# Trustworthy Reasoning and Hallucination Mitigation in Large Language Models

Maurice D. Thornton
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
mdthornton@ucf.edu

Cody Stanley
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
cstanley@uc.edu

Bjorn Welch
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
welchbjorn@binghamton.edu

**Abstract**

Large language models are increasingly embedded in decision support systems across medicine, law, science, and public administration. Their capacity to produce fluent and seemingly well-reasoned text has made them attractive components of socio-technical infrastructures. At the same time, hallucination remains a persistent source of risk, and trustworthy reasoning cannot be achieved through a single architectural or algorithmic intervention. This paper provides a systems-level analysis of reasoning reliability and hallucination mitigation in large language models. It examines why hallucination arises from the interaction between generative pretraining, probabilistic decoding, evaluation constraints, and deployment context. The discussion moves beyond model accuracy to consider the structural trade-offs associated with reasoning depth, inference cost, verification, alignment, runtime guardrails, and institutional oversight. The paper reviews the role of chain-of-thought and search-based reasoning, verifier training, process supervision, human feedback alignment, lightweight trajectory filtering, self-evaluation, and continuous deployment monitoring. It further addresses fairness, robustness, sustainability, and policy implications of building reliable language model systems. The central argument is that hallucination should be treated as an emergent property of a larger pipeline rather than a narrow defect of a single model. Trustworthy reasoning therefore requires layered technical safeguards, transparent documentation, and governance mechanisms that enable failures to be detected, contained, and corrected before they produce institutional harm. Future directions include adaptive risk-based verification, stronger separation between linguistic generation and logical inference, cross-domain evaluation, and policy frameworks that focus on systemic accountability rather than isolated benchmark performance.

**Keywords**

Trustworthy reasoning, hallucination mitigation, large language models, alignment, verification, deployment governance, socio-technical systems, fairness.

## 1. Introduction

The rapid integration of large language models into decision support, scientific research, legal analysis, medical triage, and customer operations has transformed artificial intelligence from

a laboratory concern into an infrastructural system. The breadth of these deployments means that failures in reasoning are no longer isolated technical errors; they propagate across institutional workflows and can affect individuals, communities, and organizational accountability. Previous assessments of foundation models have emphasized their generality, adaptability, and emergent capabilities [1]. Early demonstrations of few-shot learning showed that a single model could perform many natural language tasks without task-specific fine-tuning [2]. These observations encouraged broad adoption but also obscured a fundamental challenge: the fluency of a language model is not equivalent to the reliability of its reasoning.

Trustworthy reasoning requires that a system produce not only plausible text but also claims whose inferential structure can be inspected, verified, and aligned with the expectations of a given domain. Hallucination mitigation therefore extends beyond suppressing false statements at the output layer. It must address the data used for pretraining, the architecture that shapes how evidence is aggregated, the training objectives that reward certain patterns of generation, the runtime guardrails that constrain high-risk outputs, and the governance mechanisms that allocate accountability when failures occur. This paper provides a system-level examination of trustworthy reasoning and hallucination mitigation. It treats hallucination as an emergent property of a larger socio-technical pipeline rather than a narrow defect of a single model. Subsequent sections examine architectural trade-offs, verification and alignment infrastructure, deployment governance, fairness, robustness, and sustainability.

## 2. Reasoning and Hallucination in Large Language Models

Large language models are typically trained to maximize the likelihood of sequences over massive text corpora. This objective produces models that are exceptionally sensitive to statistical regularities, discourse structure, and lexical associations. However, the same objective does not explicitly force the model to maintain a stable correspondence between its generated text and an external or internal state of truth. Survey evidence on hallucination in natural language generation identifies several failure modes, including intrinsic fabrication, where the model invents content without grounding, and extrinsic mismatch, where the model draws on training knowledge incorrectly for a novel context [3]. These failure modes are compounded by the self-attention mechanisms that enable transformers to compose long-range dependencies [4]. Self-attention permits flexible contextual representation, but it also creates multiple pathways through which a confidently worded but unsupported token can influence later generation.

