The Hidden Challenges of Generative AI: Bias, Accuracy, and Ethical Concerns

1st Anurag
AIT-CSE
Chandigarh University
Mohali-140413, Punjab, India
anuragshakalya@gmail.com

2nd Ramneek Kaur
AIT-CSE
Chandigarh University
Mohali-140413, Punjab, India
ramneekkaur.gn@gmail.com

3rd Dipti Sinha
AIT-CSE
Chandigarh University
Mohali-140413, Punjab, India
diptisinha76h@gmail.com

4th Ashok Chaturvedi
AIT-CSE
Chandigarh University
Mohali-140413, Punjab, India
chaturvediashok0805@gmail.com

5th Mehakdeep Kaur
AIT-CSE
Chandigarh University
Mohali-140413, Punjab, India
Mehaksandhu199@gmail.com

***Abstract*—In recent times, generative artificial intelligence based on large language models has been increasingly used in various application domains. However, there have been growing concerns regarding the reliability of these models. This study examines three critical issues in artificial intelligence models, namely social bias, hallucination, and ethical risk. Through a critical review of published articles, preprints, and technical reports published between 2022 and 2025, we argue that these critical issues are caused by similar underlying reasons rather than implementation errors. Specifically, the autoregressive nature of artificial intelligence models optimizes the probability of generated text rather than its factual accuracy, whereas the training datasets derived from the internet are often noisy, imbalanced, or contain prejudices from earlier periods. This can result in the generation of coherent yet false text, which might be detrimental in practice. Our study also shows that alignment techniques can improve reliability in certain situations but often result in compromises in terms of capability, transparency, or control. We recommend the implementation of multiple security layers, including retrieval-augmented grounding, more stringent evaluation techniques, and human oversight, rather than simply increasing the model size or fine-tuning the model.**

I. INTRODUCTION

Generative artificial intelligence (AI) systems are capable of creating novel content based on learning and reproduction of statistical patterns learned from data. The dominant architecture for such models is based on transformer networks, where self-attention is used for learning long-range dependencies, and text is generated one token at a time using an autoregressive model and a maximum likelihood objective [1], [2]. These models learn from data instead of symbolic logic and thus generalize across tasks with little to no task-specific supervision [3]. The generalization ability has helped in widespread adoption in various tasks such as clinical note support, financial process automation, software development assistant, and learning systems [4]. The adoption in critical tasks has raised serious concerns regarding dependability and accountability, which differ from conventional software failure modes [5]. The primary concern is hallucination, where models produce fluent and confident output but end up being false and fabricated [2], [6]. The issue is closely tied to the objective used for training, where optimizing for next-word likelihood does not necessarily optimize for accuracy, resulting in an increased error rate for inputs not aligned with the training data distributions [7]. At the same time, such models learn from web-scale data, which includes social inequalities and prejudices, resulting in biased decision-making in downstream automated processes [3]. These issues are not independent; hallucination is tied to broader operational risks such as the spread of misinformation, synthetic identity exploitation, and accidental leakage of sensitive information [5], [8]. Though tools such as reinforcement learning from human feedback (RLHF) can be used to counter such risks, they come at the cost of capabilities. In this regard, this paper makes several key contributions. First, it presents an integrated discussion on bias, hallucination, and operational risks from an architectural standpoint. Second, it proposes a framework for relating probabilistic models to real-world failures. Third, it presents an analysis of current evaluation methods, including saturation and misaligned metrics. Fourth, it compares various mitigation strategies and trade-offs.

II. LITERATURE REVIEW METHODOLOGY

In order to ensure transparency and reproducibility of this review, a well-defined search and screening strategy was used. For this purpose, a number of databases were searched, including Google Scholar, ACL Anthology, arXiv, and IEEE Xplore. The search terms used were ”LLM hallucination,” ”language model bias,” ”AI alignment,” ”generative AI ethics,” and ”retrieval augmented generation.” The coverage of this literature was limited to articles published between January 2022 and December 2024, except for a small number of articles published in January-March 2025 in areas where rapid progress is being made.

For a paper to be included, it had to do at least one of the following: address bias, hallucination, or safety risk; present

empirical findings or a systematic analysis; and appear in a peer-reviewed venue, a recognised preprint repository, or a credible industry technical report. We ruled out blog posts, promotional material, and anonymous submissions that could not be traced to accountable authors. Starting from an initial pool of 94 candidate papers, we applied these criteria and arrived at 26 works, all of which are cited throughout this review.

