IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Trustwo Rthiness, Hallucination, And Evaluation In Large Language Models

[1]Mrs. Meera. S. Sawalkar, [2]Qaizar Master, [3]Samruddhi Koratkar, [4] Soumitra Kharate
[1]Assistant Professor, [2]Third Year Student, [3]Third Year Student, [4] Third Year Student
[1]Department of Artificial Intelligence and Data Science,
[1]All India Shri Shivaji Memorial Society's – Institute of Information Technology, Pune, India

Abstract **-** Large language models have moved from research curiosities to systems used by millions, yet questions about whether they can be trusted remain largely unsettled. This review brings together recent work on three closely linked problems: trustworthiness, hallucination, and evaluation. We surveyed fifteen studies published between 2020 and 2025 covering safety alignment, factual reliability, retrieval-augmented generation, benchmark design, and human evaluation methods. The picture that emerges is one of rapid progress paired with stubborn limitations. Hallucination remains common even in frontier models, partly because it is built into how these systems are trained. Evaluation methods often disagree with each other, and benchmarks tend to age quickly as models adapt to them. Trust frameworks are getting richer, but most still treat issues like bias, toxicity, and factual accuracy as separate problems rather than parts of a whole. We argue that future work needs to combine technical metrics with user-centered studies, and that retrieval grounding, while useful, is not a complete fix. The paper closes with a discussion of open challenges and what we think are the more promising research directions.

Index Terms **-** Large Language Models, Trustworthiness, Hallucination, Evaluation, Retrieval-Augmented Generation, AI Safety, Benchmarks.

1. INTRODUCTION

The release of ChatGPT in late 2022 brought large language models (LLMs) into mainstream attention almost overnight. Since then, models such as GPT-4, Claude, Gemini, and LLaMA have been deployed across customer support, healthcare triage, legal research, software engineering, and education. Adoption has outpaced our understanding of when these systems can be relied on. A model that produces confident-sounding text about a medical symptom or a legal precedent is not the same thing as a model that is correct, and the gap between fluency and accuracy is at the heart of much current concern.

Three problems dominate the literature. The first is trustworthiness, which is a broad term covering whether a model behaves safely, fairly, and predictably. The second is hallucination, the well-documented tendency of LLMs to generate content that looks plausible but is factually wrong or unsupported by any source. The third is evaluation, that is, how we actually measure these things in a way that is reproducible and meaningful. These three concerns are not independent. A model that hallucinates frequently cannot be trusted, and we cannot know how much it hallucinates without good evaluation methods.

The stakes are not abstract. There have already been documented cases of lawyers submitting briefs that cited fabricated cases generated by ChatGPT, of medical chatbots giving harmful advice, and of customer service agents being misrepresented by company-deployed bots. Each of these incidents traces back, in some way, to all three of our themes at once. The model produced false output (hallucination), users were not able to tell it was false (trust), and the team that deployed the system did not have evaluation methods sensitive enough to catch the failure mode in advance.

This review consolidates fifteen recent studies that address one or more of these themes. Our goals are modest. We do not try to solve any of the open problems, but we do try to draw connections between threads of research that often appear in separate venues. We focus on work that has shaped current practice, including foundational alignment papers, the major hallucination surveys, key benchmark proposals, and several human evaluation studies. The remainder of the paper is organized as follows. Section 2 provides background on LLMs and the technical context for the issues we discuss. Section 3 reviews the literature on trustworthiness. Section 4 turns to hallucination, its causes, and current mitigation strategies. Section 5 examines evaluation methods. Section 6 offers a comparative discussion of findings, and Section 7 outlines open challenges and concludes.

1.1 Scope and Methodology

Our selection of fifteen primary references was driven by three criteria. First, the work had to address at least one of trustworthiness, hallucination, or evaluation directly, rather than treating these as background motivation for some other contribution. Second, we favored papers that have been influential, measured by citation count and by the frequency with which their methods appear in follow-up work. Third, we tried to maintain a balance between survey papers, which give the broad shape of a subfield, and empirical papers, which provide concrete results. We excluded blog posts, technical reports without peer review, and works that focused only on a single proprietary model without generalizable findings.