Empirical evaluations of very large systems have reported seemingly coherent analytical behavior that can appear robust in conversational settings while remaining brittle under adversarial or unusual queries [5]. This discrepancy is important for trustworthy reasoning. A system may pass benchmark questions yet still fail when the distribution shifts because the model has learned to imitate the surface grammar of reasoning rather than the underlying inferential operations. The architectural and statistical properties that support broad generalization also make it difficult to determine whether a generated chain of thought corresponds to a causal process, a post hoc rationalization, or a probabilistic paraphrase of similar arguments seen during training. Consequently, hallucination mitigation cannot be approached only by increasing model size or retraining with additional text. It must be addressed through the deliberate design of reasoning procedures, verification mechanisms, and operational constraints that surround the model.

## 3. Architectural Foundations and Structural Trade-offs

The choice of reasoning architecture shapes what kinds of verification are possible. Chain-of-thought prompting demonstrated that requiring a model to produce intermediate reasoning steps can improve performance on arithmetic, commonsense, and symbolic tasks [6]. This approach increases the transparency of the generation process by exposing intermediate claims, but it also increases inference cost and creates new opportunities for subtle fabrication. A model can produce a fluent chain that concludes with an incorrect answer while appearing methodical. Tree-of-thought prompting extends this idea by allowing the model to explore multiple reasoning branches, evaluate partial solutions, and backtrack when a path appears unpromising [7]. Such search-based reasoning improves robustness on planning tasks, yet it introduces a structural trade-off between reasoning depth and compute. Systems designed for high-volume deployment may not be able to support extensive search for every request.

From a systems perspective, the architecture around the language model matters as much as the model architecture itself. Retrieval-augmented generation, tool use, and structured output layers can reduce hallucination by anchoring generation in external information. However, these additions move the system toward a distributed architecture in which the language model is one component among many. This distribution creates coordination problems: the retrieved evidence may be outdated, the tool interface may return conflicting data, and the language model may override reliable tool outputs when the prompt pressure is high. Therefore, architectural decisions must balance the transparency gained from explicit reasoning paths, the reliability gained from external grounding, and the operational cost introduced by additional components. A well-designed trustworthy system should explicitly define which parts of the answer are generated by the language model, which parts come from retrieval, and which parts are the result of deterministic checking. This separation is not only a technical requirement; it is also a governance requirement because it determines how failures can be traced and corrected.

## 4. Verification, Alignment, and Training Infrastructure

Verification is a critical layer in hallucination mitigation. Rather than relying on a single generated answer, systems can train separate verifier models or use the language model itself to evaluate candidate solutions. Early work on training verifiers for mathematical word problems showed that solution verification can outperform direct answer generation when the verifier is exposed to multiple candidate solutions [8]. Subsequent work on process supervision refined this idea by rewarding individual reasoning steps rather than only final answers [9]. Process supervision offers a more granular signal for trustworthy reasoning because it encourages the model to maintain correctness throughout the solution path. From an infrastructure standpoint, process-level verification is more expensive to collect and annotate, but it improves the ability of system operators to identify exactly where a reasoning chain diverges from an acceptable path.

Alignment methods add another layer of behavioral control. Open model releases such as Llama 2 have demonstrated that instruction tuning and human preference optimization can be applied successfully to models intended for external deployment [10]. Instruction tuning with human feedback, as operationalized in the InstructGPT pipeline, combines supervised fine-tuning with reward modeling to reduce unhelpful or unsafe outputs [11]. The reinforcement learning step often uses proximal policy optimization to update the policy while limiting large destructive updates [12]. Targeted human judgements further refine alignment by focusing on specific failure modes such as sycophancy, evasiveness, and unsafe advice [13]. These

training procedures are not simply performance improvements; they are mechanisms for encoding institutional values and domain constraints into model behavior.

