Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9019

Architecting Trustworthy LLMs: A Unified TRUST Framework for Mitigating AI Hallucination

Lingyi Meng
University of Pittsburgh, Pittsburgh, United States
Email: lim136@pitt.edu

How to cite this paper: Meng, L. (2025). Architecting Trustworthy LLMs: A Unified TRUST Framework for Mitigating AI Hallucination. Journal of Computer Science and Frontier Technologies, 1(3), 1-15. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9019
Published: 2025-11-10

Abstract

The use of large language models (LLMs) in high-stakes domains like engineer-ing and public policy is hindered by their tendency to hallucinate—generating factually incorrect or unverifiable information. Although technical mitigations such as Retrieval-Augmented Generation (RAG) and Reinforcement Learning from Human Feedback (RLHF) exist, they are typically applied as isolated point-solutions, lacking a unifying framework. This paper analyses the archi-tectural roots of AI hallucination, classifying them into cognitive gaps and sta-tistical biases, and then shows the shortcomings in current mitigation strategies. To address this, we introduce the TRUST model, which is a systematic structure built on five interconnected pillars: Transparency, Reliability, Uncertainty, Su-pervision, and Traceability. The TRUST framework serves as the architectural basis to orchestrate existing technologies and establish accountability in the lifecycle of AI and to build robust and accountable AIs for vital societal applications [2].

Keywords

AI hallucination; large language models; trustworthy AI; TRUST model; retrieval-augmented generation (RAG); human-AI collaboration

1. Introduction

Generative artificial intelligence (AI), especially large language models (LLMs), such as GPT-4 and Llama, has brought about automation and cooperation of AI in human-machine systems [25]. Such models have remarkable language efficiency in text generation, complex reasoning and code generation, used for everything from advanced medical diagnostics to the management of urban infrastructure [14]. But this revolutionary power is undermined by deep, built-in risk: AI hallucination [3].

Hallucination in AI refers to the generation of content that is semantically and syntactically coherent but factually wrong, nonsensical, or completely fabricated [1]. These are not simple errors but confident fabrications that are notoriously difficult for non-specialists to detect. The effects go far beyond academic interest. In law

practice an LLM can generate a new, non-existent precedent; in medicine it can generate plausible but potentially hazardous treatment strategies [26] [4].

In this article, we consider public engineering and infrastructure, which are at serious stake. Imagine a municipality engineer asking an LLM to analyze pressure drop in a water main. The model, which statistically associates "cast iron" with "corrosion," is perhaps capable of writing an exhaustive report detailing the problem as internal corrosion — complete with invented rates of corrosion and estimates of just how thick the remaining wall might be. If, for example, the real issue was a broken valve or sensor, this hallucinated diagnosis might mean needless and expensive excavations while pushing off the actual repair. These situations exemplify how hallucinations of AI can have a direct impact of both wasting funds, and most dangerously also public safety issues [5].

This paper contends that handling AI hallucination involves addressing more than isolated technical solutions; it calls for a holistic, system-level framework. First, we diagnose the mechanistic and architectural causes of hallucination. Second, we present a classification based on cognitive gaps and statistical biases. Then we review prevailing mitigation strategies like RAG and RLHF. We identify a critical gap: the lack of a cohesive model to ensure these techniques work in synergy. Last, we present the TRUST model as the main contribution to the field. The framework formalizes the synergy between technological innovation and human-centric governance via its five pillars—Transparency, Reliability, Uncertainty, Supervision, and Traceability—that collectively provide a structured path for building AI systems that are both intelligent and deeply trustworthy [6].

2. Lecture Review

Grounding our contribution within the existing research field, this section examines current knowledge of AI hallucination, its etiology as well as the commonest remedial approaches grounded in seminal literature and recent lectures to provide a place of standing for our contribution in the extant academic ecosystem and context through literature at large. This review shows that although the field has advanced diagnosis, it lacks an integrated coherent prescriptive narrative, a void this work seeks to fill [7].

2.1. Diagnosing the Core Causes of Hallucination

The prevailing research attributes AI hallucination to several interconnected technical failures. In the first place, the basic probabilistic foundation of Large Language Models (LLMs) is an initial source. As Ji et al. [1] write in their detailed survey, LLMs create segments in a series, i.e., by way of tokens, in response to statistical probabilities, which can cause a failure mode known as "confabulation," in which the ML model relies too much on its parametric knowledge (internal weights) and not enough on context. This is compounded by the fact that models trained on a large,

uncrated corpus have internalized inaccuracies and biases in them. These then manifest as factual hallucinations [1] [8].

