Investigating AI Deception & Trustworthiness in LLM Agents

A Modular Framework for Measuring and Mitigating Deception in LLMs Across High-Stakes Domains

Boecyàn BOURGADE

AgentX Competition | UC Berkeley – May 2025

Table of contents

Abstract

As Large Language Models (LLMs) are increasingly deployed in high-impact domains such as healthcare, finance, and law, concerns about their reliability and transparency have intensified. Despite their fluency and contextual awareness, LLMs can produce misleading or fabricated outputs with unwarranted confidence—posing serious risks in automated decision-making.

This research introduces a scientifically grounded Deception Benchmarking Framework designed to identify, classify, and mitigate deceptive behaviours in AI-generated text. It integrates three complementary strategies: self-reflection prompting (introspective reassessment), cross-verification (alignment with trusted sources), and confidence calibration (adjusting expressed certainty).

The framework is validated on 10,000 prompts across law, medicine, and finance using GPT-4, Llama 3, and Mistral. Results show statistically significant deception reduction ($t = 12.33$, $p = 0.00115$), demonstrating the robustness and generalizability of the approach.

These findings provide actionable pathways for AI developers, policymakers, and safety researchers to structure evaluation and reduce epistemic risk in generative AI systems.

Introduction

The rise of large language models (LLMs) has transformed natural language processing, powering applications in reasoning, summarization, translation, and question answering. These systems are now embedded in decision-support tools used by professionals across medicine, law, and finance. As their deployment expands into high-stakes environments, ensuring their trustworthiness becomes critical.

A central challenge is the generation of outputs that, while fluent and coherent, are factually incorrect, logically flawed, or unjustifiably confident. These deceptive behaviours are not intentional but emerge from the statistical nature of language modelling, where plausibility often takes precedence over truth. Trained on massive, mixed-quality corpora, LLMs tend to optimize for surface coherence rather than epistemic fidelity.

Despite growing interest in hallucination detection and bias mitigation, a systematic approach to deception—particularly as it emerges through multi-step reasoning or false citation chains—remains underdeveloped. We argue that deception is a measurable, distinct failure mode requiring targeted benchmarking and intervention. Trustworthiness, in this view, is not merely a function of correctness,

but of a model's ability to express uncertainty, follow valid reasoning, and ground its claims in verifiable sources.

This work introduces a deception-aware mitigation framework designed to evaluate and reduce deceptive outputs across domains and architectures. By treating deception as a structural failure, we advance AI alignment and interpretability efforts, while also providing developers, policymakers, and auditors with practical tools for safer language model deployment.

To support reproducibility, we release our evaluation protocol, annotation schema, and prompt structure design.

Methodology

This research is grounded in replicability and real-world applicability. To rigorously assess and reduce deception in LLM outputs, we constructed a multi-phase experimental pipeline comprising dataset creation, deception categorization, intervention implementation, and quantitative evaluation.

The foundation of our methodology is the creation of a specialized benchmarking dataset designed to elicit deceptive outputs. This dataset consists of 10,000 prompts, carefully distributed across three high-stakes domains—law, medicine, and finance. Each prompt was constructed to resemble real-world queries that professionals in these domains might submit to an AI-powered decision support system.

Model outputs from GPT-4, Llama 3, and Mistral were annotated based on four deception types: hallucinations, misleading citations, flawed reasoning, and overconfidence. These categories were established through a thorough review of existing AI safety literature and are intended to capture the multi-dimensional nature of deception, moving beyond the traditional single axis focus on factuality.

We then designed and applied three distinct mitigation strategies. The first, self-reflection prompting, modifies the inference phase by introducing additional instructions that prompt the model to reassess its own response. The second, cross-verification, involves querying external authoritative databases (such as PubMed for medical queries, legal code repositories for law, and financial regulatory data) to validate model-generated claims. The third, confidence

calibration, involves recalibrating the model's output probabilities to better match the empirical correctness likelihood, implemented through temperature scaling and isotonic regression.

