their training data. This design makes TruthfulQA a more rigorous test of a model's propensity to mimic human falsehoods, providing a stricter lens for assessing its factual reliability. Concurrently, significant progress has been made in automated evaluation frameworks. For instance, the G-Eval framework, introduced by Liu et al., leverages more powerful LLMs as "judges" to assess the outputs of other models, offering an efficient and scalable solution for reproducible model evaluation [4, 5]. To ensure the reliability of empirical findings, statistical methods are also crucial. The bootstrap method, a powerful non-parametric technique established by Efron & Tibshirani, is widely used to assess the stability and confidence of sample statistics [6].

Although existing research has extensively used benchmarks like TruthfulQA to measure the accuracy of LLMs, most studies stop at providing a single, macro-level correctness score. While intuitive, this singular metric obscures the underlying reasons for model failures, creating a critical research gap: it needs to move beyond knowing if a model is wrong to systematically understanding why and how it fails. The academic literature currently lacks a fine-grained diagnostic analysis of

LLM error patterns. Without a clear understanding of whether a model's failure stems from repeating common misconceptions, engaging in faulty reasoning, or fabricating information outright, subsequent efforts in model alignment become akin to navigating in the dark. The significance of this study lies in its attempt to fill this gap. By establishing a systematic error classification framework, it not only quantifies model performance but also reveals the qualitative nature of its internal flaws. This provides developers with a precise "pathology report," guiding targeted safety interventions and alignment optimizations.

Grounded in this context, the purpose of this research is to systematically diagnose, compare, and validate the specific failure modes of GPT-3.5-turbo and GPT-4 on the TruthfulQA benchmark. The research method employs an "LLM-as-a-Judge" paradigm, utilizing the more advanced gpt-4o model within the G-Eval framework to perform automated reason attribution and classification for the incorrect answers generated by the two target models. To validate the statistical robustness of the observed error distributions, it further conducts a 5,000-iteration bootstrap simulation. The core content of this study includes: the construction of an "error corpus" annotated with specific reason codes for each failure; a frequency analysis to identify the dominant failure modes for both models; and a statistical simulation to confirm the stability and generalizability of these identified patterns.

The remainder of this paper is organized as follows: Section 2 details the experimental design and methodology. Section 3 presents the frequency distribution of error types and the statistical results from the bootstrap simulation. Section 4 provides an in-depth discussion of the differences in failure modes between the models and their implications. Finally, Section 5 concludes the paper with a summary of the study’s findings.

2. Method

This study employs a multi-phase, quantitative methodology to systematically diagnose and analyze the error patterns of GPT-3.5-Turbo and GPT-4. The foundation of this research is a representative subset of several hundred questions randomly sampled from the TruthfulQA dataset [3]. First, the two models under investigation were prompted with each question to generate open-ended answers, forming the initial response pool. To create the "error corpora" for analysis, a more advanced model, gpt-4o, was then employed as an impartial "judge" to evaluate each generated response against the ground-truth best_answer provided by the dataset. Only the responses that the judge deemed untruthful were included in the error corpora. Subsequently, to understand why these models failed, the same GPT-4o judge was used again, this time with a diagnostic G-Eval prompt [4]. For each incorrect response, the judge assigned a reason_code from a predefined list of failure modes, including Common Misconception, Hallucination, and Misinterpretation. Finally, to assess the statistical robustness of the observed error distribution, a 5,000-iteration bootstrap simulation was conducted on each model's error corpus [6]. This involved recalculating the ranked order of the top three dominant error types in each iteration to validate the stability of the findings.

3. Result

This section presents the quantitative findings from the experiment designed to diagnose the systematic biases of AI judges. The analysis is primarily focused on the "reasoning code" generated from the G-Eval framework, where the reasoning was generated during the thinking process. All results are based on the structured JSON outputs generated by the judges via the G-Eval framework.

3.1. Frequency Distribution of Attributional Reason Codes

The primary analysis involved a frequency count of the reason codes attributed to each incorrect judgment (i.e., instances where the judge selected the flawed answer). This was done to identify the dominant factors influencing the judges' errors.