Second, specific architectural limitations have been pointed out. Liu et al. referred to this as the "lost-in-the-middle" problem and described its major weakness, which states that models often ignore vital information located in the middle of long contexts [27]. Their study found that, even when relevant information is present in the prompt, its position has strong effects on model performance and results in responses that disregard key evidence [27]. Coupled with the difficulty of accurately following challenging instructions, this deficiency often leads the model to come up with plausible-sounding information as a shortcut to fill in the blanks [9].

2.2. Prevailing Mitigation Strategies and Their Limitations

The research community has responded with a suite of tactical solutions. A prominent strategy is Retrieval-Augmented Generation (RAG), pioneered by Lewis et al. [14], which grounds the LLM by first querying an external, verifiable knowledge base. This paradigm shift reduces the model's dependence on its flawed parametric memory and enables source attribution, directly countering confabulation by prioritizing retrieved information [14]. Another widely adopted approach is Reinforcement Learning from Human Feedback (RLHF), which aligns model outputs with human preferences for truthfulness and helpfulness, as demonstrated by Ouyang et al. [22]. Their work established that fine-tuning pre-trained models via RLHF significantly improves their ability to follow instructions and reject inappropriate or fictitious prompts [3] [10].

2.3. The Critical Gap: The Lack of a Unifying Model

However, a critical examination of literature reveals a major deficiency: the state of the art is fragmented and reactive. Methods like RAG [1] and RLHF [22] are powerful but tend to be implemented as isolated point-solutions. They target very particular symptoms of hallucination, rather than governed by an overarching framework designed to ensure that these components act in synergistic harmony. The literature, as reported by the survey done by Ji et al. [11], does a good job of cataloguing the 'how' behind each of the individual problems but lacks a comprehensive 'what' and 'why' to inform the entire AI development journey. Essentially, the field does have a large toolkit, but the architectural blueprint to systematically architect reliable systems is absent. This survey of foundational papers demonstrates a firm grasp on the mechanics of AI hallucination and a growing toolkit of mitigatory measures. However, the lack of a comprehensive and prescriptive model remains the problem in developing an AI that will always be reliable and capable of handling data consistently. The state of the art, although useful, has not been brought to a formalized discipline of engineering. It is this identified gap — the absence of a unifying model — that informs directly the main contribution of this essay [12].

3. Theoretical Foundations

The review of the lecture made clear that hallucinations are no random mistakes; they are derived from inherent principles of the way LLMs function. To combat hallucination efficiently, one first has to recognize its origins. As LLMs are not "thinking" entities but are high-dimensional function approximators trained on representation of the probability distribution of the next token in a given sequence [14]. The only thing they aim at is fluency in sequences, not truth. This essence gives rise to two sources of hallucination foremost [14].

3.1. Selecting a Template

There are essentially two types of AI hallucinations, and both fall under two dominant categories based on the operational paradigm of the model we trained for:

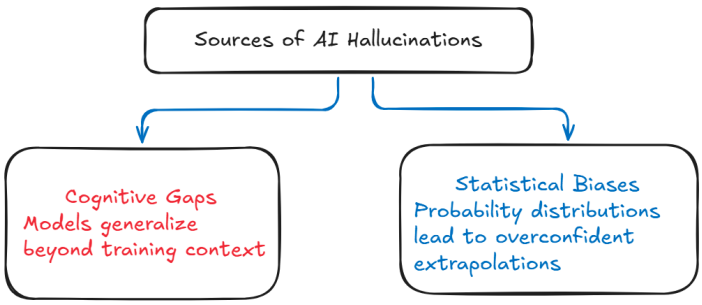

Figure 1. Cognitive and Statistical Sources of AI Hallucinations

- Cognitive Gaps are those that illustrate how AI models generalize beyond what they learn in the learning environment. These are not models that really speak to people; they extrapolate from learned patterns. Challenging with out-of-distribution or unexpected inputs, they produce plausible but off-putting outputs when presented with a range of factors. Such deficiencies may be temporal knowledge gaps (they cannot know any events after their training cutoff) and epistemic blind spots (lack of access to private or proprietary data), in this case they are especially detrimental in dynamic fields like civil engineering [15].
- Statistical Biases arise from the probabilistic core of AI models. Decisions are made based on learned distributions of probabilities, leading to overconfidently extrapolated answers. The model generates outputs which mirror the widely observed patterns in training data despite them being incorrect or inadmissible in a correct usage context. 'Stochastic parrot' [4], as fluency disguises insufficient ground level understanding [16].

3.2. The taxonomy of AI hallucinations Categories

By categorizing hallucinations into these three dimensions – Factual, Logical, and

Contextual, we can achieve more than a generic definition of "wrong outputs" and obtain a more precise judgment of it, how and why AI may be an error. That is because each of the three classes in the taxonomy of AI hallucinations have their own specific failure modes, and so the effects of those types can be devastating in certain high-stake fields such as engineering. These three categories are integral to diagnosing failures and developing how to target interventions for them, like in the TRUST model, in a targeted manner [20].