Recent work has also explored lightweight methods for reasoning path filtering and alignment in small parameter language models. For example, a scalable trajectory filtering technique has been proposed to reduce reasoning path overhead while retaining alignment benefits in resource-constrained settings [14]. This type of approach matters for deployment across edge devices, regulated sectors, and organizations that cannot operate extremely large models. By filtering candidate reasoning paths before final selection, the system can reduce the cost of search and verification while still avoiding many low-quality trajectories. At the same time, such methods introduce new governance considerations because the filtering policy itself must be validated, monitored, and updated when the operational environment changes. The overall lesson from verification and alignment research is that trustworthy reasoning is not a static property of pretraining but an ongoing process supported by training infrastructure.

## 5. Deployment-Time Guardrails and Runtime Governance

Even a well-trained model can hallucinate when exposed to ambiguous, adversarial, or distribution-shifted queries. Deployment-time guardrails are therefore essential. One useful capacity is self-evaluation: models can be prompted to estimate their own uncertainty, and empirical work has shown that language models can often distinguish between what they know and what they do not know [15]. Self-evaluation is not sufficient on its own because model confidence can be miscalibrated, but it provides an important signal for routing hard cases to specialized verifiers or human reviewers. In addition, runtime governance can use deterministic detectors, rule-based fact-checking, citation checks, and tool-based retrieval to confirm outputs before they reach users. The availability of open-source transformer libraries has reduced the integration cost for such guardrails, allowing institutions to deploy custom verification pipelines around foundation models [16].

Evaluation must be continuous and context-specific. Static benchmark suites such as the Massive Multitask Language Understanding benchmark provide broad comparison across subjects, but they do not capture the full risk profile of a deployed system [17]. A legal tool may require rigorous citation grounding, while a clinical support system may need conservative behavior under uncertain conditions. Deployment governance should combine automated tests, red-teaming exercises, human audit, and incident reporting. These mechanisms should be reviewed whenever the underlying model, retrieval corpus, or prompt configuration changes. By treating inference as a managed system rather than a single model call, organizations can detect hallucinations before they cascade into institutional decisions and can maintain evidence about system behavior for downstream accountability.

## 6. Fairness, Robustness, and Socio-Technical Trade-offs

Fairness and robustness concerns extend beyond factual accuracy. Language models trained on large internet corpora can reproduce stereotypes, amplify dominant perspectives, and produce confident outputs that reflect historical biases. Research on stochastic parrots has argued that very large language models can create the appearance of understanding while lacking grounding in human communicative intent, thereby posing risks of misuse and misrepresentation [18]. In high-stakes contexts, a hallucinated fact may combine with a biased prior to produce decisions that disproportionately affect minoritized groups. Ensuring fairness therefore requires evaluating not only whether a claim is true but also whose perspective is

represented, which data sources are prioritized, and whether the system performs consistently across demographic and linguistic communities.

Robustness is likewise a socio-technical property. A model may be robust to one type of perturbation but fragile to another, and the choice of robustness metric reflects organizational values. Model cards and structured reporting frameworks have been proposed to document intended use, training data, evaluation results, and ethical limitations [19]. Such documentation improves transparency and allows downstream users to make informed decisions about deployment. However, documentation alone does not guarantee fairness. It must be integrated with procurement review, impact assessment, recourse mechanisms, and ongoing monitoring. The structural trade-off is clear: stronger safety mechanisms often reduce throughput and increase cost, while faster deployment without such mechanisms transfers risk to affected communities. Trustworthy reasoning therefore depends on how institutions navigate this trade-off rather than on any single technical safeguard.

## 7. Sustainability and Organizational Policy

Sustainability considerations are increasingly central to large-scale reasoning systems. The computational cost of chain-of-thought search, process supervision, and repeated verification can be substantial. Lightweight trajectory filtering and small parameter alignment approaches such as that reported in [14] address part of this burden, but system-level sustainability also depends on model selection, caching, route optimization, and the use of smaller specialist models where possible. Organizations should evaluate whether the marginal reliability gain from an extremely large model justifies its energy, latency, and operational cost. In many settings, a combination of a smaller language model, rigorous retrieval, deterministic rule checks, and human review may provide a more sustainable and auditable solution than a single large generative system.