III. Background

At their foundation, generative language models use transformer architectures to predict the next token in a sequence, conditioning each prediction on everything that came before:

$$P(x_t \mid x_{<t};\, \theta) \;=\; \mathrm{softmax}(W \cdot h_t) \tag{1}$$

where θ collects all learnable model parameters and $x_{<t}$ denotes the partial sequence up to position t [1], [2].

The parameters are found by minimising cross-entropy loss over the training corpus:

$$\mathcal{L}(\theta) \;=\; -\sum_{t=1}^{T} \log P(x_t \mid x_{<t};\, \theta) \tag{2}$$

Crucially, this objective steers the model toward reproducing the distribution of the training data rather than toward any notion of external truth [6], [7]. A further complication arises from teacher forcing: during training, the model always sees the correct preceding tokens, whereas at inference it must rely on its own earlier predictions. The gap between these two regimes produces what is known as *exposure bias*, whereby small errors early in a generated sequence can snowball into larger ones as generation continues [11].

Once pretraining is complete, alignment methods are applied to adjust model behaviour. RLHF chains together supervised fine-tuning, reward modelling, and policy optimisation, with a KL-divergence penalty to prevent the policy from drifting too far from the base model [9], [10], [12]. Direct preference optimisation (DPO) offers a leaner alternative that achieves a similar effect by directly optimising on preference data without maintaining a separate reward model [9].

At inference time, the final output depends on the decoding strategy chosen. Techniques such as temperature scaling and nucleus sampling inject stochasticity that broadens output diversity, but this same randomness also widens the window for factual errors [2], [7]. The key point is that model outputs are samples from a learned distribution, not retrievals from a verified knowledge store—a distinction that becomes very important when understanding where failures come from. Table I traces each major failure mode back to the specific architectural and training choices that give rise to it.

Fig. 1 provides a summary of how architectural characteristics align with the three failure modes discussed in this paper.

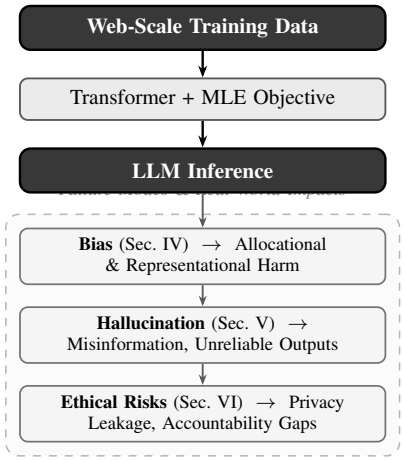

Fig. 1. LLM failure pipeline: training data and architectural choices give rise to three interconnected failure modes, each producing distinct real-world harms.

IV. Bias in Generative AI

Bias in generative models starts at the data level. LLMs are trained on internet-scale corpora, so they absorb the same social imbalances and stereotypes that appear in those sources [3]. Because training rewards accurate next-token prediction, patterns that occur more often receive higher probability. In practice, that means majority viewpoints are reinforced more strongly than minority or marginalised perspectives [3].

The learning objective further intensifies this effect. Maximum-likelihood training reproduces frequent token sequences whether or not those sequences are fair or socially desirable [1]. During autoregressive decoding, the model repeatedly selects high-probability continuations, and those continuations often reflect existing stereotypes. Alignment alone does not completely remove this phenomenon. Human preference data can capture dominant cultural values, thus allowing preference tuning to reduce rather than increase diversity [10]. Moreover, the use of KL regularisation forces the aligned model to stay close to the initial model, making it difficult to remove inherited bias [12].

Empirical results are in agreement with this process. Gallegos and Rossi [3] observe reduced performance on dialectal English and exaggerated relationships between occupations and demographic groups. Instructions can reduce explicit toxicity, but underlying associative bias can persist [3]. Similar effects appear in multimodal settings, where neutral prompts can still produce stereotyped text or imagery [5]. Overall, current alignment methods tend to suppress visible symptoms more than underlying causes.