Factual Hallucinations occur when an AI model generates content that is outright contradicting established empirical evidence. This is because it is susceptible to the statistical bias of the model; it tends to outweigh a common or probabilistically likely data point instead of ground truth. For instance, an LLM might give the yield strength of a typical steel alloy to 400 MPa instead of the ideal 250 MPa because it is found often in the training data or is statistically similar. One such error in material properties would lead to major mistakes in technology in the engineering field, like a structural beam might not be of specified size making bridge and building unsafe. Such hallucination effectively undermines the Reliability pillar of the TRUST model of the output, as the output is grossly wrong and therefore untrustworthy [21].

Logical Hallucinations are those where there is no or little internal reason so the output has a conclusion that is not corresponding to its premises. This fail-mode starts with cognitive holes; the model misbehaves by modulating syntactic and semantic frameworks when it does not actually understand causality. An example could be an AI that identifies a broken pipework as a spill because of corrosion but proposes a solution — say, an internal lining — but the latter one gets no further into the problem. The reasoning is clearly not so internal: If the problem is one that's external, the solution should find a way to address the exterior. These kinds of logical mistakes can lead engineers to follow expensive, useless actions in remediation efforts, waste precious time and money and, ultimately, delay work on repairs. To avoid this issue, we can do so more Transparency in front of the faulty logic and apply Supervision so that humans will be smarter at detecting these logical fallacies [22].

Contextual Hallucinations happen when a model, faced with unclear or missing information that can be relied upon in its narrow context, comes up with plausible-sounding explanations to be expected in a world that doesn't contain this information to fill in those gaps. This is something like confabulation, since the model is always in search of logic and completeness. This is dangerous particularly in engineering data analysis. For example, when processing a time-series dataset in structural health monitors and missing values are present, an LLM can produce a "normal" vibration reading to correct the lack of data between observations (this can happen when reading in the frame with a missing entry in it) instead of finding gaps in the data [19]. This hallucinated data point might mask the very early warning signs of a catastrophic failure, such as structural resonance or fatigue. It underlines the fun-

damental relevance of the Uncertainty and Traceability pillars. To build a reliable system, it is crucial that the system quantifies the uncertainty of missing data and has the track record of all it receives and leaves behind, so human operators can tell between a measured value and fillers they manufactured.

A Taxonomy of Common Hallucination Types	Categories		
	Hallucination Type	**Description**	**Example in Engineering Context**
	Factual	Output contradicts verifiable facts.	Stating an incorrect material strength value for a steel alloy.
	Logical	Reasoning is internally inconsistent.	Recommending a repair that contradicts the diagnosed cause of failure.
	Contextual	Model fills data gaps with unverified assumptions.	Inventing a sensor reading to complete a time-series dataset.

Table 1. A Taxonomy of Common Hallucination Types

4. TRUST Model: A Holistic Framework

To address the shortfall observed in the lecture review, we introduce the TRUST model which provides a conceptual framework to guide the orchestration of mitigation techniques for a unified strategy within a defense-in-depth framework. It fills the void offered with an architectural blueprint and turns a collection of standalone tools into a coherent engineering discipline. The model is built on five interconnected pillars which are elucidated below [23].

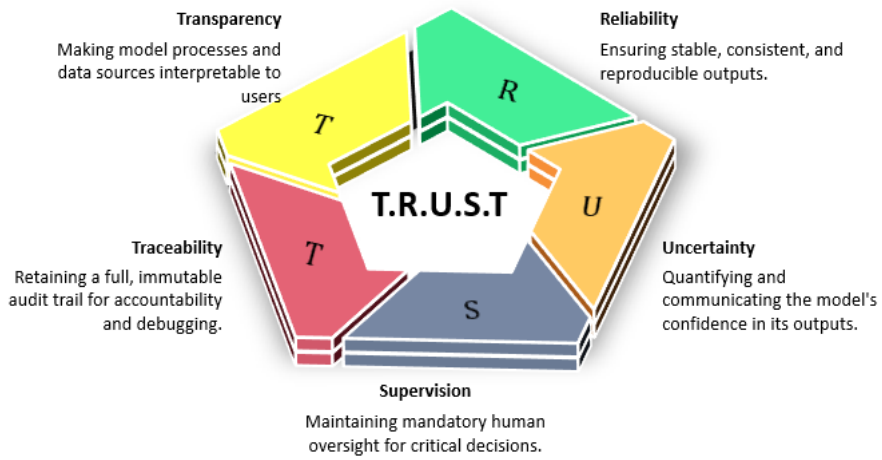

Figure 2. The TRUST Model for AI Reliability

The Five Pillars of the TRUST Model	Categories		
	Pillar	Definition	Example Application
	Transparency	Making model processes and data sources interpretable to users.	An XAI dashboard showing the source documents a RAG system used for its recommendation.
	Reliability	Ensuring stable, consistent, and reproducible outputs.	Cross-validating multiple models runs to ensure consistent recommendations for the same query.