Two limitations of this scope deserve mention up front. The literature in this area moves quickly. Several of the papers we cite have already been partly superseded by 2025 follow-up work. We have tried to indicate where this is the case. Also, our review is biased toward English-language venues, which mirrors a broader bias in the field itself. We comment on this in Section 6.

2. BACKGROUND AND CONTEXT

2.1 What Large Language Models Do

Modern LLMs are autoregressive transformer networks trained on enormous text corpora using a next-token prediction objective. After this pretraining stage, most production models go through additional fine-tuning, often involving supervised learning on curated examples followed by reinforcement learning from human feedback (RLHF) [1]. The result is a system that maps input prompts to coherent natural language responses. The transformer architecture introduced by Vaswani and colleagues in 2017 is what made this scaling possible, but the qualitative jump in capability came mostly from sheer increases in parameter count, training data, and compute [2].

It is worth emphasizing what these models are not. They are not databases, and they do not consult facts in any structured way during generation. They predict sequences of tokens that fit patterns learned during training. This distinction matters for everything that follows in this paper. When a model produces a citation that does not exist, or invents a court case, it has not made a lookup error. It has produced exactly what its training would suggest, given the prompt.

A useful way to think about this is to consider what the loss function actually optimizes. Cross-entropy loss on next-token prediction rewards plausible continuations. Plausibility is not the same thing as correctness. A continuation that mentions a famous physicist by name is more plausible than one that says "someone whose name I do not recall," even when the latter is more honest. This bias toward confident output is reinforced during fine-tuning, where helpfulness is often rewarded over caution.

2.2 The Rise of Concerns

Concerns about LLM behavior were not new in 2022. Bender and colleagues had already published their influential critique of large models in 2021, raising questions about training data, environmental cost, and the risks of mistaking fluency for understanding [3]. What changed with ChatGPT was scale of exposure. Suddenly millions of users were interacting with a system that would happily explain quantum field theory or produce a wedding speech, and the gap between perceived and actual reliability became visible to non-experts.

Around the same time, regulators and standards bodies began drafting frameworks for AI risk. The NIST AI Risk Management Framework, the EU AI Act, and various sector-specific guidelines all assume

that some form of evaluation is possible. Whether current evaluation methods are up to that job is one of the questions this review tries to address.

2.3 Terminology

Some terms are used inconsistently across the literature, and we want to be explicit about how we use them. Trustworthiness here refers to the broader property of a model being safe to rely on, including but not limited to factual accuracy. Hallucination refers specifically to outputs that are not supported by inputs or by ground truth. Evaluation refers to the process of measuring properties of a model, whether through automated metrics, human ratings, or some combination. We will note where authors we cite use these words differently.

3. TRUSTWORTHINESS IN LARGE LANGUAGE MODELS

3.1 Defining Trustworthiness

Trustworthiness is harder to define than it first appears. The TrustLLM benchmark proposed by Sun and colleagues in 2024 lists eight dimensions: truthfulness, safety, fairness, robustness, privacy, machine ethics, transparency, and accountability [4]. Other surveys propose slightly different taxonomies. Liu and coauthors split the problem into eight categories as well, but theirs include reliability and resistance to misuse [5]. The fact that researchers cannot quite agree on what trustworthiness means is itself a finding worth noting. It suggests the term has become an umbrella for any concern that does not fit neatly under accuracy or capability.

We can group these dimensions roughly into three families. The first family is about factual reliability: does the model produce true outputs? This overlaps significantly with the hallucination literature discussed in Section 4. The second family is about value alignment: does the model produce safe, fair, ethical outputs? The third family is about predictability and accountability: can users and developers reason about how the model will behave, and can someone be held responsible when it goes wrong? Most current research tackles one family at a time, which is part of the integration problem we return to in Section 6.