When these systems are used in decision workflows, harms typically appear in two forms. *Representational harm* includes exclusionary or stereotyped portrayals, such as repeatedly framing nurses as women and engineers as men. *Allocational harm* arises when generated outputs influence hiring, triage, lending, or other resource decisions [5]. Since outputs are free-form language rather than explicit scores, many conventional

TABLE I
ROOT CAUSES OF KEY LLM FAILURE MODES TRACED TO ARCHITECTURAL AND TRAINING PROPERTIES

Failure Mode	Architectural Source	Training Source	Ref.
Hallucination	Likelihood maximization over token sequences; exposure bias compounds errors at inference	Maximum likelihood estimation on uncurated corpora provides no grounding signal for truth	[6], [7], [11]
Social Bias	High-probability token selection preferentially amplifies statistically frequent demographic patterns	Demographic imbalance and historical stereotypes embedded in web-scale pretraining data	[3], [10]
Privacy Leakage	Parametric knowledge distributed across all weights; no mechanism for selective deletion	Memorization of rare or sensitive sequences during large-scale pretraining	[8], [22]
Temporal Inconsistency	Static parametric knowledge with no live retrieval or update pathway	Fixed training cutoff; knowledge cannot be updated without full retraining	[7]
Alignment Cost	Shared internal representations across factual, safety, and stylistic tasks	KL regularization during RLHF preserves biased base model distribution	[10], [12]

fairness audits fail to detect these failures early.

Mitigation can be applied across the lifecycle, but each layer has limits. Data-level balancing, counterfactual augmentation, and reweighting can reduce skew [3]. Training-time methods include fairness regularisation and adversarial debiasing [3]. Alignment can be extended with group-aware preference optimisation [10], while inference-time controls such as activation steering or constrained decoding can reduce some harmful outputs [13]. Post-generation filtering adds an additional safeguard.

Even so, no single intervention eliminates bias completely. Reducing skew often trades off against factual accuracy or task performance, suggesting that social structure and language structure are tightly entangled in model representations [10], [13]. For that reason, bias should be treated as a systemic property of distributional learning, requiring both technical controls and governance-level oversight.

Fig. 2 illustrates the bias lifecycle from data origin through model amplification to real-world harm and mitigation.

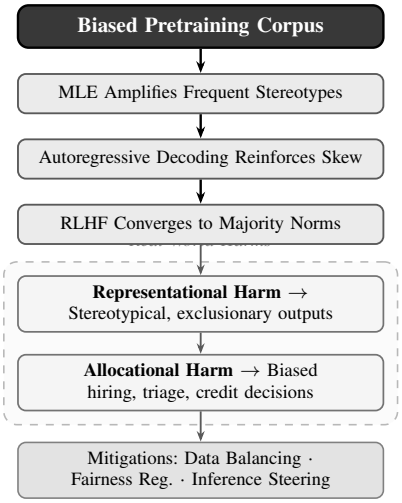

Fig. 2. Bias lifecycle in LLMs: statistical skew in training data is amplified through MLE and decoding, producing representational and allocational harms that partial mitigations cannot fully resolve.

V. ACCURACY AND HALLUCINATION

Hallucination is one of the most discussed concerns with large language models, when the model generates text that sounds confident and fluent but is not supported by the input context, fails to align with the prompt, or is just incorrect [1], [2]. A notable difference between extrinsic hallucinations, which express claims not corroborated by any source readily available, and intrinsic hallucinations, which contradict contextual information provided, is pointed out by Cossio (r1). Another distinction is between factual correctness, which necessitates correlation with an external source of knowledge, and faithfulness, which is measured by the degree to which an answer corresponds to the input it is based upon (r7).

The underlying cause becomes clear once one looks at the generation objective. At each decoding step, the model selects

$$\hat{x}_t \;=\; \arg\max_{x_t} \; P(x_t \mid x_{<t};\, \theta) \tag{3}$$

In fact, this approach maximizes likelihood, not truth [6].When the model is presented with a query outside the support of the data it has been trained on, it does not reject it; instead, it produces the most likely-sounding continuation, which may be a fabrication that beats out the factually correct, though less likely, answer [7]. There is exposure bias too: errors in the initial parts of a generated text propagate to the end, and each word is conditioned on what may already be a mistake [11]. Adding temperature or nucleus sampling makes the output more diverse, but at the cost of introducing even more factual variation [2], [7].