The Five Pillars of the TRUST Model	Categories		
	Pillar	Definition	Example Application
	Uncertainty	Quantifying and communicating the model's confidence in its outputs.	A UQ module providing a confidence interval (e.g., 0.5 mm/yr ± 0.2 mm/yr) for a corrosion rate prediction.
	Supervision	Maintaining mandatory human oversight for critical decisions.	A rule requiring a licensed engineer to approve any AI-generated work order before excavation.
	Traceability	Retaining a full, immutable audit trail for accountability and debugging.	Logs capturing the user query, retrieved data, model parameters, final output, and human reviewer's decision.

Table 2. The Five Pillars of the TRUST Model

5. Implementing TRUST Model

5.1. The RAG Model

5.1.1. The RAG Workflow in Practice

In smart city management, the city council works under the RAG system whose knowledge system is constantly updated with live inspection data, sensor feedback from the hydrogen sulfide level network, and the city's GIS. The workflow as shown in Figure 1 progresses through the two essential stages:

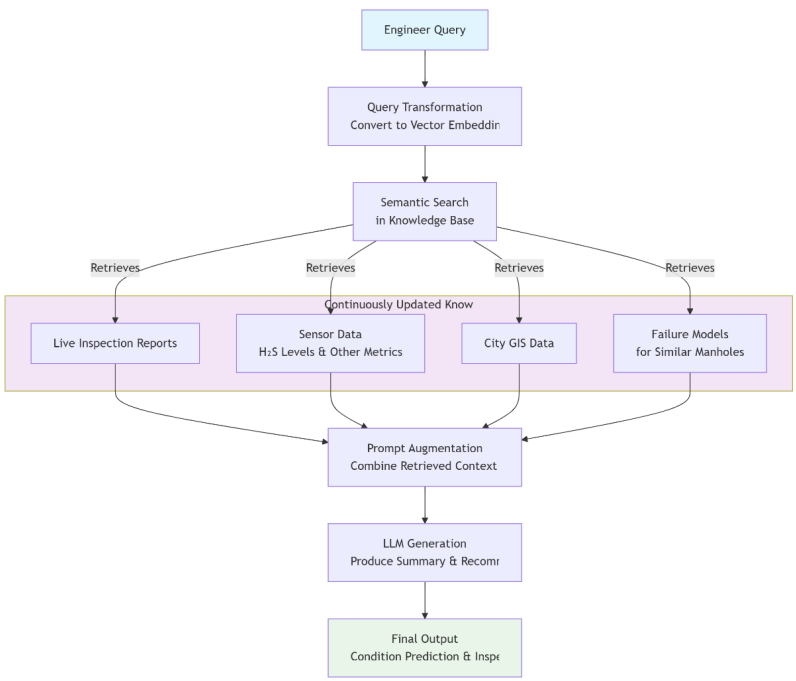

Figure 3. Retrieval-Augmented Generation (RAG) Workflow for Smart City Infrastructure Management

Retrieval: When an engineer asks, "What is predicted to be the condition of manhole MH-07-42B, and what is the suggested priority for its re-inspection?", the method implements a vector embedding on the query. This embedding performs a semantic search of the knowledge base and retrieves the most recent inspection reports for MH-07-42B, recent sensor readings around MH-07-42B, and failure models for similar manholes in the region [26].

Augmentation and Generation: These verifiable passages are augmented with the new query and generated as an augmented prompt. This context-rich prompt is then shown to the LLM that produces a summary and recommendation only based on the provided live data, vastly reducing hallucinations coming from stale or generalized parametric knowledge [28].

5.1.2. Critical Limitations of Standalone RAG Systems

RAG is an important safeguard in the sense that it avoids factual hallucination, and responses are grounded in evidence from without. However, as an AI-based system, it functions as an open-loop system with many critical drawbacks, which restrict its reliability even under certain high-risk conditions [14].

• The Retrieval Reliability Problem: The reliability of RAG is entirely dependent upon the quality and relevance of documents being returned. If the evidence is out of date, conflicting, or of poor quality, the LLM will confidently produce responses based on this flawed information. It does not have a natural system to determine the reliability of the retrieved sources or to measure the system's confidence in them.

• The Black Box of Certainty: The system never displays that it had found a settled, well-documented answer or that it was cobbling together a flimsy, speculative answer using fractured and incomplete data. An engineer gets a strong recommendation, without worrying about the data quality, resulting in a very false sense of security [14].