3.2 Alignment as a Path to Trust

RLHF, introduced in production form by Ouyang and colleagues in their InstructGPT paper [1], is the most widely deployed alignment technique. The procedure trains a reward model on human preference data, then uses reinforcement learning to fine-tune the base model so it produces outputs the reward model scores highly. RLHF clearly works for some things. It reduces toxic and obviously harmful outputs, it improves instruction following, and it makes models more useful in conversational settings.

But RLHF has also been shown to introduce its own problems. Reward hacking, where the model learns to produce outputs that score well on the reward model without actually being good, is a recurring concern [6]. There is also evidence that RLHF can degrade calibration, meaning that aligned models may express more confidence in their answers even when they are wrong. Bai and colleagues at Anthropic introduced Constitutional AI as one alternative, where instead of relying purely on human feedback, the model critiques and revises its own outputs against a set of written principles [7]. Constitutional methods are promising but rely on the principles being well-written, which is its own difficult problem.

More recent variants such as Direct Preference Optimization (DPO) try to simplify the RLHF pipeline by avoiding the explicit reward model. The reported results are competitive with full RLHF on many benchmarks at lower computational cost. Whether these approaches solve the underlying alignment issues, or just shuffle them around, is a question the community is still working through.

3.3 Beyond Alignment: Bias, Privacy, and Robustness

Trust is not only about avoiding harmful content. Bias in LLM outputs has been documented extensively. Gallegos and coauthors surveyed bias and fairness in LLMs in 2024 and found that models reproduce stereotypes about gender, race, and nationality even after standard alignment training [8]. The bias often shows up in subtle ways. A model may not produce overtly sexist text when asked directly, but its choices when given an open-ended task, like generating a story about a doctor, can still skew along stereotypical lines.

Privacy is another concern. LLMs can memorize and regurgitate training data verbatim, which creates risks when training corpora include personal information [9]. Carlini and colleagues showed that with carefully designed prompts, attackers can extract names, phone numbers, and addresses from production

models. The risk is not theoretical. Several published studies have demonstrated extraction attacks against widely deployed systems. Mitigation typically involves data filtering during training and output filtering at inference, but no current method offers a guarantee.

Robustness, in the sense of consistent behavior under prompt perturbations, also remains weak. A small change in phrasing can flip a model's answer, which is troubling for any deployment that depends on stable outputs. Adversarial prompting research, including jailbreak prompts that bypass safety training, demonstrates that current alignment is shallow in important respects. A model that reliably refuses harmful requests in standard phrasing may comply when the same request is wrapped in a fictional scenario or translated into another language.

3.4 Transparency and Accountability

Even when models behave well most of the time, users and downstream developers often cannot tell why. Mechanistic interpretability research is making progress on understanding what happens inside transformer networks, but for production purposes, current models are largely opaque. This matters for accountability. If a deployed system causes harm, tracing the cause back to a specific training decision or data source is difficult, sometimes impossible. Several frameworks have proposed model cards and system cards as partial solutions, providing structured documentation of capabilities and known limitations [4]. These are useful but voluntary, and they capture only what the developer chose to disclose.

4. HALLUCINATION IN LARGE LANGUAGE MODELS

4.1 What Counts as a Hallucination

The term hallucination is used loosely in the literature, which has caused some confusion. Ji and colleagues, in their widely cited survey, distinguish two main types: intrinsic hallucinations, where the output contradicts the input, and extrinsic hallucinations, where the output makes unsupported claims that go beyond what the input provides [10]. Huang and coauthors offer a more recent and more granular taxonomy, separating factuality hallucinations from faithfulness hallucinations [11]. The point of these distinctions is not academic. Different types have different causes and different fixes.

It is also worth noting that some researchers have argued the word hallucination itself is misleading. The term implies a perceptual error, as if the model briefly saw something that was not there. What actually happens is more banal. The model generates a continuation that is consistent with patterns in its training data but not consistent with reality. Confabulation has been suggested as an alternative, since it captures the idea of confidently producing false but coherent content. We use hallucination here because the term is dominant in the literature, but readers should understand it as a label for a behavior rather than an explanation of one.

