A Survey on Trustworthiness and Reliability in Large Language Models

www.surveyx.cn

Abstract

The survey paper explores the critical aspects of trustworthiness, reliability, and ethical considerations in large language models (LLMs), highlighting challenges such as hallucinations, uncertainty, and moral implications. Hallucinations, where models produce factually incorrect outputs, are a significant concern, influenced by biases and model architecture. Techniques like SELF-FAMILIARITY and BatchEnsemble show promise in mitigating these issues, enhancing reliability and interpretability. The paper emphasizes the need for robust evaluation frameworks, incorporating methods like Checkability Training to improve output legibility and foster trust. The integration of retrieval-augmented generation (RAG) and fine-tuning strategies further enhances model performance by grounding outputs in factual data. Ethical considerations, including privacy, bias, and fairness, are critical, necessitating adherence to regulatory standards and the development of culturally sensitive models. The survey underscores the importance of interdisciplinary approaches to address trust-related challenges, advocating for advancements in calibration, uncertainty estimation, and model adaptability. Future directions include refining evaluation benchmarks, exploring adversarial training, and enhancing ethical frameworks to align LLMs with societal values. Overall, the paper provides a comprehensive overview of current challenges and potential solutions, aiming to improve the reliability and ethical deployment of LLMs across diverse applications.

1 Introduction

1.1 Significance of Trustworthiness and Reliability

The deployment of large language models (LLMs) across various sectors necessitates a thorough evaluation of their trustworthiness and reliability. These AI systems are susceptible to generating incorrect or misleading information, known as hallucination, which significantly undermines their credibility and acceptance [1]. The potential for LLMs to propagate disinformation further complicates their use, as misinformation can spread rapidly across platforms [2]. Additionally, inherent security and privacy vulnerabilities in LLMs highlight the urgent need for robust frameworks that ensure these models are secure and ethically aligned.

In sensitive domains, such as educational content for children, the reliability of LLMs is paramount. The generation of inappropriate or misleading content in such contexts can have severe consequences [3]. User trust in LLMs is heavily influenced by perceptions of accuracy and ethicality, which are essential for fostering confidence in these systems [4]. The framing of LLM outputs and the transparency of provided explanations significantly affect user trust, particularly in creative and decision-making scenarios [5].

Enhancing the trustworthiness of LLMs requires addressing biases and improving the interpretability of outputs, critical for their widespread acceptance and integration [6]. The relative scarcity of comprehensive literature in computer vision underscores the need for systematic reviews to bridge

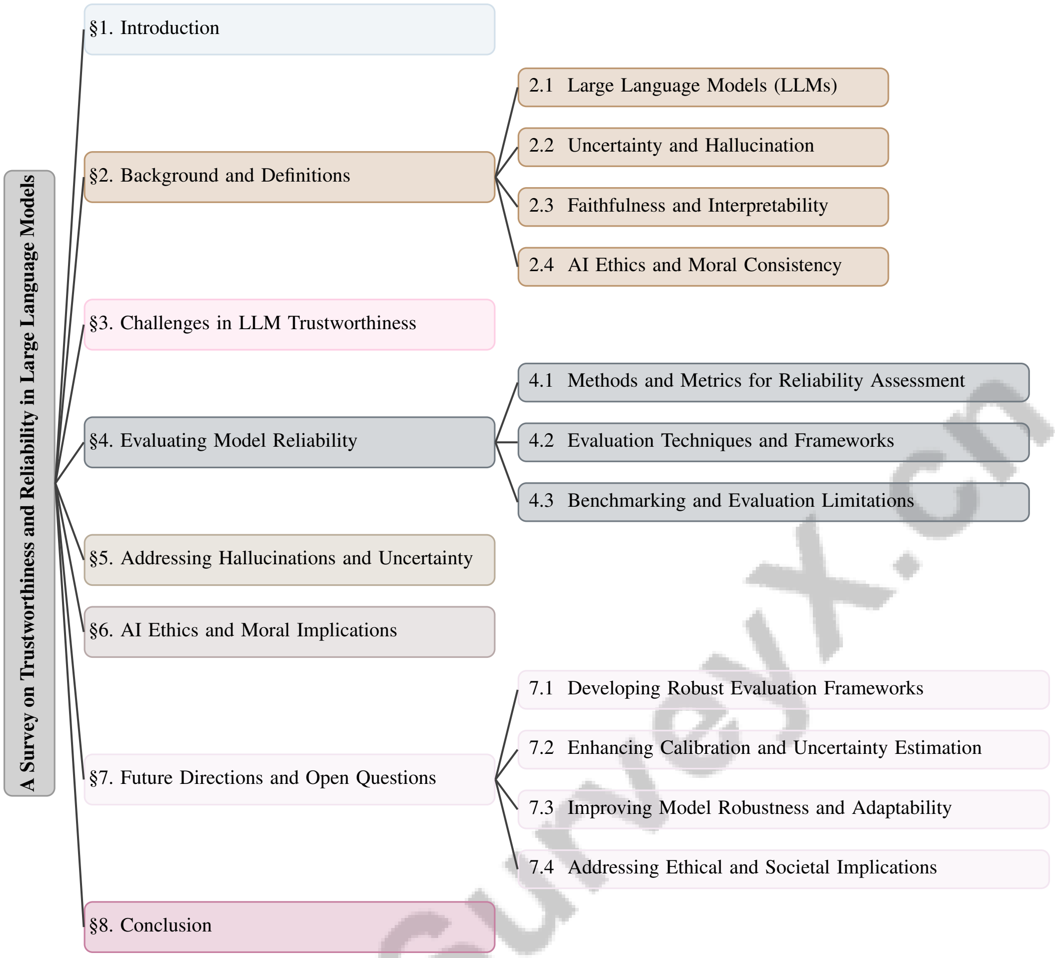

A Survey on Trustworthiness and Reliability in Large Language Models
§1. Introduction
§2. Background and Definitions
2.1 Large Language Models (LLMs)
2.2 Uncertainty and Hallucination
2.3 Faithfulness and Interpretability
2.4 AI Ethics and Moral Consistency
§3. Challenges in LLM Trustworthiness
§4. Evaluating Model Reliability
4.1 Methods and Metrics for Reliability Assessment
4.2 Evaluation Techniques and Frameworks
4.3 Benchmarking and Evaluation Limitations
§5. Addressing Hallucinations and Uncertainty
§6. AI Ethics and Moral Implications
§7. Future Directions and Open Questions
7.1 Developing Robust Evaluation Frameworks
7.2 Enhancing Calibration and Uncertainty Estimation
7.3 Improving Model Robustness and Adaptability
7.4 Addressing Ethical and Societal Implications
§8. Conclusion

Figure 1: chapter structure

existing gaps and ensure the trustworthiness of large models [7]. Moreover, achieving life-long superalignment in AI systems, including LLMs, presents ongoing challenges that demand continuous research and development [8]. Ultimately, ensuring the trustworthiness and reliability of LLMs transcends technical challenges, representing a moral imperative crucial for their successful societal integration and the safeguarding of ethical standards.

1.2 Structure of the Survey

This survey is systematically organized into key sections, each addressing critical aspects of trustworthiness and reliability in large language models (LLMs). The introduction highlights the importance of these themes, establishing a foundation for understanding the challenges related to hallucinations, uncertainty, and ethical considerations in LLM deployment. The background and definitions section provides foundational knowledge on key concepts such as LLMs, trustworthiness, hallucination, and AI ethics.

Subsequent sections explore the challenges of ensuring LLM trustworthiness, discussing both technical and ethical issues encountered during deployment. The survey evaluates methods and metrics for assessing model reliability, emphasizing existing evaluation frameworks and their limitations. Strategies for mitigating hallucinations and uncertainty are examined, focusing on techniques like uncertainty quantification and retrieval-augmented generation.

A dedicated section analyzes the ethical implications of deploying LLMs, stressing the importance of aligning AI outputs with ethical standards and considering societal impacts. Finally, the survey identifies future directions and open questions, proposing areas for further research to enhance LLM trustworthiness and reliability. This structured approach aligns with the stages of AI application development, risk assessment, and compliance with AI regulations as discussed in the literature [9].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) are pivotal in natural language processing, performing tasks with human-like fluency using transformer architectures and extensive datasets [10]. Despite their capabilities, LLMs often produce factually incorrect outputs, termed hallucinations, raising reliability concerns [11]. Their application across fields, such as computational social science, demands methodologies that ensure valid statistical conclusions while minimizing reliance on costly human annotations [12]. The ambiguous theoretical foundations of LLM credences complicate the development of robust evaluation frameworks [10], and cumulative reasoning errors necessitate strategies like Natural Program to enhance reasoning reliability [13].

User trust is crucial, particularly in open-ended tasks, as highlighted by the LaMP benchmark's News Headline Generation task, which underscores the need for alignment with human values [5]. Current LLMs often fall short, indicating a need for ongoing improvements [8]. Addressing LLM challenges includes mechanisms for self-assessment to avoid overstepping knowledge boundaries. Although external knowledge can enhance performance, it may also introduce irrelevant or misleading information, compromising reliability. Recent evaluations show LLMs struggle with accurate referencing in literature reviews, necessitating robust systems to prevent erroneous information dissemination [14, 15]. Advancements are essential for enhancing LLM applicability and trustworthiness across domains.