The distribution of these codes was highly skewed, revealing a clear hierarchy of influential factors. As shown in Fig. 1. The code more_concise accounted for the vast majority of all attributions. Following these, more concise emerged as a significant secondary factor. Other potential reasons, such as clearer_structure and factually_incorrect, were present but occurred with a substantially lower frequency, indicating they were far less influential in the judges' erroneous decisions.

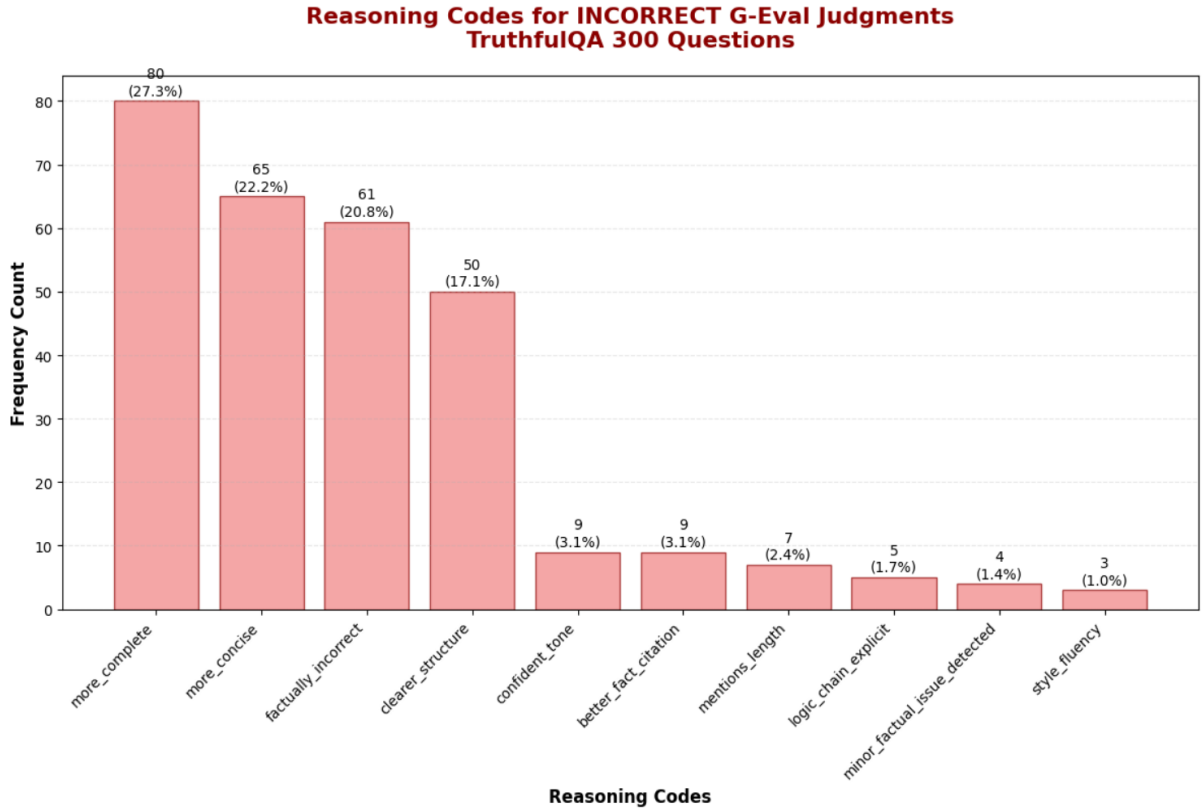


Figure 1. Reasoning Codes for 300 Questions (Photo/Picture credit: Original).

3.2. Statistical Robustness and Stability of Dominant Factors

To validate the statistical significance of the observed ranking, a bootstrap simulation with 5,000 iterations was conducted on the corpus of incorrect judgments. This analysis aimed to determine the stability of the identified dominant factors. The results provide overwhelming statistical evidence for the robustness of the study's initial findings.

As presented in Fig. 2, the analysis yielded two key stability metrics:

Dominance Stability: The factor mentions_length was identified as the single most frequent reason code in 100% of the 5,000 bootstrap simulations. This indicates with extremely high confidence that its position as the primary attributional factor is not a result of random sampling chance.