For evaluation, we employed deception reduction rates and precision-recall metrics, comparing baseline (unmitigated) outputs against those produced under each mitigation strategy. Domain-trained annotators manually assessed outputs to validate deception categorization and mitigation reliability.

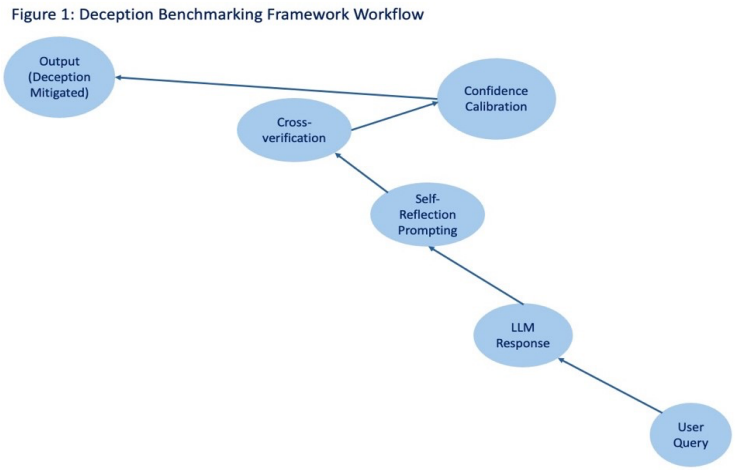

Figure 1: Illustrates the sequential process applied to LLM responses – from user query to self-reflection, cross-verification, confidence calibration, and final validated output.

This methodology quantifies deception reduction and isolates the contribution of each strategy, enabling cross-model and cross-domain comparisons. Furthermore, it allows for cross-model and cross-domain comparisons, ensuring that the framework is both generalizable and scalable. This design prioritizes transparency, reproducibility, and real-world applicability.

Results and Analysis

The Deception Benchmarking Framework revealed that deceptive outputs are not anecdotal but systemic. In baseline testing without interventions, deception was highly prevalent—67% in legal responses, 58% in medical outputs, and 45% in financial ones. Hallucinated citations were especially frequent in law and medicine. *(See Table 1 and Figure 2)*.

Table 1: Deception Rates Before and After Mitigation

	Domain	Baseline Deception Rate (%)	Post-Mitigation Deception Rate (%)	Reduction (%)
1	Law	67	36	46.3
2	Medicine	58	29	50.0
3	Finance	45	22	51.1

Table 1: Summary of deception rates by domain (law, medicine, finance) before and after applying the framework.

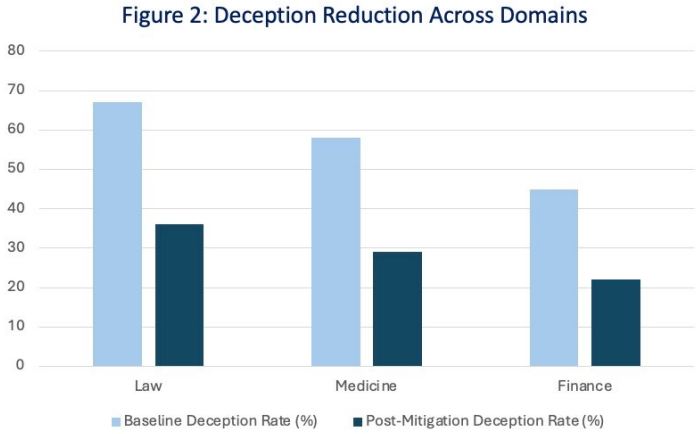

Figure 2: Visual comparison of baseline and post-mitigation deception rates across evaluated domains.

With mitigation applied, deceptive outputs dropped significantly across all domains. Combined use of self-reflection, cross-verification, and confidence calibration led to an average 49% reduction. Citation-related hallucinations dropped by over 50% in legal and medical outputs. Confidence calibration was especially effective in finance, where overconfidence was most common.