4.2 Why Hallucinations Happen

There is broad agreement on several causes. First, training data is imperfect, containing errors, contradictions, and outdated information. Second, models are trained to produce fluent text, and fluency is rewarded even when grounding is absent. Third, there is no internal mechanism that flags when the model has reached the edge of what it actually knows. Kalai and Vempala recently formalized the last point, showing that under standard training objectives, hallucination is in some sense unavoidable, because the next-token prediction loss does not penalize confident wrong answers any differently than confident right ones [12]. This is a sobering result. It suggests that hallucination cannot be eliminated by better training alone.

Beyond these structural causes, hallucination rates vary with the type of task. Open-ended generation, like writing a biographical paragraph about a less famous person, produces hallucination far more often than constrained tasks like answering multiple choice questions. Long-form generation tends to drift, with the model anchored well in fact for the first few sentences and increasingly inventive afterward. Asking the model to cite sources reliably produces fabricated citations, since the model is good at the surface form of academic references but has no way to verify whether a given combination of authors and title corresponds to a real paper.

We should also mention that hallucination is sometimes a feature, not a bug. When users ask a model to write fiction or brainstorm ideas, fabrication is exactly what they want. The problem is that the same mechanism produces invented case law and fake medical advice. The model has no way to tell which mode the user wants, and users do not always say.

4.3 Mitigation Approaches

Mitigation strategies fall into a few categories. The most popular at the moment is retrieval-augmented generation, or RAG, originally proposed by Lewis and colleagues in 2020 [13]. RAG works by retrieving relevant documents at inference time and including them in the prompt, so the model can reference real sources rather than relying entirely on memorized training data. RAG reduces hallucination on factual tasks, sometimes dramatically. But it does not eliminate it. The model can still misread retrieved documents, ignore them, or blend their content with hallucinated additions [14].

Other approaches include uncertainty estimation, where the model is asked to estimate confidence in its own outputs, and self-consistency, where multiple samples are compared. Both have shown some success but neither is a complete solution. Self-consistency in particular suffers when the model has a stable but wrong belief, since multiple samples will agree on the same incorrect answer.

A more recent line of work tries to detect hallucinations after the fact, using either trained classifiers or LLM-based judges [15]. Detection is useful for filtering outputs in production but does not fix the underlying generation problem. Some systems combine multiple approaches: retrieval grounding, followed by self-checking, followed by a separate verification step. The combinations help, but each layer adds latency and cost, and even stacked systems do not reach human-level reliability on harder factual tasks.

4.4 Hallucination as a Deployment Problem

An emerging view in the practitioner literature is that hallucination should be treated less as a problem to solve at the model level and more as a property to manage at the system level. This view accepts that current models will hallucinate at some rate and focuses on building scaffolding around them. Examples include forcing the model to cite specific retrieved passages, blocking outputs that fail a verification check, or routing high-stakes queries to a different system entirely. The shift in framing is significant. It treats LLMs as components rather than as standalone products, which seems closer to how reliable real-world systems get built.

5. EVALUATION OF LARGE LANGUAGE MODELS

5.1 Automated Benchmarks

The most common way to evaluate LLMs is through automated benchmarks. MMLU, introduced by Hendrycks and colleagues in 2021, became something of a default for measuring general knowledge across academic domains [2]. BIG-bench, with its 200-plus tasks, took a broader approach. HELM, the Holistic Evaluation of Language Models framework from Stanford, tried to standardize evaluation across many dimensions including accuracy, calibration, robustness, and bias [6]. These benchmarks have been valuable, but they share well-known weaknesses.