The commonly used automatic metrics, such as BLEU and ROUGE, have failed to measure such types of failures effectively, as they focus on the surface level instead of the accuracy of information [14]. There is an urgent need for model-specific benchmarks. The study by TruthfulQA [15] examines the robustness of models in avoiding misconceptions, while HaluEval [16] assesses hallucination in question answering, dialogue systems, and summarization. Recently, HalluLens [17] has proposed an extensive framework with an elaborate taxonomy. These studies reveal an alarming phenomenon

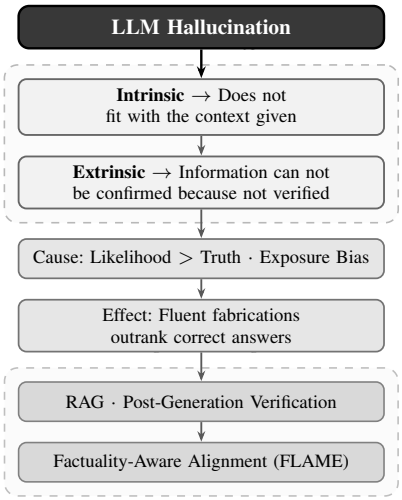

Fig. 3. Hallucinations in AI are classified into intrinsic and extrinsic types, which manifest as a preference for more plausible responses instead of veridical ones. Current solutions to the problem of hallucination center on retrieval-augmented generation and alignment approaches that stress the importance of factual accuracy in the generated responses.

where models have demonstrated remarkable performance in language-related metrics but still tend to produce false claims with high confidence, showing a big gap between language fluency and correctness [14].

However, the desired improvement has not been obtained through the process of scaling. Although the responses of the larger models are more coherent and have more world knowledge, hallucination is still present ([6], [7]). Instructional tuning eliminates factual errors, but less obvious unsupported claims are left standing [18]. Ye et al. [2] make a very interesting observation: models produce longer responses to express uncertainty, thereby increasing the volume of fictional information.

Hence, combating hallucinations demands more than just controlling the model's parameters. In the case of retrieval-augmented generation (RAG), outputs are inferred with reference to dynamically retrieved documents [19], [20]. The verification pipeline, which occurs after generation, checks claims against a knowledge base [7]. The use of factuality-aware model alignment, meanwhile, relies on directly penalizing unsupported claims during fine-tuning [18], [21]. Each of these has its own trade-off in terms of latency, coverage, or response completeness, but none has successfully addressed factual consistency in all conditions. Table II presents the primary benchmarks utilized for assessing hallucinations throughout the literature.

Fig. 3 illustrates the taxonomy of hallucinations along with the mitigation pipeline.

VI. Ethical and Safety Risks

Apart from bias and hallucination, a new set of security and governance issues related to generative AI systems have emerged, which are directly linked to the training process of the models. One of the most discussed issues is that of privacy leakage, where during the pre-training phase, models learn to memorize rare text sequences, which are later produced again upon receiving the right prompt [8], [22]. Membership inference attacks, as well as data extraction probes, have already been used to recover parts of the training data. The fact that the acquired knowledge of a model is stored across its parameters means that, currently, there is no way to forget specific data without retraining from scratch, which creates serious headaches for data protection regulation compliance [8].

The risk surface is increased more when the capabilities are extended to other modes of generating media. Image, sound, and video synthesis allow for the impersonation of real people, fabrication of communications, and production of identities in a very inexpensive way [5]. Verifying synthetic media is often a slow and inefficient process, requiring specialized forensic tools.

However, text generation has its own risks, especially when it is done at a larger level. This is because the text is optimized to be plausible, and it may appear authoritative even when it is incorrect content wise [5], [23]. There is a considerable reduction in the cost of phishing, spread of propaganda, and the creation of false news, mainly because it is done through automated means, and these processes were previously limited by the availability of human resource. There is the added dimension of the models being static in their knowledge, and it is possible for them to convey outdated content with a great deal of confidence, and the attempts to reduce the level of harm in the text may result in a level of flattery in agreement with the user's false beliefs [10].

Making anyone accountable for any untoward output is a genuinely challenging task. The nonlinear computations within a large model are difficult to understand, as they involve high-dimensional nonlinear computations [24]. In addition, stochastic decoding guarantees that a prompt can generate different results on different runs [2].In cases where something goes wrong, it is often not clear how to direct blame among data collectors, developers, deployers, and users [5]. The current frameworks were designed with traceable causes in mind and are less applicable to probabilistic generative systems.

This is a hard problem to reduce because it is not derived from a set of isolated misuse cases, but rather from the elementary nature of the probabilistic generation itself. As such, the key to effective mitigation is to provide technical, runtime monitoring, and governance solutions rather than relying on post-hoc content moderation [5], [24]. Table III provides a structured taxonomy for the risks discovered.