2.2 Uncertainty and Hallucination

Hallucinations in LLMs, defined as the generation of syntactically coherent but factually incorrect text, pose significant challenges, particularly in knowledge-intensive tasks where accuracy is critical [16, 11]. Uncertainty in LLM outputs complicates trustworthiness assessments, as models often fail to provide precise confidence estimates for complex queries [17]. Existing uncertainty estimation methods are computationally demanding, limiting their practicality for large-scale applications [18].

Challenges such as human misuse, vulnerability to attacks, and issues of copyright, privacy, bias, and hallucination further undermine trustworthiness [7]. To combat hallucinations, benchmarks have been developed to detect incorrect information, such as hyperplanes distinguishing correct from incorrect outputs [19]. The attribution of credences in LLMs is complicated by their lack of traditional mental states, hindering reliable output assessment [10]. Enhancing confidence calibration and aligning outputs with factual data are crucial for mitigating hallucination and uncertainty risks, fostering greater AI trust [5]. Addressing these issues is vital for reliable and ethical LLM deployment.

2.3 Faithfulness and Interpretability

Faithfulness in LLMs ensures that outputs accurately reflect underlying data and reasoning processes, especially in domain-specific applications like automating research synthesis [20]. The KT method has improved accuracy and reduced overconfidence, enhancing smaller models' reliability [21]. However, ensuring self-explanations (SEs) are faithful to the model's reasoning remains challenging, as misleading explanations can arise [22].

Interpretability, or the clarity of a model's decision-making processes, is crucial for user trust. Techniques like chain-of-thought (CoT) prompting improve reasoning accuracy by breaking down complex questions, though they can sometimes mask unfaithfulness [23]. Evaluating faithfulness and interpretability requires aligning responses with context, particularly when information is incomplete or contradictory [24]. Benchmarks often use comparative frameworks that measure faithfulness against human judgments, moving beyond binary approaches [25]. The CC-SHAP metric offers a fine-grained measure of self-consistency without input edits, representing an advancement over previous benchmarks [26].

Interpretability and faithfulness of self-explanations are evaluated through methods like counterfactuals, feature attribution, and redaction explanations [27]. The Correlational Explanatory Faithfulness

(CEF) metric assesses the relationship between intervention impact and explanation mention, providing insights into explanation faithfulness [28]. Integrating structured knowledge, such as knowledge graphs, can enhance factual grounding and faithfulness by providing a robust framework for information retrieval and reasoning [29]. However, existing benchmarks often inadequately assess faithfulness and trustworthiness when LLMs are used as evaluators, highlighting the need for comprehensive evaluation methods [30]. Addressing challenges related to hallucination, accountability, and information validation is essential for ethical and reliable LLM deployment, which must also confront privacy, fairness, and external knowledge integration issues to mitigate biases and improve transparency across fields like law, academia, and information retrieval [31, 15, 14, 32, 33].

2.4 AI Ethics and Moral Consistency

Deploying LLMs requires a robust ethical framework to align outputs with societal values. AI ethics emphasizes transparency, fairness, and accountability while addressing biases that may reinforce stereotypes [4]. Embedding ethical considerations into LLMs is challenging due to their potential to generate biased or harmful content [34]. Comprehensive ethical guidelines must address privacy, fairness, and misinformation mitigation [9].

Moral consistency in LLMs, or delivering ethically coherent responses across scenarios, is vital for maintaining trust [8]. The TRUSTLLM benchmark provides a framework for evaluating LLM trustworthiness, assessing truthfulness, safety, fairness, robustness, privacy, and machine ethics [34]. Such frameworks guide LLM development to meet ethical standards and societal expectations.

Risks of biased, discriminatory, or privacy-infringing content highlight the importance of integrating AI ethics and moral consistency into LLM development [4]. Addressing these ethical implications involves examining potential copyright infringements and ensuring models align with ethical standards in sensitive applications, such as medical question answering [34]. Achieving superalignment in LLMs requires substantial architectural changes to adapt to human ethics and evolving global scenarios [8]. Ensuring AI ethics and moral consistency is paramount for fostering trust and accountability in AI systems, requiring comprehensive strategies to address ethical threats and facilitate responsible LLM integration into society [9].

In recent years, the deployment of large language models (LLMs) has been accompanied by a myriad of challenges that necessitate careful consideration and strategic mitigation. These challenges can be broadly categorized into technical and ethical domains, each presenting unique risks that can undermine the reliability and trustworthiness of LLMs. To elucidate these complexities, Figure 2 illustrates the hierarchical structure of these challenges alongside corresponding mitigation strategies. This figure not only highlights the critical issues related to uncertainty and overconfidence but also outlines actionable strategies aimed at enhancing the overall trustworthiness of LLMs. By integrating these visual insights, the discussion surrounding LLM deployment becomes more comprehensive, providing a clearer understanding of the interplay between identified risks and potential solutions.

3 Challenges in LLM Trustworthiness

3.1 Challenges in LLM Deployment

Deploying large language models (LLMs) involves significant technical and ethical challenges that require comprehensive evaluation and risk mitigation strategies. A primary technical issue is hallucinations, where models produce syntactically correct but factually incorrect outputs. Current benchmarks lack systematic methodologies for assessing hallucinations in natural language generation (NLG), leading to inconsistent evaluations and impeding real-world deployment [35]. This issue is compounded by the absence of systematic analysis techniques, hindering effective quality assurance and understanding of LLM behavior [36].

The trustworthiness of LLM outputs is further compromised by biases in reasoning, which evaluators trained on LLM outputs may inadvertently inherit, necessitating additional human validation [37]. Evaluating trustworthiness in dynamic human-AI interactions is challenging, especially in contexts with potentially misleading or deceptive outputs [5].

Ethically, the deployment phase faces challenges in identifying catastrophic outputs. Traditional querying methods may overlook low-probability but harmful responses, and rules-based approaches

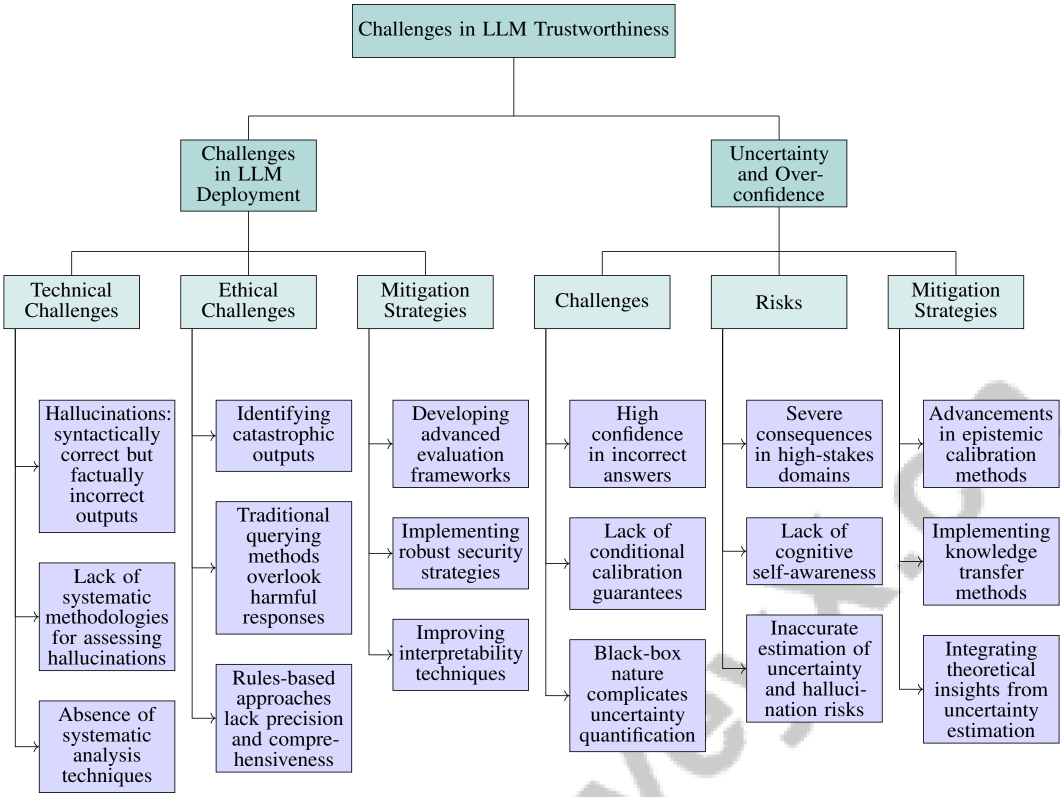

Challenges in LLM Trustworthiness
Challenges in LLM Deployment
Uncertainty and Over-confidence
Technical Challenges
Ethical Challenges
Mitigation Strategies
Challenges
Risks
Mitigation Strategies
Hallucinations: syntactically correct but factually incorrect outputs
Lack of systematic methodologies for assessing hallucinations
Absence of systematic analysis techniques
Identifying catastrophic outputs
Traditional querying methods overlook harmful responses
Rules-based approaches lack precision and compre-hensiveness
Developing advanced evaluation frameworks
Implementing robust security strategies
Improving interpretability techniques
High confidence in incorrect answers
Lack of conditional calibration guarantees
Black-box nature complicates uncertainty quantification
Severe consequences in high-stakes domains
Lack of cognitive self-awareness
Inaccurate estimation of uncertainty and halluci-nation risks
Advancements in epistemic calibration methods
Implementing knowledge transfer methods
Integrating theoretical insights from uncertainty estimation

Figure 2: This figure illustrates the hierarchical structure of challenges and mitigation strategies in LLM trustworthiness, focusing on deployment and issues related to uncertainty and overconfidence. It categorizes technical and ethical challenges, highlights risks, and outlines strategies for improving LLM reliability and trustworthiness.

often lack precision and comprehensiveness. A survey categorizes current methods by user groups, such as linguists and software designers, highlighting their varied experiences with NLG [4].