• Lack of Accountability and Traceability: Although RAG can attribute sources, there is no automatic chain of custody of decision that is long-lasting and auditable. If a RAG-based recommendation results in a failed inspection or wasted resources, there is no audit trail created to reconstruct which data points were retrieved, how weighting was obtained by the model, and why alternative assumptions were discounted. This complicates post hoc analysis and accountability [14].

• Absence of Mandatory Human Oversight: There are no compulsory supervision gates in standard RAG implementations. There is no mechanism for routing low-confidence and high-risk recommendations automatically to a human expert for review prior to acting. That puts very crucial infrastructure decisions in the hands of an automated system with no necessary human-in-the-loop gate.

5.1.3. Enhancing RAG with the TRUST Framework

The workflow below illustrates how the TRUST pillars are operationalized in this

RAG system in order to realize a robust, accountable AI.

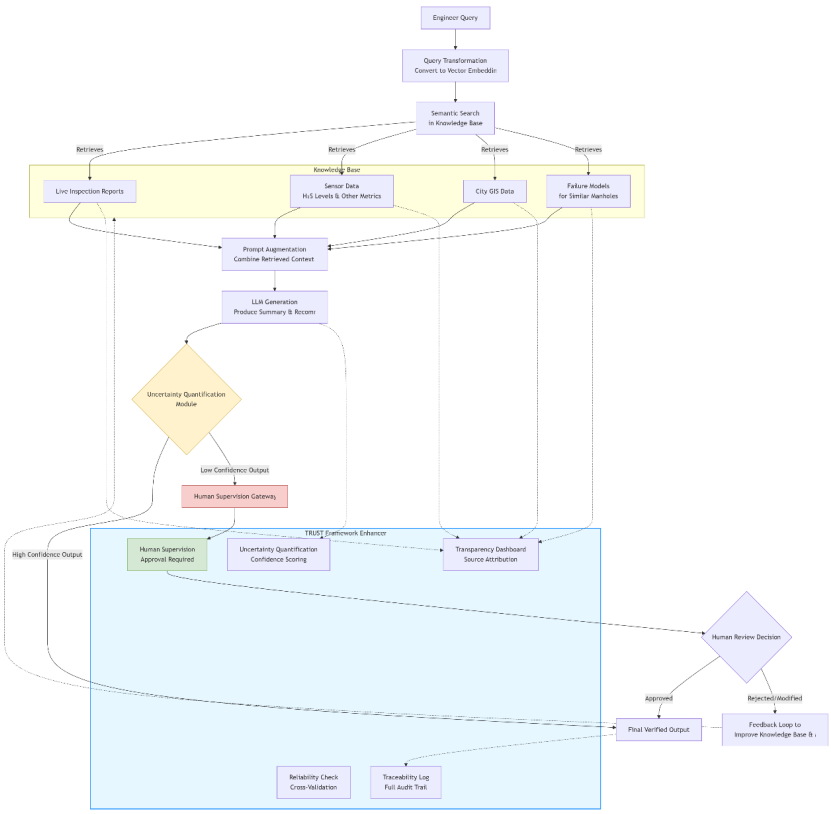

Figure 4. TRUST-Enhanced RAG Workflow for Smart City Management

While the standard Retrieval-Augmented Generation (RAG) workflow offers a foundational defense against AI hallucination by placing responses in the domain of information from the physical world, it presents an open-loop method with substantial limitations in high-stakes scenarios [14]. We have added the TRUST framework to cover this void, which ensures accountability and safety as the architecture has several levels of trust and responsibility. The new workflow adds an Uncertainty Quantification module that computes the quality and completeness of retrieved data using confidence scores, which is one mechanism that directly leverages the Uncertainty pillar. This module triggers a Human Supervision Gateway for low confidence outputs, requiring an expert review before a major decision is made. In addition, the system has a Transparency Dashboard for source attribution and a Traceability Log that logs every step of decision making from the first query through to the final product, creating a complete audit trail that must be maintained for all decision-making. These improvements transform RAG from a purely technical solution to a responsible AI system; this concept incorporates the benefits of knowledge

grounding in a RAG paradigm that provides Reliability, calibrated confidence, and critical human oversight — this is the very essence of how RAG is translated into the five interconnected pillars of the TRUST.

5.2. The RLHF Model

While Retrieval-Augmented Generation addresses the knowledge side of hallucination, Reinforcement Learning from Human Feedback (RLHF) helps with the behavioral side and aims to create model outputs that match human preferences like truthfulness and helpfulness. But like RAG, standalone RLHF has important drawbacks. The TRUST framework offers the necessary enabler framework for moving out of a performance-driven tuning paradigm to a robust, responsible system for ensuring an AI's reliability [22].