Policy must also address accountability. When a system produces a hallucinated legal citation or medical instruction, the failure cuts across model provider, application developer, data broker, and deploying institution. Governance mechanisms should define roles for each actor and require clear documentation of model version, prompt design, retrieval sources, and verification procedures. Continuous monitoring should include semantic drift, user feedback, adversarial probes, and changes in external knowledge bases. These practices are not purely technical; they require organizational capacity, regulatory alignment, and cross-disciplinary expertise. The goal is not to eliminate every possible hallucination, which is likely unattainable for large generative models, but to build a system in which hallucinations are more likely to be detected, contained, and corrected before they cause harm.

## 8. Future Directions

Future research should move toward adaptive verification that adjusts reasoning depth to risk. Low-risk queries can be handled by lightweight filtering, while high-risk queries can be routed through deeper search, external retrieval, and human review. This adaptive model requires better uncertainty estimation, context-aware risk classification, and standardized audit trails. It also requires stronger connections between language generation and structured knowledge representations. Retrieval-augmented systems already provide partial grounding, but future architectures may need more explicit separation between linguistic generation and logical inference. Such separation could help reduce cases in which a model produces text that looks like reasoning but lacks coherent inferential content. At the same time, the field must avoid the assumption that more explicit reasoning always increases trust. Explicit chains

can be fabricated, and increased transparency may lead to overtrust if users do not inspect the underlying evidence. Therefore, future systems should be evaluated not only on answer accuracy but also on whether the reasoning trail supports appropriate human verification and whether users can identify when the system is uncertain.

Cross-domain comparisons are also needed. Medical, legal, scientific, and financial environments impose different standards of evidence, different tolerance for uncertainty, and different institutional review processes. A reasoning architecture that is acceptable in customer service may not be acceptable in clinical triage. Research should examine how these domain-specific requirements shape verification, documentation, and human oversight. Finally, policy research should keep pace with technical change. Regulatory frameworks that focus only on model size or benchmark performance will miss the systemic factors that determine trustworthiness in practice. Effective governance will require interoperable incident reporting, independent auditing, public benchmarking of failure modes, and clear pathways for affected individuals to contest automated decisions. These developments must be considered alongside the economic and computational costs of safer systems so that trustworthy reasoning becomes a sustainable norm rather than an occasional high-cost configuration.

## 9. Conclusion

This paper has argued that trustworthy reasoning in large language models cannot be achieved through a single architectural or algorithmic intervention. Hallucination emerges from the interaction between generative training, probabilistic decoding, evaluation limitations, and deployment context. Mitigation requires a layered system that includes explicit reasoning procedures, verifier training, process supervision, alignment, lightweight trajectory filtering, runtime guardrails, human oversight, and governance. Structural trade-offs must be acknowledged: deeper reasoning and stronger verification increase reliability but also raise computational cost, latency, and operational complexity. Open model ecosystems enable customization but place greater responsibility on deploying institutions. Fairness and sustainability further complicate the design space because safety mechanisms can impose disproportionate costs and may not distribute benefits equally. The future of trustworthy reasoning lies not in eliminating hallucination entirely but in designing systems that can detect, contain, and learn from failures. This requires interdisciplinary collaboration among machine learning researchers, systems engineers, domain experts, auditors, and policymakers. Only through such collaboration can large language models become reliable components of socio-technical infrastructures.

## References


1. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

3. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

5. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

6. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

7. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36, 11809-11822.

8. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2023). Let's verify step by step. arXiv preprint arXiv:2305.20050.

9. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

10. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

12. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

13. Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P.-S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., ... Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. arXiv preprint arXiv:2209.14375.

14. Li, Q. (2026, May). TrajLite PPO: Scalable Reasoning Path Filtering and Alignment for Small Parameter Large Language Models. In 2026 2nd International Conference on Artificial Intelligence and Digital Ethics (ICAIDE) (pp. 29-32). IEEE.

15. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones,

A., Elhage, N., Hume, T., Chen, A., Bai, Y, Bowman, S., Fort, S., ... Kaplan, J. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

16. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45). Association for Computational Linguistics.

17. 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.

18. 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 (pp. 610-623). ACM.

19. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229). ACM.