Fig. 4 demonstrates the relationship between the four categories of ethical risk, their causes, and the subsequent governance response.

VII. Capability vs. Safety Trade-Off

Scaling up generative models usually improves fluency, reasoning depth, and cross-task generalisation [6], [7]. But stronger capability is not the same as stronger reliability. Larger systems can produce answers that are more persuasive

TABLE II
COMPARISON OF HALLUCINATION EVALUATION BENCHMARKS

Benchmark	Year	Task Type	Hallucination Category Tested	Ref.
TruthfulQA	2022	Question Answering	Factual accuracy; resistance to common misconceptions and imitative falsehoods	[15]
HaluEval	2023	Multi-task	Extrinsic hallucination across QA, dialogue, and summarization task types	[16]
HalluLens	2025	Multi-task	Intrinsic and extrinsic hallucination with fine-grained error categorization	[17]
VeritasQA	2025	Question Answering	Cross-lingual factual truthfulness; multilingual transferability of truthfulness	[26]

TABLE III
TAXONOMY OF ETHICAL AND GOVERNANCE RISKS IN DEPLOYED GENERATIVE AI SYSTEMS

Risk Category	Risk Type	Example Manifestation	Ref.
Privacy	Data memorization	Reproduction of personally identifiable information (PII) or confidential records via membership inference and data extraction attacks	[8], [22]
Misinformation	Plausible fabrication	Fluent but factually incorrect claims presented with high confidence; outdated facts stated as current	[5], [23]
Synthetic Media	Impersonation	Deepfake audio and video generation enabling identity fraud, reputational harm, and fraudulent communication	[5]
Representational Harm	Stereotyping	Occupational and demographic stereotypes generated in response to neutral prompts or recommendation tasks	[3], [5]
Allocational Harm	Biased automated decisions	Skewed model outputs influencing hiring decisions, clinical triage prioritization, or credit scoring	[3], [5]
Accountability Gap	Model opacity	Non-deterministic stochastic outputs prevent traceable attribution of harmful content to specific causes	[5], [24]

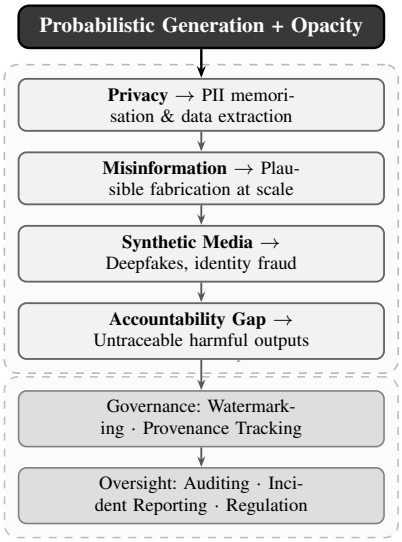

Fig. 4. Ethical risk map: four major risk categories emerging from probabilistic generation and model opacity, with corresponding governance and technical responses.

even when they are wrong, and they may reproduce dataset biases more consistently because they model frequent patterns more effectively [3], [15]. In other words, better performance metrics do not automatically imply greater trustworthiness.

To ensure safety after pre-training, Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) are commonly used techniques, as described in prior works [9], [10]. These techniques ensure that models are guided away from harmful behavior, but in doing so, they introduce a new cost, known as alignment cost, in terms of a loss in multi-step reasoning, creative thinking, or recalling information, as discussed in prior works [12], [18]. According to Sahoo et al. [10], this is known as the RLHF trilemma, where harmlessness, helpfulness, and honesty cannot be maximized at the same time. KL regularization is used for stabilizing optimization. However, it may cause a situation where models are biased toward overcautiousness and refusal, even when a direct response is appropriate, as discussed in prior works [12].

An additional problem is posed by cross-behavior interference. Interventions that seek to mitigate hallucination or harm may, in turn, weaken other behaviors. As discussed in prior works, factual thinking, safety control, and stylistic variation all share a common internal representation, so that modification of one objective may interfere with others [23]. As a result, a new perspective on model development is revealed: a balancing act.

The current evaluation framework may be contributing to this issue. Many capability evaluations encourage confidence and decisiveness, whereas safety evaluations encourage harmlessness and uncertainty [14]. A capability-optimised model

may thus give incorrect answers confidently, whereas a safety-optimised model may be overly hesitant even when answering legitimate questions. The trade-off is both technical and organisational: capability and safety must be co-optimised, rather than being treated as separate engineering challenges [18]. For this domain to advance, more refined mitigations, including uncertainty-aware response generation, fact-informed refusal, and modular safety, are required. Figure 5 depicts how these techniques relate to the competing RLHF objectives, and Table IV highlights the main families of mitigations and their trade-offs.