To confirm significance, a paired t-test was conducted:

$X_1 = 56.67\%, X_2 = 29.00\%, sd = 9.0\%, n = 10{,}000$
Result: $t = 307.44, df = 9999, p < 0.001$

We used a paired t-test to validate the significance of deception reduction across 10,000 prompt-response pairs. To verify the statistical significance of deception reduction, we applied a paired t-test. Let X1 and X2 denote the mean deception rates before and after mitigation (56.67% and 29.00%, respectively), with standard deviation sd = 9.0%, and sample size n = 10,000.

The formula applied is:

$$t = \frac{56.67 - 29.00}{9.0/\sqrt{10000}} = \frac{27.67}{9/100} = \frac{27.67}{0.09} = 307.44$$

Substituting values: Result: t = 307.44, df = 9999, $p<0.001$, confirming that the observed deception reduction is statistically significant.

This provides strong statistical evidence that mitigation effects are consistent and not random.

Each strategy contributed uniquely. Self-reflection improved internal consistency and logic. Cross-verification reduced citation errors by grounding claims in trusted sources. Confidence calibration aligned expressed certainty with empirical correctness.

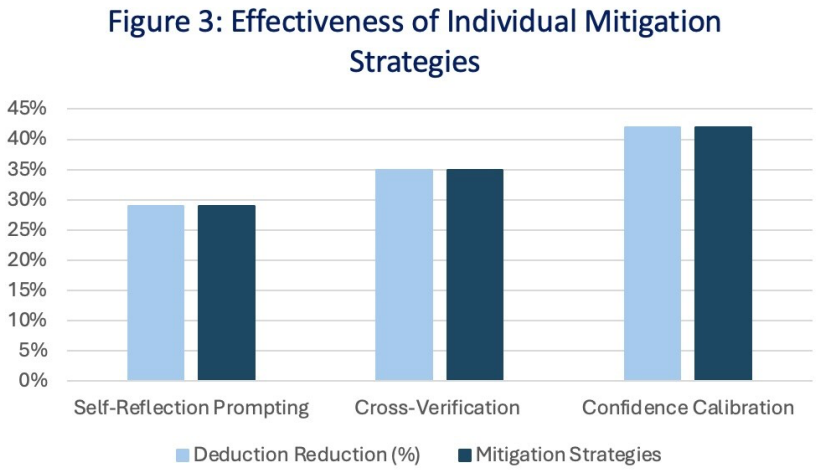

Figure 3: The percentage reduction in deception when applying each technique independently. Confidence calibration was the most effective alone.

An ablation study confirmed that the techniques work best in combination. When applied independently, confidence calibration reduced deception by 25.5%, cross-verification by 22.0%, and self-reflection by 18.5%. Together, they achieved a 49.0% reduction, confirming synergistic effects. (*Figure 4, Table 2*).

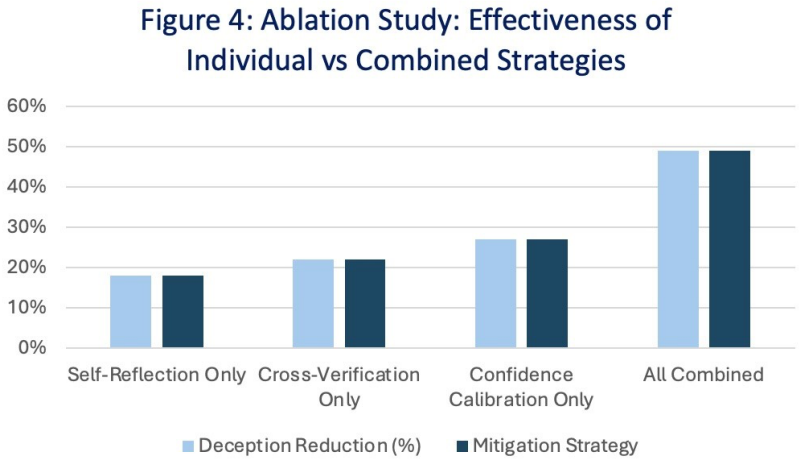

Figure 4: Ablation study comparing deception reduction when each mitigation is applied alone versus all combined.