Benchmark contamination is one of the bigger issues. As benchmarks become popular, they leak into training data, and models start scoring well not because they have learned the underlying skill but because they have seen the test questions. Saturation is another problem. Frontier models now score near ceiling on many established benchmarks, which makes the benchmarks less informative for distinguishing top systems. The community has responded by introducing harder benchmarks, but this is a treadmill rather than a solution.

There are also concerns about construct validity. A benchmark on multiple-choice questions about physics might be measuring physics knowledge, or it might be measuring the model's ability to recognize phrasing patterns common in physics textbooks. When models pass benchmarks, we usually want to claim they have the underlying capability, but the evidence for that claim depends on the benchmark being well-designed in ways that are not always carefully checked.

5.2 Hallucination-Specific Benchmarks

Specialized benchmarks for hallucination have emerged. TruthfulQA, by Lin and colleagues, tests whether models repeat common misconceptions [14]. HaluEval offers a larger collection of hallucination examples across multiple task types [11]. FActScore evaluates the factual accuracy of long-form generation by breaking outputs into atomic facts and checking each one [15]. Each of these has trade-offs. TruthfulQA is small and somewhat narrow. HaluEval is broader but its labels rely on automatic methods that may themselves be unreliable. FActScore is more rigorous but expensive to run.

Cross-cutting these benchmarks, an important methodological question is what counts as ground truth. For some questions, the answer is in a reference document or a knowledge graph. For others, the answer depends on what a careful expert would say, and experts disagree. Evaluation that pretends ground truth is unambiguous risks penalizing models for outputs that are actually correct, or rewarding models for outputs that happen to match an imperfect reference.

5.3 Human Evaluation

Human evaluation is often treated as the gold standard, but it has its own difficulties. Inter-annotator agreement is typically low for open-ended tasks like summarization, especially when raters disagree about what counts as a hallucination [10]. Recruiting and training annotators is expensive.

There is also a more philosophical problem: human raters can be wrong, and they can be biased by superficial features like fluency or length. A response that sounds polished may receive higher ratings even if it contains errors that less polished responses lack.

A growing line of work uses LLMs themselves as judges, an approach sometimes called LLM-as-a-judge [14]. This is cheaper than human evaluation and seems to correlate reasonably well with human judgments on many tasks. But the approach inherits whatever biases the judge model has, and it can favor outputs that look like the judge would have produced them. Trusting an LLM to evaluate another LLM, when both share similar failure modes, is not a stable foundation.

5.4 Beyond Static Benchmarks

There is increasing recognition that static benchmarks are not enough. Real users interact with models in messy, context-rich ways that single-turn benchmarks cannot capture. Some researchers have proposed dynamic evaluation, where the test set is regenerated periodically, or red-teaming evaluations, where adversarial prompters try to elicit failures [4]. These methods are more realistic but harder to standardize, which is part of why automated benchmarks remain dominant despite their limitations.

Another approach gaining traction is task-grounded evaluation. Instead of testing whether the model can answer abstract questions, the evaluator measures whether the model helps a real user complete a real task. This is closer to what end users actually care about, but it requires defining the task carefully and recruiting users, which makes it slow and expensive. The trade-off between realism and scalability runs through almost every evaluation choice in the literature.

6. COMPARATIVE DISCUSSION

Looking across the fifteen studies surveyed, a few patterns are worth highlighting. The first is that progress on these problems is genuinely real but uneven. Hallucination rates on factual question answering have dropped significantly since 2020, partly because of RAG and partly because of better training. Alignment techniques have made models noticeably safer. Yet the same models still fail in ways that resemble older failures, and new capabilities tend to bring new categories of failure with them. Tool use, agentic behavior, and longer context windows have each introduced fresh failure modes that the earlier literature did not anticipate.

Table 1 summarizes a representative subset of the works we surveyed, organized by primary focus area, methodological approach, and headline finding. The table is not exhaustive but gives a sense of how the field divides labor across the three themes of this review.

The second pattern is that the three areas we have surveyed are connected in ways that are not always reflected in the literature. Many trustworthiness papers treat hallucination as one item on a list. Many hallucination papers treat evaluation as a methodological detail. We think this fragmentation is a problem.