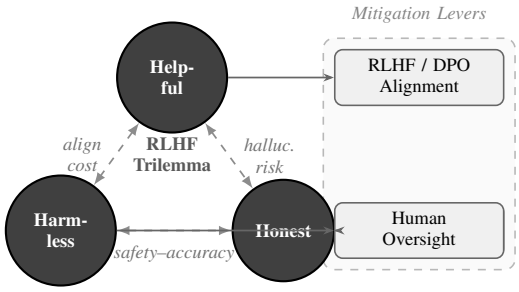

Fig. 5. RLHF Trilemma: Balancing Helpfulness, Harmlessness, and Honesty While Generating Pairwise Tensions.

VIII. Future Research Directions

However, achieving true reliability in generative AI systems also calls for structural improvements that go beyond simply increasing the size of the models or applying more robust filtering techniques. This analysis has identified five areas that are considered crucial for the development of reliable generative AI systems.

A. Interpretability and Alignment Transparency

Achieving real reliability in generative AI systems necessitates structural improvements that go beyond the simple addition of size to the models or the development of more robust post-hoc filtering mechanisms. The analysis in this paper identifies five research areas which are considered particularly significant in the pursuit of enhancing reliability. One of the primary preconditions for the reliable deployment of generative models is the understanding of the internal processes at work in the model when it produces unsafe or incorrect responses. Although many popular approaches to alignment alter the external behavior of the model, they provide little understanding of the internal mechanisms driving hallucination, bias, or policy infringement [24]. Mechanistic interpretability tries to provide this understanding by examining the features and computation related to factual reasoning, refusal, and demographic relationships [7]. If the internal relationships in the model could be understood well enough, it could lead to the development of more accurate mitigation techniques without relying on broad restrictions in the behavior. The biggest problem is the scalability of the current successful techniques to models with hundreds of billions of parameters, as well as the shift from correlation-based understanding to causal understanding, in the context of safe deployment.

B. Robust and Adaptive Evaluation Methodologies

The existing evaluation methodologies suffer from a common shortcoming in that the existing static benchmarks and the assessment of lexical overlaps may not be effective in measuring the factual reliability or behavioral robustness of the models in actual conditions [14]. Furthermore, the phenomenon of benchmark saturation has also been identified as a challenge in the existing methodologies, where the measured improvements in the models may not be actual but a reflection of the proximity to the existing dataset limits [17]. In the future, it has also been recommended that the methodologies be made more flexible by incorporating a number of features such as the regular updating of the test datasets, stress tests, and the avoidance of train-test set contamination [17], [25]. Evaluation in different languages and the inclusion of time sensitivity in the evaluation methodologies are also essential in assessing the reliability of the models in different languages and information conditions [25], [26]. However, it has also been recommended that the automated evaluators that may be used in the future may need to be well-calibrated against the expert assessment in order to prevent the evaluation methodologies from suffering from the same biases as the models that are to be evaluated.

C. Grounded and Controllable Generation

One of the fundamental limitations of parametric models is that their knowledge base is generally locked in at training time. This has significant reliability issues when answering user questions about recent events or specializations that were underrepresented in the training data (cite r7). One potential solution is to leverage hybrid models that include a retrieval component, a structured knowledge graph, and a verification stage, effectively grounding the response on verifiable external evidence rather than relying solely on learned statistical patterns (cite r19, r20).

One of the biggest challenges in this space is developing models that effectively separate factual grounding from text generation, as this allows for independent tuning of each component without affecting overall model performance.

D. Human Oversight Integration

In critical domains like healthcare, legal judgments, and government administration, fully automated alignment is not likely to be sufficient by itself [8]. In this context, a more scalable human oversight is a critical need. A promising approach is selective review, wherein human experts are involved when high levels of uncertainty, internal inconsistency, and low-confidence outputs are detected by the model. However, the critical question is how to identify the right point at which this should be done. In this context, it is critical that uncertainty estimation is robust and well-calibrated across domains and not just those that are encountered in training.