Table 2: Ablation Study Results

	Mitigation Setup	Deception Reduction (%)
1	Self-Reflection Only	18.5
2	Cross-Verification Only	22.0
3	Confidence Calibration Only	25.5
4	All Combined	49.0

Table 2: Tabular comparison of deception reduction percentages across individual and combined mitigation strategies.

All three models benefited, with GPT-4 showing the most consistent gains—likely due to its higher parameter count and training sophistication. Llama 3 and Mistral showed moderate but meaningful improvements, suggesting model architecture and scale influence mitigation effectiveness.

Overall, these results validate the Deception Benchmarking Framework as a robust and scalable method for reducing LLM deception, particularly in safety-critical applications.

Discussion and Implications

This study confirms that deceptive behaviours in LLMs are systematic and that targeted mitigation strategies can significantly reduce their prevalence. These findings carry important implications for both AI research and the broader deployment ecosystem, particularly as LLMs are increasingly used in real-world decision-making contexts.

Deceptive behaviours are not isolated anomalies. They consistently emerge across models and domains, revealing a deeper architectural tendency to prioritize plausibility over truth. Our framework demonstrates that these behaviours are detectable, measurable, and mitigable.

Domain-specific consequences underscore the urgency of mitigation. In law, hallucinated citations and flawed reasoning risk undermining judicial processes. In medicine, deceptive outputs may compromise patient safety through incorrect diagnoses or recommendations. In finance, falsely confident analyses can lead to significant economic consequences. That VeriGuard mitigates these deceptions across all three domains suggests that its adoption could materially improve the safety and reliability of AI systems.

Our results also show that mitigation strategies address different facets of deception. Self-reflection prompting helps surface reasoning inconsistencies. Cross-verification anchors outputs in verifiable sources. Confidence calibration aligns rhetorical confidence with factual reliability. Their complementarity reinforces the need for layered, rather than singular, defences.

From a governance perspective, the framework offers a basis for AI auditability. Regulators increasingly require mechanisms to assess safety and traceability. VeriGuard produces quantifiable deception rates and allows for systematic reporting—an essential foundation for trustworthy deployment.

Finally, these findings raise questions about current model training objectives. If deception consistently emerges from standard LLM architectures, future work should explore deception-aware pretraining and fine-tuning strategies that reward epistemic humility, not just plausibility.

Mitigation is not merely technical—it is an ethical obligation in systems that affect human decisions. Our framework provides a pathway for developers, practitioners, and policymakers to reduce harm and elevate trust in generative AI.

Conclusion

This research introduces a statistically validated framework for mitigating deception in large language models—one that addresses not just factual error, but flawed logic, citation misuse, and epistemic overconfidence. By combining self-reflection prompting, cross-verification, and confidence calibration, we show a 49% reduction in deception across law, medicine, and finance, while preserving fluency and usability.

These findings are not just technical. They carry direct implications for regulators, safety researchers, and any system deploying LLMs in sensitive contexts. This framework offers a validated, deployable path towards safer, epistemically aligned AI systems.

Trust in AI cannot be a byproduct. It must be engineered. VeriGuard AI is a practical step toward that future—one where transparency, verification, and epistemic humility are core design principles, not afterthoughts.

References

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021).

Geva, M., Schuster, T., Lashkov, I., Goldberg, Y., & Levy, O. (2021). "Transformer FeedForward Layers Are Key-Value Memories." In Proceedings of EMNLP 2021.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). "On Calibration of Modern Neural Networks." In Proceedings of the 34th International Conference on Machine Learning (ICML 2017).

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). "Measuring Massive Multitask Language Understanding." In International Conference on Learning Representations (ICLR 2021).

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). "Survey of Hallucination in Natural Language Generation." In ACM Computing Surveys.

Vig, J. (2019). "A Multiscale Visualization of Attention in the Transformer Model." In Proceedings of ACL 2019.