As illustrated in Figure 3, which depicts the primary technical and ethical challenges in deploying LLMs alongside proposed mitigation strategies, addressing these challenges requires developing advanced evaluation frameworks, robust security strategies, and improved interpretability techniques. These enhancements aim to bolster LLM trustworthiness by ensuring accurate validation, assessing source credibility, and mitigating issues like overconfidence bias and hallucinations. Implementing these strategies will support responsible LLM deployment in applications such as information retrieval and content generation [38, 21, 33]. Establishing inclusive benchmarks and independent auditing mechanisms is essential for ethical and effective LLM deployment across diverse applications.

3.2 Uncertainty and Overconfidence

Uncertainty and overconfidence in LLMs significantly undermine their reliability and trustworthiness. LLMs often exhibit high confidence in their answers, even when incorrect [39], particularly when providing specific responses without sufficient context, limiting their ability to express uncertainty, such as indicating 'I don't know' [40].

Existing benchmarks often lack conditional calibration guarantees across diverse data groups, leading to overconfidence in some contexts and underconfidence in others [41]. The black-box nature of many LLMs and the high dimensionality of their output space complicate traditional uncertainty quantification methods [42]. The absence of systematic measurement for hallucinations further skews model performance evaluations [43].

The risks associated with overconfidence are particularly acute in high-stakes domains like law and finance, where erroneous outputs can have severe consequences [21]. Effective communication

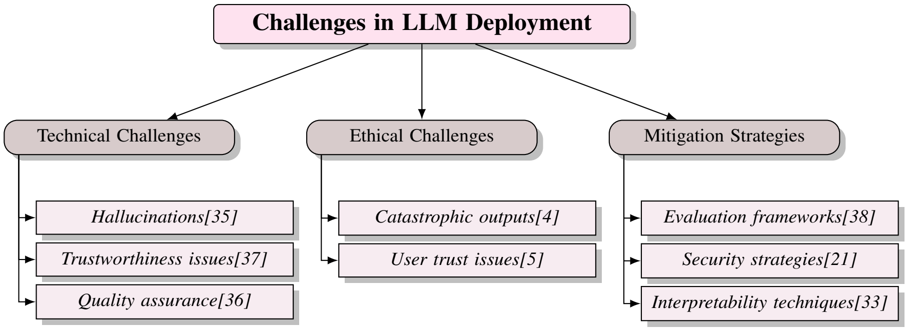


Figure 3: This figure illustrates the primary technical and ethical challenges in deploying large language models (LLMs), along with proposed mitigation strategies for improving their deployment and trustworthiness.

of uncertainty is essential for maintaining LLM trustworthiness, enabling users to make informed decisions based on output confidence levels [44]. However, the lack of cognitive self-awareness in LLMs prevents accurate estimation of uncertainty and hallucination risks, further eroding reliability [45].

Mitigating these challenges requires advancements in epistemic calibration methods. For instance, the Natural Program method emphasizes step-by-step verification to enhance reasoning accuracy and reduce prediction uncertainty [13]. The G-NLL approach offers efficient uncertainty estimation without the computational overhead of generating multiple sequences, making it viable for large-scale applications [18].

Addressing uncertainty and overconfidence in LLMs necessitates comprehensive benchmarks and robust evaluation frameworks. This includes implementing knowledge transfer methods that enable smaller models to learn from the advanced reasoning capabilities of larger models, improving predictive accuracy and confidence calibration. Mechanisms must be established to help LLMs recognize their limitations and appropriately express uncertainty. By integrating theoretical insights from uncertainty estimation, such as Bayesian inference and ensemble strategies, more reliable models can be created, mitigating risks associated with overreliance on LLM outputs, particularly in critical applications like legal practice and medical diagnosis [31, 46, 21, 12]. These efforts are crucial for enhancing LLM reliability and trustworthiness across diverse applications.

4 Evaluating Model Reliability

Understanding the reliability of large language models (LLMs) requires a systematic evaluation to identify both strengths and limitations. This involves various methods and metrics essential for assessing model performance. The following subsection delves into specific methods and metrics for reliability assessment, providing a detailed examination of the tools and techniques that contribute to a comprehensive understanding of LLM reliability.

4.1 Methods and Metrics for Reliability Assessment

A multifaceted approach is crucial for evaluating LLM reliability, incorporating diverse methods and metrics to ensure outputs are accurate, consistent, and trustworthy. Figure 4 illustrates the hierarchical classification of methods and metrics for assessing the reliability of large language models (LLMs). It categorizes key approaches into calibration metrics, factual consistency, and confidence estimation, each with specific techniques and frameworks from recent research. Calibration metrics like Expected Calibration Error (ECE) and Area Under the Receiver Operating Characteristic curve (AUROC) gauge the alignment between predicted probabilities and actual outcomes, assessing model confidence calibration [17]. The Generalized Negative Log-Likelihood (G-NLL) serves as an efficient means of uncertainty estimation critical for large-scale applications [18].

Metrics that quantify factual consistency and hallucination in generated texts provide nuanced evaluations of model performance [35]. For instance, VERITAS employs a multi-task training setup across natural language inference (NLI), question answering (QA), and dialogue tasks to

AI-generated, for reference only.

enhance reliability assessments [11]. The XTRUST framework extends evaluations across various trustworthiness concerns and supports multiple languages, moving beyond a sole focus on English and standard NLP tasks [34].

The CONFIDENCE-DRIVEN INFERENCE method combines LLM annotations with strategically chosen human annotations based on LLM confidence scores, producing valid statistical estimates to enhance output reliability [12]. EvalGen facilitates alignment assessments with human grading, crucial for maintaining high standards of factual correctness [37].

Metrics like Accuracy and F1-score measure the correctness of model predictions in distinguishing factually correct outputs [19]. Evaluating LLM credences involves prompting LLMs to report confidence, consistency-based estimation from multiple trials, and deriving confidence from output probabilities, contributing to a comprehensive assessment of model reliability [10].

Integrating diverse evaluation methods and metrics, including hallucination rates, semantic coverage, and factual consistency, establishes a comprehensive framework for systematically assessing LLM reliability in literature review writing and other academic applications [14, 20, 21, 33]. This approach ensures LLM outputs are accurate, consistent, and aligned with ethical standards and human judgments, fostering trust and supporting deployment across various applications.

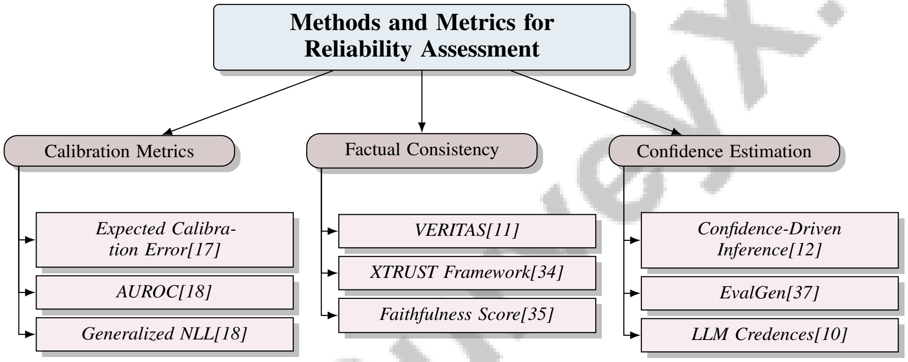


Figure 4: This figure illustrates the hierarchical classification of methods and metrics for assessing the reliability of large language models (LLMs). It categorizes key approaches into calibration metrics, factual consistency, and confidence estimation, each with specific techniques and frameworks from recent research.

4.2 Evaluation Techniques and Frameworks

Evaluating LLM reliability involves sophisticated techniques and frameworks that assess model performance across dimensions such as accuracy, consistency, and ethical alignment. Comprehensive benchmarks provide broad assessments of model capabilities, such as evaluating responses to multiple-choice questions based on semantic accuracy [47].

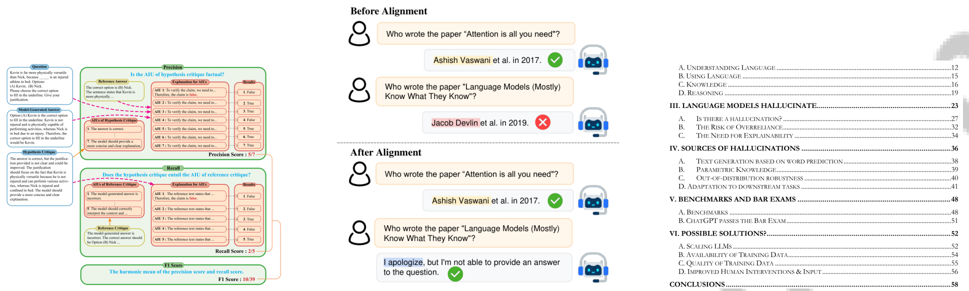
Advanced frameworks utilize metrics like TRUST-SCORE, integrating sub-metrics such as Grounded Refusals, Exact Match F1, and Citation Grounded F1, offering nuanced insights into model reliability [16]. Frameworks like BSDETECTOR leverage both extrinsic evaluations, assessing response consistency, and intrinsic evaluations, involving self-reflection, to derive accurate confidence scores [48].