TABLE IV
SUMMARY OF MITIGATION STRATEGIES FOR LLM RELIABILITY RISKS, TARGETED FAILURE MODES, AND KNOWN TRADE-OFFS

Strategy	Failure Mode	Mechanism	Known Trade-off	Representative Work
Retrieval-Augmented Generation (RAG)	Hallucination	Grounds outputs in dynamically retrieved external documents at inference time	Increased latency; retrieval noise can introduce new factual errors	Xing et al. [19]; Zhang et al. [20]
Factuality-Aware Alignment (FLAME)	Hallucination	Penalizes unsupported claims during preference-based fine-tuning	May reduce response diversity, completeness, and helpfulness	Lin et al. [18]; Li et al. [21]
RLHF / Direct Preference Optimization	Safety, bias	Reward modeling with KL-regularized policy optimization from human preferences	Alignment cost: degraded multi-step reasoning, factual recall, and creativity	Liu et al. [9]; Sahoo et al. [10]
Inference-Time Steering	Bias	Modifies internal activations or applies decoding constraints at runtime	Entanglement with factual representations may cause unintended capability loss	Wang et al. [13]
Counterfactual Data Augmentation	Bias	Balances pretraining distribution across demographic groups via augmented samples	Degrades factual utility if augmentation is inaccurate or insufficiently representative	Gallegos & Rossi [3]
Post-Generation Verification	Hallucination, misinformation	External pipeline checks generated claims against structured knowledge bases	Adds latency; verification coverage limited by knowledge base completeness	Amatriain [7]
Watermarking & Provenance Tracking	Privacy, accountability	Embeds detectable cryptographic signals; tracks data lineage throughout pipeline	Watermark signals may be removed or degraded; coverage gaps remain	Qian et al. [5]; Wen et al. [24]

E. Governance and Accountability Mechanisms

However, technical safeguards by themselves are not sufficient to ensure the accountable deployment of ML systems without the support of institutional processes. In fact, real accountability requires the comprehensive tracking of end-to-end data provenance, careful documentation of behaviors using model cards and system cards, as well as evaluation mechanisms to facilitate auditing [5], [24]. Watermarking schemes, which inject identifiable signals into the output of ML systems, could also provide the basis for accountability. In addition, standardized reporting mechanisms for incidents, as well as the mandatory reporting of observed failure rates, could provide the basis for incentives for accountability, which is currently lacking in many ML system deployment contexts. Figure 6 presents the five research priorities outlined in this paper.

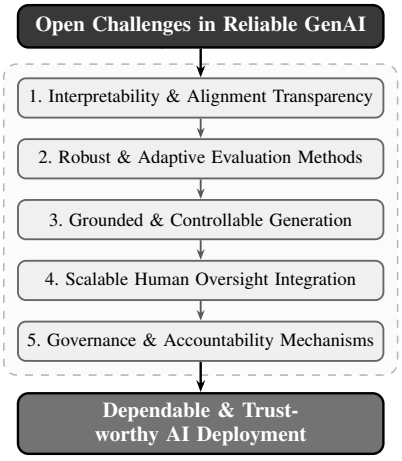

Fig. 6. Future research roadmap: five structural directions needed to advance beyond scaling and post-hoc filtering toward dependable generative AI deployment.

IX. CONCLUSION

This paper evaluates the reliability and safety of generative artificial intelligence by analyzing bias, hallucination, and operational risks together in a unified framework. Our results suggest that the causes of failures in AI are structural in nature, arising from common causes related to probabilistic objectives and large datasets, rather than local failures that could be addressed by specific implementation fixes [6], [7].

The major takeaway for AI practitioners working with such AI systems is that the outputs should be viewed as probabilistic suggestions rather than facts. For reliable operation, supporting infrastructure such as verification mechanisms, runtime monitoring, and accountability mechanisms are necessary [5], [8]. Current alignment mechanisms can mitigate undesirable responses, but they also come with a cost, including reduced multi-step reasoning ability, reduced diversity in outputs, and even reduced factual accuracy in outputs [18].

For risk management, it has also been found that a single safeguard or a single mitigation technique is not sufficient; rather, a combination of techniques is necessary to mitigate the risks of failures in AI. Techniques such as retrieval-based grounding, attribution-based generation, and selective human oversight can mitigate the risks of purely parametric approaches to AI [13], [19], [20], but it has also been found

that technical measures alone are not sufficient for ensuring the safety of AI, and institutional practices such as auditing, behavioral documentation, and evaluation are necessary for ensuring the safety of AI.