A trust framework that does not measure hallucination carefully cannot really claim to assess truthfulness. An evaluation method that does not consider robustness or fairness is incomplete. The most useful work tends to be the work that crosses these lines, but it is the minority.

Third, there is a recurring tension between automated and human-centered approaches. Automated methods scale, but they miss things. Human methods catch things, but they do not scale and they introduce their own variability. Hybrid approaches, where automated metrics are validated against human judgments on a sample and then used at scale, seem like the most promising path. Several recent papers have moved in this direction [11], [15].

Ref.	Focus Area	Method	Key Finding
[1]	Alignment	RLHF training	Improves helpfulness
[3]	Critique of LLMs	Position paper	Risks of scale
[4]	Trustworthiness	Multi-dim. benchmark	Eight trust pillars
[7]	Alignment	Constitutional AI	Self-critique works
[10]	Hallucination	Survey	Two main types
[12]	Hallucination	Theoretical	Inevitability result
[13]	Mitigation (RAG)	Architecture	Retrieval helps
[15]	Evaluation	FActScore metric	Atomic fact check

Table 1: *Summary of selected primary references and their contributions*

Fourth, the field has not converged on what good evaluation looks like at the system level, as opposed to the model level. Most benchmarks evaluate a model in isolation. But real deployments involve prompts, retrieval pipelines, output filters, fallback systems, and human oversight. A model that performs poorly on a benchmark may produce reliable outputs in a well-designed system, and a model that benchmarks brilliantly can still fail spectacularly when deployed naively. Bridging the gap between model evaluation and system evaluation is one of the more practically important open questions.

Finally, we note that almost all current work focuses on English and on a relatively narrow band of high-resource languages. Multilingual evaluation is sparse, and what exists tends to find that performance, hallucination rates, and bias all vary across languages in ways that monolingual evaluation will miss [8]. This is not a minor caveat. If LLMs are deployed globally but evaluated mainly in English, we are mostly checking what happens for a fraction of users.

6.1 Illustrative Case Studies

To make the patterns above more concrete, it helps to look at a few well-documented incidents. The first is the 2023 case in which a U.S. lawyer submitted a brief that cited several court decisions invented by ChatGPT. The model produced realistic-looking case names and reporter citations that did not exist. This is a textbook extrinsic hallucination [10]. It is also a trust failure, because the user assumed the output had the same status as a normal legal search. And it is an evaluation failure, because no benchmark in routine use at the time would have flagged this kind of error in a typical workflow.

A second example is the well-known issue of medical advice. Studies of LLM responses to health questions have found that frontier models are correct most of the time but produce dangerous advice in a non-trivial minority of cases.

The fluency of the wrong answers makes them harder to spot. Here the trust dimension is acute, because users seeking medical information are often less able to evaluate the response than the model is to generate it. RAG over curated medical databases helps but does not eliminate the issue [13], [15].

A third example, drawn from the alignment literature, involves jailbreak prompts. Researchers have repeatedly shown that the same model that refuses a direct request for instructions on producing a harmful

chemical may comply when the request is framed as fiction or wrapped in a code-like syntax [4]. This is a robustness failure that is also a trust failure, because users and operators cannot easily predict where the safety boundary actually sits.

6.2 What the Pattern Suggests

Reading these incidents alongside the survey literature, one impression is that the failures are rarely surprising in retrospect. The mechanisms that produce them have been described in the academic record. What is missing is the operational discipline of treating the academic findings as relevant to deployment decisions. There is a gap between knowing, in general, that LLMs hallucinate and refusing to deploy them in a high-stakes context where this property would cause harm. Closing that gap is partly a research problem and partly a question of professional practice.