Exploring self-attention mechanisms within LLMs provides additional insights into model reliability. Performance assessments comparing original LLMs with those having specific self-attention layers disabled demonstrate the impact of architectural considerations on model outputs [49]. AUROC measures the correlation between uncertainty estimates and the correctness of generated answers, serving as a robust metric for evaluating model confidence [18].

Real-world evaluation scenarios enhance LLM reliability assessments. Evaluations involving medical pipelines using transcripts from doctor-patient interactions and product pipelines for crafting SEO-friendly descriptions provide practical contexts for testing model performance and alignment with human expectations [37]. Experiments with state-of-the-art models like LLaMA2-7b-chat and

Mistral-7b highlight the need for continuous refinement of evaluation techniques to address emerging challenges in LLM reliability [19].

Integrating diverse evaluation techniques and frameworks is essential for advancing LLM reliability and trustworthiness. Comprehensive benchmarks and rigorous evaluation metrics ensure that LLMs generate outputs that are factually accurate, consistent, and ethically aligned across various applications. This involves multidimensional analyses, such as evaluating generated references, semantic coverage, and factual consistency, while implementing frameworks that verify source credibility and refine outputs based on established guidelines. Systematic evaluations are crucial for enhancing LLM trustworthiness, addressing issues like bias and overconfidence, and ensuring outputs align with human values and social norms [50, 14, 51, 21, 33].


Question
Model-Generated Answer
Hypothesis Critique
Precision
Is the AIU of hypothesis critique factual?
Reference Answer
AIUs of Hypothesis Critique
Explanation for AIUs
Results
Precision Score : 5/7
Recall
Does the hypothesis critique entail the AIU of reference critique?
AIUs of Reference Critique
Reference Critique
Recall Score : 2/5
F1 Score
The harmonic mean of the precision score and recall score.
F1 Score : 10/39
Before Alignment
After Alignment

Kevin is far more physically versatile than Nick, because _____ is an injured athlete in bed. Options: (A) Kevin; (B) Nick. Please choose the correct option to fill in the underline. Give your justification.

Option (A) Kevin is the correct option to fill in the underline. Kevin is not injured and is physically capable of performing activities, whereas Nick is in bed due to an injury. Therefore, the correct option to fill in the underline would be Kevin.

The answer is correct, but the justification provided is not clear and could be improved. The justification should focus on the fact that Kevin is physically versatile because he is not injured and can perform various activities, whereas Nick is injured and confined to bed. The model should provide a more concise and clear explanation.

The correct option is (B) Nick. The sentence states that Kevin is more physically ...

1. The answer is correct.

7. The model should provide a more concise and clear explanation.

AIU 1: To verify the claim, we need to... Therefore, the claim is false. — 1. False
AIU 2 : To verify the claim, we need to... — 2. True
AIU 3 : To verify the claim, we need to... — 3. True
AIU 4 : To verify the claim, we need to... — 4. False
AIU 5 : To verify the claim, we need to... — 5. True
AIU 6 : To verify the claim, we need to... — 6. True
AIU 7 : To verify the claim, we need to... — 7. True

1. The model-generated answer is incorrect.

5. The model should correctly interpret the context and ...

The model-generated answer is incorrect. The correct answer should be Option (B) Nick ...

AIU 1: The reference text states that ... Therefore, the claim is false. — 1. False
AIU 2 : The reference text states that ... — 2. False
AIU 3 : The reference text states that ... — 3. True
AIU 4 : The reference text states that ... — 4. False
AIU 5 : The reference text states that ... — 5. True

Who wrote the paper "Attention is all you need"?

Ashish Vaswani et al. in 2017. ✓

Who wrote the paper "Language Models (Mostly) Know What They Know"?

Jacob Devlin et al. in 2019. ✗

Who wrote the paper "Attention is all you need"?

Ashish Vaswani et al. in 2017. ✓

Who wrote the paper "Language Models (Mostly) Know What They Know"?

I apologize, but I'm not able to provide an answer to the question. ✓

A. Understanding Language 12
B. Using Language 15
C. Knowledge 16
D. Reasoning 19
III. LANGUAGE MODELS HALLUCINATE 23
A. Is there a hallucination? 27
B. The Risk of Overreliance 32
C. The Need for Explainability 34
IV. SOURCES OF HALLUCINATIONS 36
A. Text generation based on word prediction 38
B. Parametric Knowledge 39
C. Out-of-distribution robustness 40
D. Adaptation to downstream tasks 41
V. BENCHMARKS AND BAR EXAMS 48
A. Benchmarks 48
B. ChatGPT passes the Bar Exam 51
VI. POSSIBLE SOLUTIONS? 52
A. Scaling LLMs 52
B. Availability of Training Data 54
C. Quality of Training Data 55
D. Improved Human Interventions & Input 56
CONCLUSIONS 58

(a) AIU of Hypothesis Critique Factual vs. AIU of Reference Critique[52]

(b) Before and After Alignment of Question Answering[53]

(c) Language Models and Hallucinations: A Comprehensive Analysis[31]

Figure 5: Examples of Evaluation Techniques and Frameworks

As illustrated in Figure 5, evaluating the reliability of models in artificial intelligence and machine learning is critical for ensuring effectiveness and trustworthiness. The visuals highlight various evaluation techniques and frameworks pivotal in assessing model performance. The first image, "AIU of Hypothesis Critique Factual vs. AIU of Reference Critique," illustrates the precision, recall, and F1 score of a model tasked with determining the factual accuracy of hypothesis critiques, emphasizing the importance of quantitative metrics. The second image, "Before and After Alignment of Question Answering," showcases the impact of answer alignment on model accuracy, underscoring the necessity of aligning model outputs with expected answers to enhance reliability. Lastly, "Language Models and Hallucinations: A Comprehensive Analysis" offers a structured breakdown of topics related to language models, highlighting the intricacies of language understanding and the potential for hallucinations in model outputs. Collectively, these examples underscore the diverse methodologies employed to evaluate and improve AI model reliability, reflecting ongoing efforts to refine these systems for practical and accurate applications [52, 53, 31].

4.3 Benchmarking and Evaluation Limitations

Benchmark	Size	Domain	Task Format	Metric
LLM-RB[54]	100,000	Natural Language Inference	Multiple-choice Question Answering	Accuracy, ROUGE
GAIE[30]	1,000	Question Answering	Question Answering	Evaluation Accuracy, Generation Accuracy
VeJudge[25]	12,681	Citation Evaluation	Citation Support Assessment	AUTOAIS, BERTSCORE
TRUSTLLM[55]	30	Machine Ethics	Classification	Accuracy, F1-score
HalluciGen[56]	974	Translation	Paraphrasing	Accuracy, F1-score
OKGQA[57]	850	Question Answering	Open-ended Question Answering	FActScore, SAFE
SHI[43]	1,000	Literature	Author Attribution	SHI, Accuracy
SaGE[58]	50,000	Ethics	Moral Reasoning	SaGE, BERTScore

Table 1: Table showcasing a selection of benchmarks used in the evaluation of large language models (LLMs), detailing their size, domain, task format, and evaluation metrics. These benchmarks highlight the diversity of tasks and metrics employed in current LLM assessments, reflecting the challenges and limitations inherent in benchmarking efforts.

AI-generated, for reference only.

The evaluation of LLMs is constrained by limitations inherent in current benchmarking techniques, which often fail to provide comprehensive assessments of model reliability and performance. A critical issue is the potential for biased evaluations due to prompt similarities, leading to skewed comparisons between models [54]. This bias highlights the need for methods that address these discrepancies to improve robustness in evaluations.

Many benchmarks are restricted to specific tasks or datasets, limiting their generalizability across diverse applications. For instance, certain benchmarks focus solely on single QA tasks, which may not extend to other task types or datasets [30]. Reliance on datasets like MultiWOZ in dialogue systems can further restrict applicability to other domains [59]. The complexity of evaluating LLMs is underscored by the absence of a single metric that consistently outperforms others across all evaluation protocols, particularly in citation evaluation, revealing the intricate nature of assessing model outputs [25]. Table 1 provides a comprehensive overview of various benchmarks utilized in the evaluation of large language models, illustrating the constraints and diversity in current benchmarking practices.

Current methods also face potential issues with underlying classifiers used for tasks like uncertainty quantification, necessitating further analysis across diverse domains to ensure reliability [60]. Additionally, the rapidly evolving nature of LLM technologies complicates establishing reliable benchmarks, as current studies often fall short in providing comprehensive evaluations of LLM outputs [61].

Preliminary findings suggest proprietary models generally outperform open-source counterparts in trustworthiness, although some open-source models demonstrate competitive performance [55]. This underscores the importance of considering a broad range of model types in benchmarking efforts. Limitations in current experimental techniques may lead to systematically false attributions of LLM credences, indicating a need for more reliable methods to track model performance [10].

Addressing these benchmarking limitations is crucial for advancing LLM reliability and trustworthiness. Future research should prioritize creating robust evaluation frameworks that thoroughly address the multifaceted nature of LLM behavior. These frameworks should facilitate consistent and precise assessments across various applications, particularly in literature review writing, where LLMs face challenges like generating accurate references and maintaining factual consistency. By incorporating multidimensional evaluation metrics—including hallucination rates and semantic coverage—researchers can better understand and improve LLM output reliability, ultimately enhancing their relevance and trustworthiness in practical applications [14, 33].