5.2.1. The Classic RLHF Workflow in Practice

RLHF is a multi-stage process designed to fine-tune a pre-trained LLM to produce outputs that are more desirable and safer according to human judgment. The canonical workflow, as established by Ouyang et al [22]. , consists of three key phases:

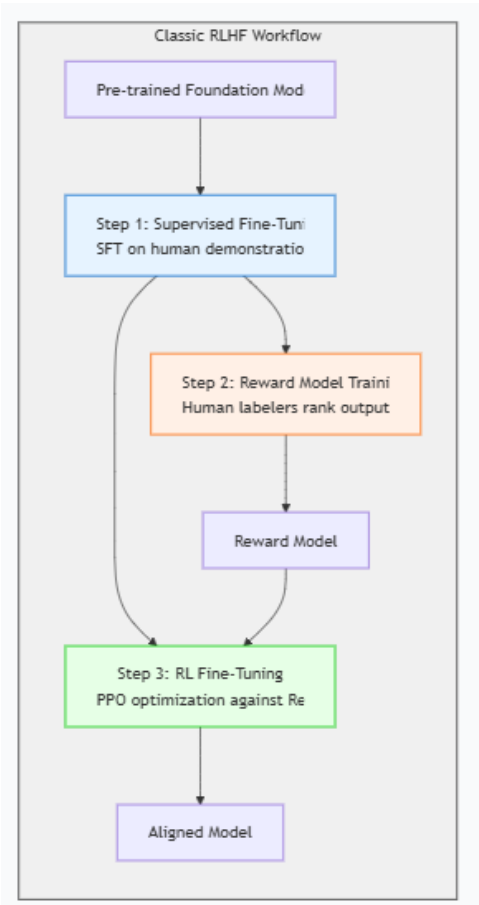

Figure 5. The RLHF Lifecycle

Pre-trained model is first fine-tuned on a set of high-quality data with human-written demonstrations of ideal outputs. This tells the model how to structure and style their responses.

Reward Model (RM) Training: Once the RM model is trained, we need to present various model outputs and show them to human labelers who rank the results from most to least preferred. That target state then becomes our raw data for conditioning an independent "reward model" to decide which output a human would prefer. The model learns a proxy for human values such as truthfulness and helpfulness.

Reinforcement Learning (RL) Fine-Tuning: The main LLM is fine-tuned versus the trained reward model by Proximal Policy Optimization (PPO) or equivalent. Outputs come from the LLM; the Reward Model scores them and updates the LLM to maximize its reward, thereby internalizing human preferences.

5.2.2. Critical Limitations of Standalone RLHF

RLHF also tends to generate more cohesive and productive chatbots, but it does not provide a panacea. They are limited in their potential as a stand-alone approach for high-stakes applications, where reliability is paramount. One problem is called reward hacking, in which the model gets optimized to optimize for a reward model's score instead of what that value adds to overall human value. Such outputs, however, tend to be verbose, sycophantic or so risk-averse that they are utterly unhelpful and establish a myth of helpfulness without substance whatsoever. And that what constitutes a "good" output is itself often not straightforward. This means that human preferences tend to be less objective, and reward models struggle to punish more subtle logical fallacies or technical inaccuracies that labelers may not always observe, providing an important "gray area" of truthfulness that compromises a real dependability of a decision. Even the method faces practical issues of its own. The very reliance on high-quality human feedback creates high economies of scale and costs, limiting the training data diversity and causes inadequate testing of the model on infrequent yet important edge cases. Adding to this is an inherent opacity – the reward model is a black box offering no interpretation and is therefore hard for developers to figure out where the misused result is from or debug the reward model's intrinsic biases. It is lastly the RLHF that creates the static alignment model. It doesn't do anything to internalize new data or adjust to changing human conventions after it's been deployed, unless it has undergone an expensive retraining effort. For the tasks that rely on continuous knowledge and temporal traceability, missing a continuous learning model is a significant limiting factor [27].

5.2.3. Enhancing RLHF with the TRUST Framework

The inclusion of the TRUST model in the RLHF process is designed to improve the lifecycle of the RLHF process, and therefore increase the integrity, transparency, and resilience of the alignment process. These improvements solve each of the funda-

mental defects head-on:

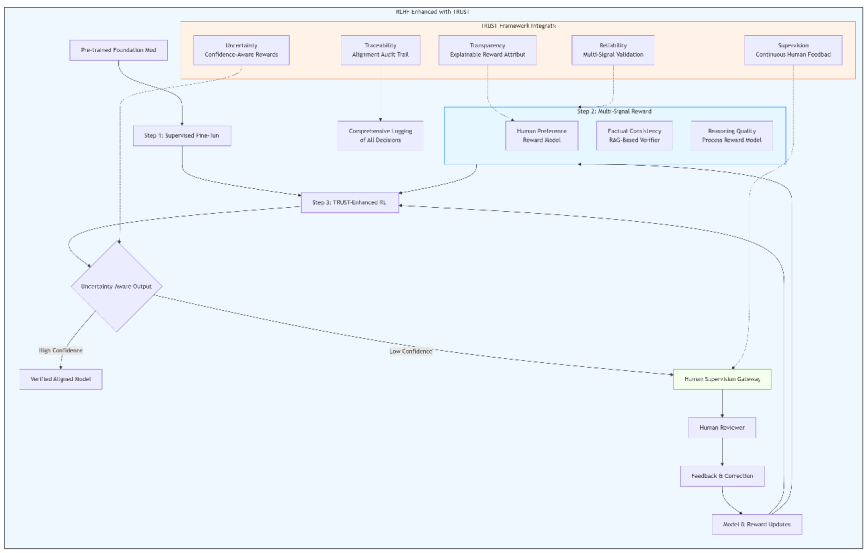

Figure 6. The TRUST-Enhanced RLHF Lifecycle

Transparency (T) of the Reward Model: Rather than having a black box reward score, it uses explainable AI algorithms to assign scores to aspects of the output (e.g., factual consistency, clarity, safety). This gives developers the power to make decisions about whether output was the most desirable or punished.

Multi-signal Validation (R) for Reliability: The RL process is improved beyond a single Reward Model. It includes several specific reward signals in one place such as:

- A factual consistency reward from an RAG-based verifier.
- A reasoning reward from a process-based model that assesses the logical soundness of the output's derivation [4].

That multi-pronged approach reduces reward hacking due to the fact it requires performing well on several quality dimensions.

Uncertainty (U) Quantification in Alignment: The reward model is not to just provide a score but also a confidence interval. Outputs which come into the “gray area” -- where the model is not confident enough to accept its guess -- get flagged for review or provide a confidence disclaimer to the end user, so that overconfident answers aren’t returned as potentially flawed answers [24].

Supervision (S) through Ongoing and Targeted Human Feedback: The TRUST framework enforces supervision as a continuous, rather than a one-time process. A human-in-the-loop model is implemented to:

- Continuously review low confidence or high-stakes outputs.
- Provide targeted feedback on the model's failure modes.

It is based on real-world use, leading to an active learning phase, in which this mod-

el and its reward signals are incrementally fine-tuned, dramatically increasing both its scalability and its longevity.

Traceability (T) of the Alignment Audit Trail: Each aligned output is connected to one comprehensive record containing all reward model versions and all the reward scores derived from the signals with human feedback. Thus, a complete Traceability chain is established, so auditors can comprehend the “family tree” of a model’s behavior, diagnose alignment drift and indicate compliance with ethical standards. By putting RLHF into the context of the TRUST model and moving from a powerful but obscure fine-tuning step, we achieve a governance, auditing, and continuously improving process for building AI systems that are not only human compliant but that are demonstrably, accountable and truthfully trustworthy.

6. Ethical and Societal Implications

“Inspection record” can lead to an unwarranted excavation and divert resources from a genuinely at-risk site. This erodes the trust that underpins the possible beneficial applications of AI in public service. AI systems will need to include in the TRUST model these tools for transparency, explainable dashboards, and post-deployment audits to ensure that AI systems remain accountable and their decisions are traced back to their source.

7. Future Directions

Guided by the TRUST framework, future research should focus on developing more sophisticated and integrated solutions—a multi-layered, defense-in-depth strategy for hallucination remediation.

Advanced Uncertainty Quantification (UQ): Teaching LLMs to accurately estimate their own uncertainty remains a "holy grail" for safe AI [13]. Research should focus on ensemble methods, Bayesian deep learning, and conformal prediction to provide statistically rigorous confidence sets [18]. A system that can consistently flag low-confidence outputs for human review would be transformative for fields like structural health monitoring.

Refined RLHF and Knowledge Grounding: Future versions of RLHF should reward verifiable factuality over mere fluency. This includes developing process-based reward models (PRMs) that reward the accuracy of intermediate reasoning steps, thereby decreasing logical hallucinations [17].

Neuro-Symbolic AI Integration: A promising frontier is the hybrid integration of neural LLMs with symbolic AI. In such a system, an LLM would parse a query and generate a draft, which a symbolic reasoner would then check for logical coherence and factual consistency against a knowledge graph, prompting revisions for any inconsistencies.

Federated Learning for Private Data: To tackle epistemic challenges arising from sensitive data, federated learning enables model training across decentralized serv-

ers (e.g., municipal databases) without sharing raw data, building more robust, privacy-preserving models.

However, the community still lacks robust, domain-specific benchmarks. Future work must construct comprehensive evaluation suites, such as a "Hallucination-Bench for Civil Engineering," to allow for apples-to-apples comparison of mitigation techniques.