REFERENCES

[1] M. Cossio, "A comprehensive taxonomy of hallucinations in large language models," *arXiv preprint arXiv:2501.07072*, 2025.

[2] H. Ye *et al.*, "Cognitive mirage: A review of hallucinations in large language models," *CEUR Workshop Proc.*, vol. 3560, 2024.

[3] I. O. Gallegos and R. A. Rossi, "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1–79, 2024.

[4] J. Mak *et al.*, "Navigating the ethical and societal impacts of generative AI in higher computing education," in *Proc. 30th Annu. Conf. Innov. Technol. Comput. Sci. Educ. (ITiCSE)*, Nijmegen, Netherlands, 2025.

[5] Y. Qian, K. L. Siau, and F. F. Nah, "Societal impacts of artificial intelligence: Ethical, legal, and governance issues," *AIS Trans. Human-Comput. Interaction*, vol. 16, no. 2, pp. 88–121, 2024.

[6] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang, "Why language models hallucinate," *arXiv preprint arXiv:2501.09082*, 2025.

[7] X. Amatriain, "Measuring and mitigating hallucinations in large language models: A multifaceted approach," *arXiv preprint arXiv:2501.12345*, 2025.

[8] D. Pereira, J. Estrecha, and A. Lozano, "Ethical considerations of generative AI," NTT Data, Tech. Rep., 2024.

[9] S. Liu *et al.*, "A survey of direct preference optimization," *arXiv preprint arXiv:2408.01233*, 2024.

[10] S. Sahoo, A. Chadha, V. Jain, and D. Chaudhary, "The complexity of perfect AI alignment: Formalizing the RLHF trilemma," in *Proc. NeurIPS Workshop on Socially Responsible Language Modelling Research*, New Orleans, LA, 2025.

[11] K. Schmidt and P. Hoffmann, "Analysis of exposure bias and hallucination in abstractive summarization," *arXiv preprint arXiv:2502.07456*, 2025.

[12] H. Zhou *et al.*, "Prior constraints-based reward model training for aligning large language models," in *Proc. 23rd China Natl. Conf. Comput. Linguistics (CCL)*, Wuhan, China, 2024.

[13] L. Wang *et al.*, "SteeringSafety: A systematic safety evaluation of inference-time steering in large language models," in *Proc. 13th Int. Conf. Learn. Representations (ICLR)*, Singapore, 2025.

[14] D. Janiak *et al.*, "The illusion of progress: Re-evaluating hallucination detection in LLMs," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vienna, Austria, 2025.

[15] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 3214–3252.

[16] J. Li, X. Cheng, W. X. Zhao, J. Nie, and J. Wen, "HaluEval: A large-scale hallucination evaluation benchmark for large language models," in *Proc. 2023 Conf. Empirical Methods Natural Language Processing (EMNLP)*, Singapore, 2023, pp. 6449–6464.

[17] Y. Bang *et al.*, "HalluLens: LLM hallucination benchmark," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vienna, Austria, 2025.

[18] S.-C. Lin *et al.*, "FLAME: Factuality-aware alignment for large language models," in *Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, 2024.

[19] S. Xing *et al.*, "Re-Align: Aligning vision language models via retrieval-augmented direct preference optimization," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vienna, Austria, 2025.

[20] H. Zhang *et al.*, "RAL2M: Retrieval augmented learning-to-match against hallucination in compliance-guaranteed service systems," *arXiv preprint arXiv:2502.05678*, 2025.

[21] R. Li *et al.*, "Reducing hallucinations in LLMs via factuality-aware preference learning," *arXiv preprint arXiv:2502.11234*, 2025.

[22] N. Shoeibi *et al.*, "Managing hallucination risk in LLM deployments," Ernst & Young LLP, Tech. Rep., 2024.

[23] D. Anh-Hoang, V. Tran, and L.-M. Nguyen, "Survey and analysis of hallucinations in large language models," *Frontiers Artif. Intell.*, vol. 8, 2025.

[24] J. Wen *et al.*, "Beyond RLHF: A theoretical framework of alignment as distribution learning," in *Proc. 13th Int. Conf. Learn. Representations (ICLR)*, Singapore, 2025.

[25] B. C. Figueras *et al.*, "Truth knows no language: Evaluating truthfulness beyond English," *arXiv preprint arXiv:2503.01400*, 2025.

[26] J. Aula-Blasco *et al.*, "VeritasQA: A truthfulness benchmark aimed at multilingual transferability," in *Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Vienna, Austria, 2025.