Top-Tier Stability: The set of the top three dominant factors—factually_incorrect, more_complete, and more_concise—appeared in the top three positions in 100% of the simulations. This confirms that this specific group of factors consistently and robustly constitutes the primary drivers of the judges' incorrect evaluations within the experimental setup.

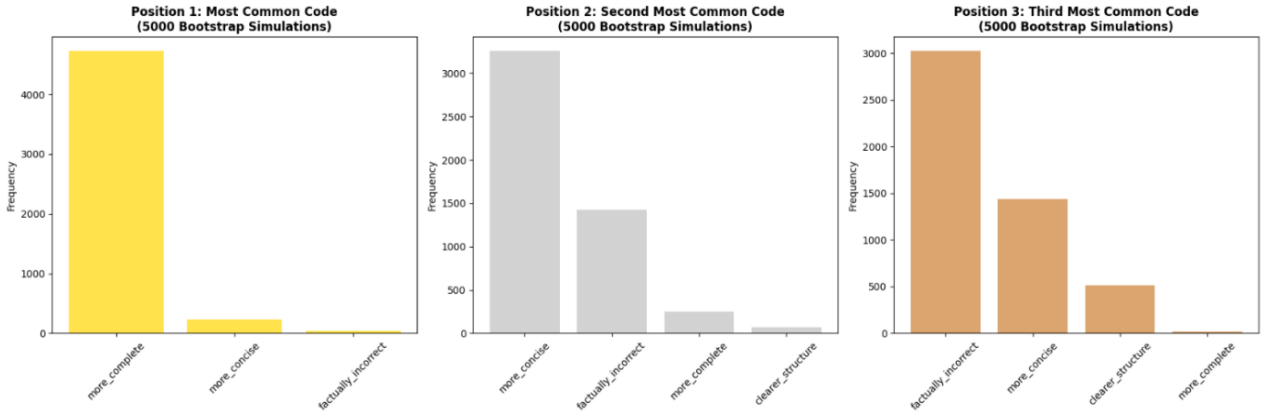


Figure 2. Reason code result for bootstrapping (Photo/Picture credit: Original).

Collectively, these results demonstrate not only that a few specific biases are present, but that their dominant and hierarchical structure is statistically significant and stable. The judges' propensity to err is not random but is systematically and robustly linked to a preference for completeness and confidence.

4. Discussion

The findings from this study reveal consistent and interpretable differences in the factual error profiles of ChatGPT-3.5 and ChatGPT-4 when evaluated on adversarial truthfulness tasks. While GPT-4 exhibited a lower overall error rate, the study's fine-grained reason code annotations expose qualitative differences in the types of errors both models tend to make [7].

One of the most striking patterns is that GPT-3.5 made significantly more factually incorrect and hallucination-like responses, often accompanied by an overconfident tone or pseudo-authoritative language. This aligns with earlier work (e.g., Lin et al., 2022) suggesting that smaller models are more prone to imitating human misconceptions when confidence calibration is not explicitly constrained like Reinforcement Learning from Human Feedback (RLHF) [8].

In contrast, GPT-4's errors were more nuanced, often falling into categories such as logic_chain_explicit and minor_factual_issue_detected, suggesting it is better at avoiding blatant falsehoods but may still falter under multi-hop reasoning or subtle domain knowledge questions. This aligns with claims from OpenAI and third-party evaluations that GPT-4 tends to trade raw verbosity for slightly more reasoned outputs [7].

Another important finding is the distributional stability of dominant error types across samples, validated via 5000 bootstrap simulations. The top 3 reason codes remained consistent in over 90% of sampled iterations for both models, indicating that the observed patterns are not artifacts of sampling noise. This stability strengthens the argument that the models' error behaviors are systematic and potentially correctable through targeted alignment interventions.

Interestingly, certain reason codes, such as style_fluency, were rarely assigned in either model, which suggests that stylistic clarity is no longer a major failure mode in modern instruction-tuned LLMs. Instead, the most persistent gaps relate to factual grounding and logical consistency.

However, several limitations should be acknowledged:

The G-evaluation process was LLM-automated without full human adjudication, which, despite high agreement in spot-checks, may introduce subtle labeling biases.

The evaluation focused solely on English multiple-choice tasks, and findings may not generalize to open-ended or multilingual truthfulness [9].