7. OPEN CHALLENGES AND CONCLUSION

7.1 Open Challenges

Several questions remain open. First, can we develop evaluation methods that do not become obsolete as soon as models adapt to them? Adaptive or contamination-resistant benchmarks are a partial answer, but the deeper issue is how to measure something as fuzzy as trust without reducing it to a single number that can be gamed. Second, can hallucination be reduced to acceptable levels, or is it a fundamental property that must be managed through deployment design rather than fixed at the model level? Recent theoretical work suggests the latter [12].

Third, how should we handle the trade-offs that come with safety training? Safer models are often more cautious, and over-cautious models are less useful. Finding the right point on this curve is partly a research question and partly a values question, which means it cannot be resolved by technical work alone. Fourth, who decides what counts as good behavior? Current practice involves a small number of labs setting norms that affect millions of users worldwide. There are obvious legitimacy concerns here that the technical literature mostly does not address.

Fifth, evaluation infrastructure has not kept up with capability. The field has poured resources into training larger models. The corresponding investment in evaluation has been much smaller. Better tooling, more diverse benchmarks, better support for human-in-the-loop assessment, and reproducible reporting standards would all help. None of this is glamorous work, but the field cannot make confident claims about progress without it.

7.2 Conclusion

This review has surveyed fifteen recent contributions to the literature on trustworthiness, hallucination, and evaluation in large language models. We have argued that these three problems are tightly linked and benefit from being studied together. Significant progress has been made in each area, but each also has stubborn open questions that will not be solved by incremental improvements. Retrieval grounding, alignment techniques, and better benchmarks all help. None is a complete answer.

For practitioners building systems on top of LLMs, the practical implication is that defense in depth matters. No single technique will guarantee trustworthy behavior. Combining alignment, retrieval, output filtering, human oversight on high-stakes outputs, and ongoing evaluation is more likely to produce reliable systems than relying on any one approach. For researchers, the most fruitful work seems to be at the intersections, where alignment, hallucination measurement, and evaluation methodology meet. We hope this review helps others navigate these intersections.

ACKNOWLEDGEMENT

The authors thank their guide for ongoing feedback throughout this work, and the faculty of the department for providing the academic environment in which this review took shape. We also acknowledge the broader research community whose openly published work made a survey of this kind possible.

REFERENCES

[1] L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback," in Advances in Neural Information Processing Systems, vol. 35, pp. 27730-27744, 2022.

[2] D. Hendrycks, C. Burns, S. Basart, et al., "Measuring massive multitask language understanding," in Proceedings of the International Conference on Learning Representations (ICLR), 2021.

[3] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "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), pp. 610-623, 2021.

[4] L. Sun, Y. Huang, H. Wang, et al., "TrustLLM: Trustworthiness in large language models," in Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

[5] Y. Liu, Y. Yao, J.-F. Ton, et al., "Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment," arXiv preprint arXiv:2308.05374, 2023.

[6] P. Liang, R. Bommasani, T. Lee, et al., "Holistic evaluation of language models (HELM)," Transactions on Machine Learning Research, 2023.

[7] Y. Bai, S. Kadavath, S. Kundu, et al., "Constitutional AI: Harmlessness from AI feedback," arXiv preprint arXiv:2212.08073, 2022.

[8] I. O. Gallegos, R. A. Rossi, J. Barrow, et al., "Bias and fairness in large language models: A survey," Computational Linguistics, vol. 50, no. 3, pp. 1097-1179, 2024.

[9] N. Carlini, F. Tramer, E. Wallace, et al., "Extracting training data from large language models," in Proceedings of the 30th USENIX Security Symposium, pp. 2633-2650, 2021.

[10] Z. Ji, N. Lee, R. Frieske, et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1-38, 2023.

[11] L. Huang, W. Yu, W. Ma, et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Transactions on Information Systems, vol. 43, no. 2, pp. 1-55, 2025.

[12] A. T. Kalai and S. S. Vempala, "Calibrated language models must hallucinate," in Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC), pp. 160-171, 2024.

[13] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[14] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3214-3252, 2022.

[15] S. Min, K. Krishna, X. Lyu, et al., "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 12076-12100, 2023.