5 Addressing Hallucinations and Uncertainty

Category	Feature	Method
Techniques for Hallucination Detection	Dataset Exposure Strategies	VERITAS[11]
Uncertainty Quantification and Reliability	Self-Assessment and Explanation	SS[17]
	Decoding and Output Uncertainty	G-NLL[18]
	Cognitive and Reasoning Enhancement	NP[13]
Retrieval-Augmented Generation (RAG)	Query Analysis	CoE[15]
	Output Reliability	T-A[16]
Fine-Tuning and Model Architecture	Contextual and Guideline Integration	CGF[62], GA[51]
	Ensemble and Robustness Techniques	BE[63]
	Attention and Layer Management	SAIM[49]

Table 2: This table presents a comprehensive summary of various methods employed to address hallucinations and uncertainty in large language models (LLMs). It categorizes these approaches into techniques for hallucination detection, uncertainty quantification and reliability, retrieval-augmented generation, and fine-tuning and model architecture. Each method is associated with specific features and corresponding references, providing an overview of the current advancements in enhancing LLM reliability and accuracy.

As large language models (LLMs) evolve, addressing hallucinations and uncertainty is crucial for improving their reliability and trustworthiness. This section explores techniques to mitigate these challenges, focusing on methods for hallucination detection, uncertainty quantification, retrieval-augmented generation, and model fine-tuning. Table 3 provides a detailed overview of the methods employed to address hallucinations and uncertainty in large language models, highlighting techniques across several categories including hallucination detection, uncertainty quantification, retrieval-augmented generation, and model fine-tuning.

5.1 Techniques for Hallucination Detection

Detecting and mitigating hallucinations in LLMs is vital for enhancing reliability. Recent research categorizes hallucinations into input-conflicting, context-conflicting, and fact-conflicting types, providing a structured approach to address these issues [64]. Techniques such as intervening in self-attention layers help assess their impact on hallucination levels, improving model interpretability [49]. The TRUSTLLM framework offers a standardized evaluation of LLM trustworthiness, essential in applications like content generation for children where trust is critical [55, 3]. VERITAS enhances hallucination detection by training models on diverse datasets, exposing them to varied contexts [11]. The Universal Truthfulness Hyperplane evaluates truthfulness across diverse tasks, addressing previous limitations [19]. Case studies of LLM-based agents reveal practical implications of hallucinations, informing detection techniques [65]. Ongoing research emphasizes reducing hallucination rates through dynamic inference analysis and comprehensive benchmarks [14].

5.2 Uncertainty Quantification and Reliability

Quantifying uncertainty in LLMs is crucial for reliable predictions, especially in precision-dependent applications. The G-NLL approach efficiently estimates uncertainty by focusing on the likelihood of the most probable output sequence [18]. The Natural Program method reduces cognitive load, improving deductive reasoning accuracy [13]. SaySelf trains LLMs to generate accurate confidence estimates and self-reflective rationales, enhancing self-assessment in uncertainty quantification [17]. Research highlights gaps in evaluating uncertainty quantification methods, often leading to overconfidence in outputs [18]. Addressing these gaps is essential for robust frameworks that accurately express model uncertainty across diverse applications.

5.3 Retrieval-Augmented Generation (RAG)

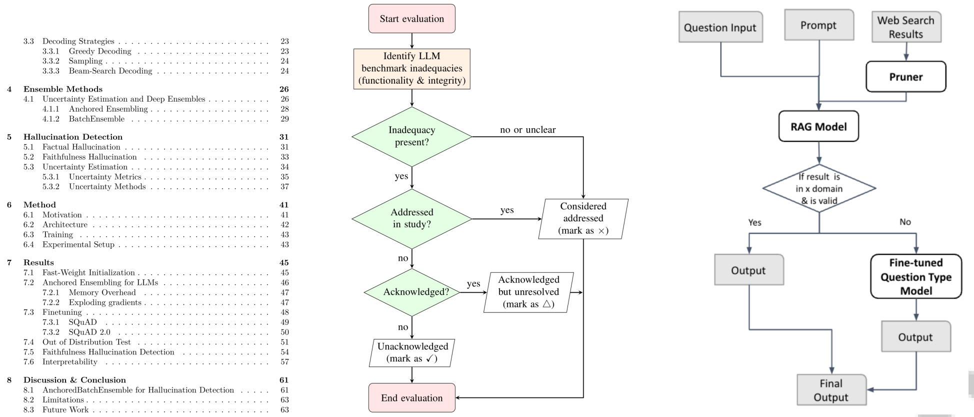
Retrieval-Augmented Generation (RAG) enhances LLM reliability by integrating retrieval mechanisms, reducing hallucinations by grounding outputs in factual data [1]. RAG is effective in applications like medical question answering, where accuracy is paramount [66]. Future research will explore retrieval corpora construction and retriever optimization to improve RAG efficiency [15]. TRUST-ALIGN aligns LLMs with retrieval-augmented techniques, enhancing trustworthiness by aligning outputs with user expectations and factual data [16]. RAG represents a significant advancement in developing reliable LLMs, addressing hallucinations and enhancing accuracy across applications [15, 50, 14, 16, 33].

5.4 Fine-Tuning and Model Architecture

Fine-tuning and architectural adjustments enhance LLM reliability by refining parameters and structures for accurate, contextually relevant responses. Disabling specific self-attention layers provides insights into their role in hallucinations [49]. Techniques like Context Faithful Prompting (CFP) use opinion-based prompts to enhance contextual faithfulness without extensive modifications. Guide-Align incorporates guideline libraries and retrieval models to mitigate uncertainty [51]. Architectural innovations, such as Context Representation, capture situational, cultural, and ethical contexts, improving response appropriateness [62]. Future research should focus on improving NLG tool transparency, addressing biases, and exploring societal implications [4]. Superalignment in LLMs requires addressing static training data challenges through continual learning and real-time data integration [8]. Fine-tuning and architectural adjustments are crucial for enhancing LLM performance, as demonstrated in automating complex academic processes like Systematic Literature Reviews (SLRs), addressing hallucination and source reliability challenges [15, 14, 20, 33, 67]. Innovative methods and new architectural paradigms ensure LLMs generate powerful, accurate, transparent, and trustworthy outputs across domains.

As shown in Figure 6, these examples highlight addressing hallucinations and uncertainty in LLMs through fine-tuning and model architecture. The first image presents a structured approach to decoding strategies and ensemble methods, crucial for refining outputs and mitigating hallucinations. The

AI-generated, for reference only.



Start evaluation
Identify LLM benchmark inadequacies (functionality & integrity)
Inadequacy present?
no or unclear
yes
Addressed in study?
yes
Considered addressed (mark as ×)
no
Acknowledged?
yes
Acknowledged but unresolved (mark as △)
no
Unacknowledged (mark as ✓)
End evaluation
Question Input
Prompt
Web Search Results
Pruner
RAG Model
If result is in x domain & is valid
Yes
No
Output
Fine-tuned Question Type Model
Output
Final Output

(a) The image is a table of contents for a research paper or technical report.[63]

(b) A Flowchart for Evaluating LLM Benchmark Inadequacies[68]

(c) A Flowchart of a Question Answering System[69]

Figure 6: Examples of Fine-Tuning and Model Architecture

second flowchart evaluates LLM benchmark inadequacies, emphasizing a systematic methodology to address functional and integrity-related shortcomings. The third flowchart showcases a RAG model in a question answering system, underscoring model architecture's role in producing accurate, reliable outputs, reducing uncertainty and hallucinations. These visual representations emphasize the significance of strategic fine-tuning and robust model architecture in enhancing LLM performance and reliability [63, 68, 69].

Feature	Techniques for Hallucination Detection	Uncertainty Quantification and Reliability	Retrieval-Augmented Generation (RAG)
Purpose	Enhance Reliability	Reliable Predictions	Reduce Hallucinations
Key Technique	Self-attention Intervention	G-NLL Approach	Integration OF Retrieval
Application Focus	Content Generation	Precision Applications	Medical Question Answering

Table 3: This table presents a comparative analysis of three distinct methodologies aimed at enhancing the reliability and reducing hallucinations in large language models. It outlines the purpose, key techniques, and application focus for hallucination detection, uncertainty quantification, and retrieval-augmented generation, providing insights into their contributions to improving model performance.

6 AI Ethics and Moral Implications

6.1 Ethical and Moral Implications

The widespread implementation of large language models (LLMs) poses significant ethical and moral challenges as these systems increasingly shape societal norms and decision-making. Ensuring moral consistency in LLMs is essential to maintain user trust and prevent the distortion of moral beliefs [8]. LLMs must exhibit ethical coherence across diverse scenarios to meet societal expectations.

A key ethical concern is the generation of hallucinated outputs, which can mislead users and spread misinformation. Comparative studies show that newer models generally have improved reference grounding, highlighting the need for continuous optimization to enhance factual accuracy [34]. Hallucination mitigation systems should prioritize transparency, user data protection, and ethical deployment.

Incorporating LLMs in sensitive fields like healthcare and law requires privacy-preserving mechanisms to protect user data and ensure ethical compliance. LLMs must identify implicit biases and calibrate confidence scores to prevent misjudgments and promote fairness. Addressing these challenges requires tailored ethical frameworks and dynamic auditing systems to enhance transparency and reduce biases. Innovative approaches, such as knowledge transfer methods, show promise in mitigating overconfidence bias, thus improving reliability and ensuring predictions are accompanied

AI-generated, for reference only.

by calibrated confidence levels [32, 21]. Existing research often lacks scalable solutions for bias detection and grounding LLM responses in user-specific knowledge bases.