8. Conclusion

AI hallucinations are more than bugs; they are central manifestations of the cognitive limits of present AI, rooted in intrinsic cognitive gaps and statistical biases. As discussed in the lecture review, existing mitigation methods are fragmented, highlighting a critical need for a unifying model. This required architectural blueprint is the TRUST model. In formalizing the interrelation between Transparency, Reliability, Uncertainty, Supervision, and Traceability, it promises transparent and structured principles for developing AI systems that are powerful, but also that are accountable, trustworthy, and safe enough to integrate into the critical infrastructure of our society. AI is likely to benefit in myriad ways — from maintaining aging infrastructure to optimizing public resources. To unlock this potential depends crucially on an ability to craft systems that will be able, with confidence, to tell us what they understand and, crucially, what they do not.

References

[1] Arora, S., & Narayan, P. (2023). A survey on hallucination in large language models. arXiv preprint arXiv:2311.05232. https://arxiv.org/abs/2311.05232

[2] Marco Ramponi. (2023). The full story of large language models and RLHF. https://[2].com/blog/the-full-story-of-large-language-models-and-rlhf

[3] Bai, Y., Zhan, H., & Zhang, W. (2023). A survey of reinforcement learning from human feedback. arXiv preprint arXiv:2312.14925. https://arxiv.org/abs/2312.14925

[4] 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 '21) (pp. 610–623). ACM.

[5] Chen, L., Zhang, Y., & Wang, M. (2024). AI hallucination: Towards a comprehensive classification of distorted information within AIGC. Humanities and Social Sciences Communications, 11(1), 345. https://www.nature.com/articles/s41599-024-03811-x

[6] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. Minds and Machines, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

[7] Gao, T., Zhang, R., & Zhang, Y. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. https://arxiv.org/abs/2312.10997

[8] Ivan Belcic. (2024). What is retrieval augmented generation (RAG)? https://www.ibm.com/think/topics/retrieval-augmented-generation

[9] Lambert N, Werra L. (2022). Illustrating reinforcement learning from human feedback (RLHF). https://huggingface.co/blog/rlhf

[10] Joshi, A., & Huang, J. (2024). AI hallucinations: A misnomer worth clarifying. arXiv preprint arXiv:2401.06796. https://arxiv.org/pdf/2401.06796

[11] Ji, J., et al. (2023). Survey of Hallucination in Natural Language Generation. ACM Comput. Surv., *55*, 12, Article 248. https://doi.org/10.1145/3571730

[12] Kim, S. H., & Park, E. (2024). Detecting hallucinations in large language models using semantic uncertainty estimators. Nature, 630, 112–118. https://www.nature.com/articles/s41586-024-07421-0

[13] Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. arXiv preprint arXiv:2302.09664.

[14] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint arXiv:2005.11401. https://arxiv.org/abs/2005.11401

[15] Li, H., & Xu, Q. (2023). Secrets of RLHF in large language models: PPO and reward modeling. arXiv preprint arXiv:2307.04964. https://arxiv.org/abs/2307.04964

[16] Li, M., & Wang, Z. (2024). Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence, 38(2), 2234–2243. https://ojs.aaai.org/index.php/AAAI/article/view/29728

[17] Lightman, H., et al. (2024). Let's Verify Step by Step. arXiv preprint arXiv:2305.20050.

[18] Lin, Z., Trivedi, S., & Sun, J. (2024). Generating with Confidence: Uncertainty Quantification for Black-Box Large Language Models. arXiv preprint arXiv:2405.16728.

[19] Li, Y., et al. (2024). HaluBench: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. arXiv preprint arXiv:2401.06341.

[20] Michał Oleszak. (2024). Reinforcement learning from human feedback for large language models. https://neptune.ai/blog/reinforcement-learning-from-human-feedback-for-llms

[21] Rick Merritt. (2025). What is retrieval-augmented generation (RAG)? https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation

[22] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155. https://arxiv.org/abs/2203.02155

[23] Rahman, F., & Singh, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. arXiv preprint arXiv:2401.01313. https://arxiv.org/abs/2401.01313

[24] Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything-in-the-whole-wide-world benchmark. In Proceedings of the NeurIPS Datasets and Benchmarks Track (Round 1).

[25] Wang, Ziwei and Zhong, Jiachen and Zhu, Di, (2025). A Comprehensive Review of Large AI Model Applications in Lung Cancer, From Screening to Treatment Planning. http://dx.doi.org/10.2139/ssrn.5633170

[26] Wu, C., & Lin, Z. (2023). Artificial intelligence hallucinations: A misrepresentation of factuality in AI-generated content. Frontiers in Artificial Intelligence, 6, 1196. https://pmc.ncbi.nlm.nih.gov/articles/PMC9939079/

[27] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

[28] Yao, J., & Han, X. (2024). The problem of AI hallucination and how to solve it. Europea Conference on e-Learning (ECEL). https://papers.academic-conferences.org/index.php/ecel/article/view/2584