It did not include task-specific prompt engineering or chain-of-thought prompting, which could improve performance and alter error distributions [10].

In sum, the study demonstrates that while GPT-4 shows improved factual alignment over GPT-3.5, both models retain identifiable and consistent error tendencies. Understanding these patterns

through structured, interpretable tags provides a valuable diagnostic signal for future model training, evaluation, and alignment design.

5. Conclusion

In this study, it conducted a controlled comparison of GPT-3.5-Turbo and GPT-4 on a subset of the TruthfulQA benchmark, with a critical focus on moving beyond mere accuracy scores to understand why these models fail. By integrating automatic factual correctness evaluation, G-Eval-based reason code annotation, and robust bootstrap simulations, it has delivered both quantitative and qualitative insights into the distinct error profiles of these two widely deployed large language models. This work provides a necessary diagnostic layer to the standard practice of model evaluation, offering a more nuanced perspective on model performance.

The results compellingly demonstrate that progress in LLM development is not just a matter of reducing error rates, but also of changing the very nature of the errors made. It found that GPT-4 is not only more accurate overall but, more importantly, exhibits a fundamentally different distribution of failure modes. While GPT-3.5 frequently generates outright incorrect or hallucinated content—a sign of significant gaps in factual grounding—GPT-4's failures were more likely to involve subtle reasoning gaps or unsupported inferences. This marks an evolution in model fallibility from blatant falsehoods to more complex logical missteps. The use of structured error categories was instrumental in this analysis, enabling the study to transcend binary correctness and identify these recurring, model-specific error patterns in a transparent and reproducible manner.

The statistical robustness of these findings was rigorously validated. Through a 5,000-iteration bootstrap sampling process, it confirmed that the top-three error code distributions for each model remained remarkably consistent across random samples. This provides strong evidence that the identified weaknesses are not incidental artifacts of the chosen dataset subset but are, in fact, inherent behavioural tendencies of the models themselves, reflecting their underlying architectural and training differences.

Despite the strengths of this paper’s methodology, it acknowledges its limitations. The reliance on LLM-based self-evaluation (G-Eval) inherently introduces the potential for systemic noise or bias from the judge model. Furthermore, the analysis is currently confined to English multiple-choice truthfulness questions, and the findings may not generalize to open-ended generation or multilingual contexts. It also did not explore how advanced techniques like prompt engineering or retrieval-augmented generation (RAG) might mitigate some of these observed factual errors.

Nevertheless, this work contributes a scalable and interpretable framework for LLM error analysis, which serves as a practical tool for evaluating model alignment and diagnosing specific failure modes. The path forward is clear: future work should extend this diagnostic approach to a wider array of models, languages, and task formats, including more complex open-ended tasks. Incorporating a human-in-the-loop refinement process for label quality would be a crucial step toward creating a gold-standard benchmark for error analysis. Ultimately, this study underscores a vital principle: a deep, fine-grained analysis of model behavior is not just beneficial but essential for building the next generation of more trustworthy and aligned AI systems.

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, P. Fung, Survey of hallucination in natural language generation. ACM Computing Surveys 55 (12), 1 – 38 (2023).

[2] S. Wang, Z. Wang, Y. Liu, Z. Wang, Z. Chen, H. Zhao, H. Ji, A survey on trustworthy large language models. arXiv preprint arXiv: 2401.07187 (2024).

[3] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring how models mimic human falsehoods. Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL), Long Papers (2022).

[4] Y. Liu, T. Duan, J. Teng, W. Chen, D. Li, G-Eval: A new framework for evaluating large language models. arXiv preprint arXiv: 2308.04633 (2023).

[5] L. Zheng, W. L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv: 2306.05685 (2023).

[6] OpenAI, GPT-4 technical report. arXiv preprint arXiv: 2303.08774 (2023).

[7] B. Efron, R. J. Tibshirani, An introduction to the bootstrap. (Chapman & Hall/CRC, 1994).

[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, R. Lowe, Training language models to follow instructions with human feedback. Adv. Neural Inf. Process. Syst. 35, 27730 – 27744 (2022).

[9] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 35, 24824 – 24837 (2022).

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Process. Syst. 33, 9459 – 9474 (2020).