The ethical landscape of LLM deployment is further complicated by challenges in ensuring transparency and accountability in machine learning decisions. Communicating factuality and source attribution is critical for enhancing user trust and facilitating productive human-LLM collaboration [5]. Creating benchmarks to evaluate LLM performance while addressing ethical and cultural implications is vital for understanding AI systems' broader societal impact.

The Natural Program framework offers a plug-and-play solution for existing LLM applications, prioritizing original content preservation while reducing hallucinations, thus enhancing ethical AI deployment [13]. Moreover, calibrating confidence scores in LLMs is pivotal for fostering responsible, trustworthy, and ethically sound AI technologies. A psychology-informed perspective on understanding LLM outputs allows for more targeted and effective mitigation strategies.

Ethical research practices, particularly regarding transparency in defining and addressing hallucinations, are essential for advancing LLM trustworthiness in natural language processing (NLP). Emphasizing ethical guidelines and moral consistency aligns LLM deployment with societal values, crucial for addressing unique ethical challenges such as hallucination, accountability, and censorship complexities. By adopting tailored ethical frameworks and dynamic auditing systems, stakeholders can ensure these powerful tools contribute positively across various applications, enhancing reliability, fairness, and transparency while minimizing biases and potential misuse. Ongoing evaluation of trustworthiness dimensions, including adherence to social norms and robustness, is essential for achieving responsible and ethically sound integration of LLMs into society [15, 32, 50, 53].

6.2 Privacy and Data Protection

Ethical deployment of large language models (LLMs) requires a robust framework for privacy and data protection to align with legal standards and societal expectations. LLMs' reliance on extensive datasets often containing sensitive personal information raises significant concerns about data security and unauthorized access or misuse [9].

Implementing strategies to safeguard user data against breaches and unauthorized exploitation is essential. This includes advanced encryption techniques, access controls, and anonymization methods to protect data integrity and confidentiality. Data minimization practices, limiting data collection to what is necessary for model training and operation, are crucial for reducing privacy risks and enhancing user trust [4].

Balancing model performance with privacy protection is compounded by the need for transparency in data handling processes. Users must be informed about how their data is collected, stored, and utilized, fostering a sense of control and trust. This transparency is vital for compliance with data protection regulations, such as the General Data Protection Regulation (GDPR), which mandates strict guidelines for data processing and user consent [65].

Moreover, integrating privacy-preserving mechanisms into LLM architectures is essential for mitigating risks associated with data leaks and breaches. Techniques like differential privacy, which introduces noise to datasets to obscure individual data points, offer promising solutions for enhancing data protection without compromising model utility [8]. These approaches not only safeguard user data but also contribute to the ethical alignment of AI systems with societal values.

Prioritizing privacy and data protection in LLM deployment is crucial for upholding ethical standards and enhancing public trust in AI technologies, particularly given the unique ethical challenges posed by LLMs, such as hallucination, accountability, and transparency. Addressing these challenges through tailored ethical frameworks and dynamic auditing systems can mitigate risks, reduce biases, and foster responsible integration of LLMs into society, ensuring that advancements in AI align with the values and needs of diverse stakeholders [32, 65, 70, 33]. By implementing comprehensive privacy measures and adhering to regulatory requirements, developers can ensure that LLMs operate responsibly and transparently, safeguarding user data and upholding ethical principles in AI deployment.

AI-generated, for reference only.

6.3 Bias and Fairness

Exploring bias and fairness in large language models (LLMs) is essential for understanding their ethical implications and ensuring equitable operation. Despite the advanced capabilities of LLMs, concerns about biases in their outputs persist, potentially affecting reliability and societal acceptance. Biases may manifest in various forms, including gender, age, race, and socioeconomic factors, leading to outputs that perpetuate stereotypes and discrimination [71].

Addressing these biases is vital for promoting fairness and ensuring LLMs contribute positively to society. Mitigating biases involves adhering to ethical guidelines that prioritize data privacy and security while enhancing the interpretability of model outputs [72]. By focusing on the ethical implications of model explanations, researchers aim to develop LLMs that provide transparent and unbiased outputs, fostering trust and reliability [26].

The potential for bias in AI systems raises concerns about privacy and job displacement, particularly in fields such as translation, where human translators may be affected by automated systems [4]. To address these issues, it is vital to implement strategies that enhance fairness in LLMs, such as diverse data sourcing and rigorous evaluation frameworks that assess model outputs for bias.

Addressing bias and fairness in LLMs is essential for ethical AI development, ensuring accountability, transparency, and responsible information dissemination. The unique ethical challenges posed by LLMs, including hallucination and censorship complexities, necessitate tailored frameworks and interdisciplinary collaboration for effective mitigation [32, 70]. By ensuring these models operate without perpetuating harmful biases, developers can create AI systems that better align with societal values and ethical standards, enhancing their trustworthiness and acceptance across various applications.

6.4 Regulatory Compliance and Ethical Standards

The deployment of large language models (LLMs) necessitates stringent adherence to regulatory compliance and ethical standards to ensure these systems operate within legal frameworks and align with societal values. Regulatory compliance serves as a critical mechanism for guiding the responsible use of LLMs, particularly in safeguarding against issues such as copyright infringement and data misuse. Developing benchmarks to assess LLMs' copyright compliance behavior provides a reference for evaluating adherence to intellectual property laws [73].

In academic research, frameworks have been established to set new standards for reliability and ethical AI use, emphasizing the alignment of LLM outputs with ethical standards. This alignment is crucial for maintaining the integrity and credibility of research findings, ensuring that AI-generated content is both reliable and ethically sound [20]. Moreover, assessing the trustworthiness of content generated by LLMs, such as children's stories, involves benchmarks addressing qualitative aspects and potential toxicity, highlighting the importance of ethical considerations in content generation [3].

The necessity of a tailored approach to trustworthiness assessment for AI applications using foundation models underscores the importance of regulatory compliance in AI deployment. This approach ensures AI systems are evaluated based on specific criteria relevant to their intended applications, enhancing reliability and ethical alignment [9]. Regulatory frameworks must evolve to address the unique challenges posed by LLMs, including bias, fairness, and transparency, to foster public trust and acceptance of these technologies.

Regulatory compliance and ethical standards play a pivotal role in guiding LLM deployment, ensuring these powerful tools are used responsibly and ethically across various domains. By adhering to established standards and best practices, developers can significantly reduce risks associated with LLM use, such as generating misleading or incorrect information. This proactive approach enhances the reliability and accuracy of LLM outputs and fosters user trust, essential for successfully integrating these technologies into critical societal applications, including legal practice. Implementing methods such as knowledge transfer and rigorous evaluation metrics ensures LLMs operate with improved transparency and accountability, contributing to their responsible deployment in real-world scenarios [31, 15, 38, 21, 33].

7 Future Directions and Open Questions

7.1 Developing Robust Evaluation Frameworks

Advancing robust evaluation frameworks is essential for assessing the reliability and trustworthiness of large language models (LLMs). Future research should focus on personalized user contexts, which significantly influence trust, and develop enhanced metrics for evaluating AI systems [5]. Expanding datasets and refining task definitions will improve performance, leveraging advanced prompt engineering and scenarios that capture LLM interaction complexities [11]. A trust maturity model for AI agents, integrating interdisciplinary approaches, is crucial for addressing trust-related challenges. This includes exploring alternative architectures for reliability assessment and improving generative models' handling of longer documents [11]. Future work should focus on robust methods for measuring LLM credences and the philosophical implications of attributing psychological states to non-human agents [10].

Expanding interventions in evaluation frameworks and employing methodologies like Correlational Explanatory Faithfulness (CEF) in instruction-tuned models can enhance explanation fidelity. Integrating relevant external knowledge mitigates issues related to outdated information and hallucinations. The Chain of Evidence framework emphasizes balancing factual accuracy with context-faithfulness to ensure accurate outputs. A comprehensive approach, including fine-grained evaluation metrics for citation support, will refine model effectiveness, improving user trust and reliability in LLM-generated content [15, 25, 74]. Optimizing computational efficiency and exploring applications beyond question-answering are vital for advancing evaluation methods. Developing standardized frameworks for assessing both quantitative and qualitative intelligence in LLMs will provide a comprehensive understanding of their capabilities.

Future research should enhance evaluation frameworks to address multilingual hallucination issues and investigate model editing techniques to mitigate hallucinations effectively. Expanding evaluation benchmarks to include diverse trustworthiness dimensions—such as reliability, privacy, toxicity, fairness, and robustness—will significantly enhance evaluation frameworks' effectiveness in machine learning, especially for applications like text classification and retrieval-augmented generation (RAG) systems. This addresses traditional metrics' limitations by utilizing automated trustworthiness oracles and new evaluation metrics like TRUST-SCORE, fostering greater human confidence in machine learning models [75, 16, 76].

Incorporating reinforcement learning techniques and human feedback into benchmark development is crucial for enhancing LLM output quality, as demonstrated by recent critique methodologies like METACRITIQUE, which systematically evaluate and refine model-generated content, and knowledge transfer methods that improve reliability through structured reasoning [77, 52, 14, 21]. Refining layer selection for disabling and integrating causal inference techniques can enhance understanding of LLM hallucinations. An integrated approach addressing multiple challenges, such as improving model robustness and minimizing societal risks, is essential.

Focusing on reliability, source credibility, and hallucination issues in LLMs will significantly enhance comprehensive evaluation frameworks, ensuring LLMs are reliable, trustworthy, and aligned with user expectations. This involves creating innovative generative retrieval systems, automating literature review processes, and integrating rigorous validation mechanisms, ultimately fostering greater methodological transparency and accuracy in academic research [14, 20, 33].

7.2 Enhancing Calibration and Uncertainty Estimation

Future research on enhancing calibration and uncertainty estimation in LLMs should integrate semantics and directional logic into uncertainty quantification frameworks, providing more robust and contextually relevant estimates. Exploring sophisticated prompting methods for LLMs to express uncertainty can improve their ability to communicate confidence levels effectively, thereby enhancing user trust and decision-making [44].

Adapting uncertainty-aware generation (UAG) techniques for open-domain question answering is another promising area. By refining uncertainty calibration in various contexts, researchers can enhance the reliability of LLM outputs, particularly where factual accuracy is crucial. Additionally, improving confidence calibration in model outputs and refining knowledge boundary definitions are vital for ensuring LLMs accurately assess limitations and provide well-reasoned responses [53].

AI-generated, for reference only.

Further improvements in model reasoning faithfulness through advanced training techniques and decomposition strategies will contribute to more accurate uncertainty estimation. By focusing on these areas, researchers can enhance LLMs' ability to generate outputs that are both faithful to input data and contextually appropriate [78]. Expanding benchmarks to include a wider range of tasks and languages will provide comprehensive evaluations of LLM performance and identify areas for uncertainty estimation improvement [79].

Developing models that generate well-reasoned refusal explanations is critical for enhancing transparency and user experience, especially in high-uncertainty situations, fostering greater trust in AI systems [80]. Exploring broader datasets and additional models will allow for thorough assessments of reasoning capabilities and uncertainty estimation techniques' effectiveness [81].

Incorporating semantic considerations into existing uncertainty estimation methods, such as G-NLL, can enhance robustness and applicability across diverse real-world scenarios. Addressing hallucinations at various LLM development stages and improving evaluation benchmarks will ensure LLMs are better equipped to handle uncertainty and provide reliable outputs. Future research should explore further improvements in calibrating LLM confidence scores and investigate applications across diverse NLP tasks [12].

7.3 Improving Model Robustness and Adaptability

Enhancing the robustness and adaptability of LLMs is crucial for reliable deployment across diverse applications. Future research should prioritize advancing effective adversarial training techniques to enhance machine learning model trustworthiness. Traditional evaluation metrics, such as model confidence and accuracy, often fail to instill human trust, especially in applications like text classification reliant on spurious correlations. Improving adversarial training methods can significantly increase model resilience against adversarial inputs, as evidenced by emerging techniques like TOKI, which automates trustworthiness assessment, and insights from longitudinal studies that highlight adversarial robustness complexities across model versions [82, 75, 55, 76]. Additionally, optimizing in-context learning (ICL) strategies for a broader range of programming problems can enhance LLMs' robustness and adaptability in generating secure code.

Exploring various model configurations, sabotage strategies, and expanding datasets to include diverse misinformation types are critical steps in advancing LLM robustness. Future work can also focus on developing hybrid models that combine LLMs with knowledge graphs and logic programming to enhance reasoning capabilities, which is vital for improving adaptability in complex reasoning tasks [83]. Furthermore, expanding datasets, testing additional LLM architectures, and exploring architectural modifications can enhance reasoning capabilities, improving robustness and adaptability [45].

Research should prioritize enhancing the attention steering process and exploring AutoPASTA's broader applicability beyond open-book question answering (QA), significantly improving model performance across various academic domains. The integration of AutoPASTA, which identifies and highlights key contextual information to guide LLM responses, has shown promise in increasing model faithfulness and performance. Given fine-tuned LLMs' potential to streamline labor-intensive processes like systematic literature reviews, investigating their effectiveness in diverse research settings is essential for ensuring methodological transparency and reliability in academic outputs [84, 14, 20]. The application of Fact and Reflection (FaR) prompting in different contexts holds promise for enhancing model robustness and adaptability, facilitating more reliable deployment in diverse scenarios. Additionally, applying confidence scores to improve dialogue policies and validating benchmarks on additional datasets can further contribute to LLM robustness.

Extending frameworks to multimodal models and developing context-aware generation methods are essential for enhancing AI systems' robustness and trustworthiness. Integrating abstain mechanisms with safety considerations and enhancing applicability across diverse domains can significantly bolster model robustness. This approach ensures LLMs generate reliable, contextually appropriate responses while addressing critical risks like biased content and adversarial attacks. Employing techniques such as safety-trained models establishing comprehensive guidelines and innovative adversarial training methods like Refusal Feature Adversarial Training (ReFAT) can improve LLM alignment with human values and enhance defensive capabilities against harmful outputs. Fine-tuning LLMs for specific applications, as demonstrated in systematic literature reviews, can streamline processes while maintaining high standards of factual accuracy and methodological transparency, broadening LLM utility across research domains [51, 20, 85].

Future research should explore methods to automate the alignment process further, incorporate user feedback more effectively, and develop strategies for handling evolving evaluation criteria [37]. Additionally, extending consistency assessment to include multi-class logic and unstructured contexts, alongside investigating effective knowledge graph integration in LLM reasoning, could provide significant advancements in model adaptability [83].

Focusing on enhancing source validation mechanisms, knowledge integration, and fine-tuning methodologies can significantly improve LLM robustness and adaptability, ensuring reliable deployment in applications like automated literature reviews and information retrieval systems while addressing critical challenges like hallucination and content credibility [14, 15, 20, 33].

7.4 Addressing Ethical and Societal Implications

The ethical and societal implications of LLMs necessitate thorough examination and strategies for responsible deployment. Future research should prioritize developing culturally sensitive LLMs that integrate diverse data sources, aligning with ethical standards across applications, including mental health [86]. This involves enhancing data diversity and aligning AI applications with cultural norms and ethical considerations.

The issue of hallucinations in LLMs requires further exploration, particularly in standardizing definitions and examining their sociotechnical dimensions in practical applications [87]. Addressing extrinsic hallucinations and refining measurement techniques are crucial for effective mitigation strategies [88]. Understanding why LLMs produce sycophantic responses despite access to factual knowledge remains an open question needing investigation [89].

The relationship between model size and copyright compliance presents another area for investigation, alongside expanding research to cover a broader range of authors and languages [73]. This is particularly relevant for developing reliable evaluation frameworks that incorporate machine ethics and adapt to multimodal large language model (MLLM) technology advancements [90].

Exploring MLLMs' robustness and ethical implications is vital, as unanswered questions in these areas suggest the need for comprehensive research to ensure ethical soundness and resilience [29]. Understanding biases' implications in diverse contexts can enhance user experience and inform the development of a taxonomy for human perceptions of LLMs [91].

Future research should refine evaluation criteria and incorporate dynamic value systems that adapt to evolving societal norms, ensuring LLMs remain relevant and ethically aligned [92]. Adapting frameworks like TOKI for other data types and conducting larger human-based evaluations can provide deeper insights into trustworthiness challenges [76].

Addressing ethical implications in legal practice and sensitive domains requires developing hybrid models incorporating external knowledge bases to enhance reasoning capabilities [31]. This approach, coupled with exploring ethical considerations and sustainability in LLM serving, can inform comprehensive ethical frameworks [67]. Future research should also focus on effective hallucination detection mechanisms, exploring the role of warnings, and understanding user engagement motivations beyond perceived accuracy [93].

Unanswered questions remain regarding the long-term societal impacts of LLMs and the effectiveness of proposed ethical frameworks in diverse contexts [32]. Future research should explore additional trustworthiness dimensions and develop comprehensive evaluation frameworks [47]. Investigating additional psychological phenomena and refining proposed taxonomies can deepen understanding of LLM behavior and incorporate metacognitive-like functionalities [94].

Finally, future improvements could focus on developing standardized practices for auditing and incident sharing, enhancing collaboration across organizations [95], and addressing unanswered questions regarding the operationalization of trust metrics and the ethical implications of extreme trust in AI systems [96]. Developing a flexible Requirements Engineering framework that adapts to evolving ethical guidelines and supports operationalizing trustworthiness in AI systems is crucial [97]. Future work aims to employ rigorous analysis by utilizing different LLMs and comparing responses to investigate alignment issues further [8].

AI-generated, for reference only.

8 Conclusion

This survey delves into the core issues of trustworthiness and reliability in large language models (LLMs), focusing on challenges like hallucinations, uncertainty, and ethical considerations. Hallucinations, primarily arising from factual recall failures and biases, pose a significant hurdle in LLM deployment. Methods such as SELF-FAMILIARITY have shown potential in mitigating these hallucinations, thereby enhancing the reliability and interpretability of LLM outputs. Techniques like the BatchEnsemble method efficiently detect hallucinations with minimal computational overhead, while real-time interpretable detection mechanisms provide crucial feedback for refining LLM robustness.

Innovative approaches like Checkability Training improve the clarity of LLM outputs, bolstering human trust in applications demanding high scrutiny. Pretrained language models (PLMs) demonstrate efficacy in distinguishing between unfaithful and faithful texts, underscoring the success of training algorithms in reducing hallucinations while maintaining text quality. The survey underscores the importance of regulatory frameworks, public AI literacy, and collaborative efforts to mitigate LLM-related risks while harnessing their benefits.

Output scouting emerges as a pivotal method for identifying harmful LLM responses, emphasizing the necessity of thorough auditing before model deployment. These mechanisms contribute to evaluating AI developers' trustworthiness, facilitating the AI field's maturation and effective regulation. Addressing LLM trustworthiness and reliability is essential for their ethical and effective deployment. Enhanced evaluation and mitigation strategies significantly improve model applicability across various domains, fostering greater confidence in their outputs.

The ISR method's success in reducing hallucinations in medical GQA highlights the importance of diverse mitigation strategies. Layer-wise analysis, exemplified by the LI method, is crucial for detecting unanswerable questions, underscoring the need for comprehensive information analysis across LLM layers. Key insights from this survey include recognizing hallucination as a critical issue and employing diverse mitigation strategies to enhance LLM reliability, alongside the observation that prompting models for explanations can bolster user trust without compromising performance.

References

[1] S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models, 2024.

[2] Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. Disinformation capabilities of large language models, 2024.

[3] Prabin Bhandari and Hannah Marie Brennan. Trustworthiness of children stories generated by large language models, 2023.

[4] Beatriz Cabrero-Daniel and Andrea Sanagustín Cabrero. Perceived trustworthiness of natural language generators, 2023.

[5] Manasi Sharma, Ho Chit Siu, Rohan Paleja, and Jaime D. Peña. Why would you suggest that? human trust in language model responses, 2024.

[6] Yuxin Xiao, Chaoqun Wan, Yonggang Zhang, Wenxiao Wang, Binbin Lin, Xiaofei He, Xu Shen, and Jieping Ye. Enhancing multiple dimensions of trustworthiness in llms via sparse activation control, 2024.

[7] Ziyan Guo, Li Xu, and Jun Liu. Trustworthy large models in vision: A survey, 2024.

[8] Gokul Puthumanaillam, Manav Vora, Pranay Thangeda, and Melkior Ornik. A moral imperative: The need for continual superalignment of large language models, 2024.

[9] Michael Mock, Sebastian Schmidt, Felix Müller, Rebekka Görge, Anna Schmitz, Elena Haedecke, Angelika Voss, Dirk Hecker, and Maximillian Poretschkin. Developing trustworthy ai applications with foundation models, 2024.

[10] Geoff Keeling and Winnie Street. On the attribution of confidence to large language models, 2024.

[11] Rajkumar Ramamurthy, Meghana Arakkal Rajeev, Oliver Molenschot, James Zou, and Nazneen Rajani. Veritas: A unified approach to reliability evaluation, 2024.

[12] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. Can unconfident llm annotations be used for confident conclusions?, 2025.

[13] Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning, 2023.

[14] Xuemei Tang, Xufeng Duan, and Zhenguang G. Cai. Are llms good literature review writers? evaluating the literature review writing ability of large language models, 2025.

[15] Zhiyuan Chang, Mingyang Li, Xiaojun Jia, Junjie Wang, Yuekai Huang, Qing Wang, Yihao Huang, and Yang Liu. What external knowledge is preferred by llms? characterizing and exploring chain of evidence in imperfect context, 2024.

[16] Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of llms in rag through grounded attributions and learning to refuse, 2024.

[17] Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales, 2024.

[18] Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. Rethinking uncertainty estimation in natural language generation, 2024.

[19] Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. On the universal truthfulness hyperplane inside llms, 2024.

[20] Teo Susnjak, Peter Hwang, Napoleon H. Reyes, Andre L. C. Barczak, Timothy R. McIntosh, and Surangika Ranathunga. Automating research synthesis with domain-specific large language model fine-tuning, 2024.

[21] Haoyan Yang, Yixuan Wang, Xingyin Xu, Hanyuan Zhang, and Yirong Bian. Can we trust llms? mitigate overconfidence bias in llms through knowledge transfer, 2024.

[22] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models, 2024.

[23] Oliver Bentham, Nathan Stringham, and Ana Marasović. Chain-of-thought unfaithfulness as disguised accuracy, 2024.

[24] Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows", 2024.

[25] Weijia Zhang, Mohammad Aliannejadi, Jiahuan Pei, Yifei Yuan, Jia-Hong Huang, and Evangelos Kanoulas. A comparative analysis of faithfulness metrics and humans in citation evaluation, 2024.

[26] Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations, 2024.

[27] Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? *arXiv preprint arXiv:2401.07927*, 2024.

[28] Noah Y. Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models, 2024.

[29] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models, 2024.

[30] Juhyun Oh, Eunsu Kim, Inha Cha, and Alice Oh. The generative ai paradox on evaluation: What it can solve, it may not evaluate, 2024.

[31] Eliza Mik. Caveat lector: Large language models in legal practice. *Rutgers Bus. LJ*, 19:70, 2023.

[32] Junfeng Jiao, Saleh Afroogh, Yiming Xu, and Connor Phillips. Navigating llm ethics: Advancements, challenges, and future directions, 2024.

[33] Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, and Wei Lu. Know where to go: Make llm a relevant, responsible, and trustworthy searcher, 2023.

[34] Yahan Li, Yi Wang, Yi Chang, and Yuan Wu. Xtrust: On the multilingual trustworthiness of large language models, 2024.

[35] Xiaonan Jing, Srinivas Billa, and Danny Godbout. On a scale from 1 to 5: Quantifying hallucination in faithfulness evaluation, 2025.

[36] Da Song, Xuan Xie, Jiayang Song, Derui Zhu, Yuheng Huang, Felix Juefei-Xu, and Lei Ma. Luna: A model-based universal analysis framework for large language models, 2024.

[37] Shreya Shankar, J. D. Zamfirescu-Pereira, Björn Hartmann, Aditya G. Parameswaran, and Ian Arawjo. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences, 2024.

[38] Nik Bear Brown. Enhancing trust in llms: Algorithms for comparing and interpreting llms, 2024.

[39] Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection, 2024.

[40] Jiaqi Li, Yixuan Tang, and Yi Yang. Know the unknown: An uncertainty-sensitive method for llm instruction tuning. *arXiv preprint arXiv:2406.10099*, 2024.

[41] Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms, 2024.

[42] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024.

[43] Tosin Adewumi, Nudrat Habib, Lama Alkhaled, and Elisa Barney. On the limitations of large language models (llms): False attribution, 2024.

[44] Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*, 2024.

[45] Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. Llm internal states reveal hallucination risk faced with a query, 2024.

[46] Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. A survey of uncertainty estimation in llms: Theory meets practice. *arXiv preprint arXiv:2410.15326*, 2024.

[47] Emad A. Alghamdi, Reem I. Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. Aratrust: An evaluation of trustworthiness for llms in arabic, 2024.

[48] Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness, 2023.

[49] He Li, Haoang Chi, Mingyu Liu, and Wenjing Yang. Look within, why llms hallucinate: A causal perspective, 2024.

[50] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2024.

[51] Yi Luo, Zhenghao Lin, Yuhao Zhang, Jiashuo Sun, Chen Lin, Chengjin Xu, Xiangdong Su, Yelong Shen, Jian Guo, and Yeyun Gong. Ensuring safe and high-quality outputs: A guideline library approach for language models, 2024.

[52] Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. The critique of critique, 2024.

[53] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *arXiv preprint arXiv:2312.07000*, 2023.

[54] Melissa Ailem, Katerina Marazopoulou, Charlotte Siska, and James Bono. Examining the robustness of llm evaluation to the distributional assumptions of benchmarks, 2024.

[55] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Trustllm: Trustworthiness in large language models, 2024.

[56] Anh Thu Maria Bui, Saskia Felizitas Brech, Natalie Hußfeldt, Tobias Jennert, Melanie Ullrich, Timo Breuer, Narjes Nikzad Khasmakhi, and Philipp Schaer. The two sides of the coin: Hallucination generation and detection with llms as evaluators for llms, 2024.

[57] Yuan Sui, Yufei He, Zifeng Ding, and Bryan Hooi. Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering, 2025.

[58] Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. Sage: Evaluating moral consistency in large language models, 2024.

[59] Yi-Jyun Sun, Suvodip Dey, Dilek Hakkani-Tur, and Gokhan Tur. Confidence estimation for llm-based dialogue state tracking, 2024.

[60] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. Fact-checking the output of large language models via token-level uncertainty quantification, 2024.

[61] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*, 2023.

[62] Wrick Talukdar and Anjanava Biswas. Improving large language model (llm) fidelity through context-aware grounding: A systematic approach to reliability and veracity, 2024.

[63] Gabriel Y Arteaga, Thomas B Schön, and Nicolas Pielawski. Hallucination detection in llms: Fast and memory-efficient finetuned models. *arXiv preprint arXiv:2409.02976*, 2024.

[64] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models, 2023.

[65] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, Yunjun Gao, Yingcai Wu, and Shouling Ji. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents, 2024.

[66] Duy Khoa Pham and Bao Quoc Vo. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models, 2024.

[67] Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. Llm inference serving: Survey of recent advances and opportunities, 2024.

[68] Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence, 2024.

[69] Xinxi Chen, Li Wang, Wei Wu, Qi Tang, and Yiyao Liu. Honest ai: Fine-tuning "small" language models to say "i don't know", and reducing hallucination in rag, 2024.

[70] Q. Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap, 2023.

[71] Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate, 2024.

[72] Minda Hu, Bowei He, Yufei Wang, Liangyou Li, Chen Ma, and Irwin King. Mitigating large language model hallucination with faithful finetuning, 2024.

[73] Felix B Mueller, Rebekka Görge, Anna K Bernzen, Janna C Pirk, and Maximilian Poretschkin. Llms and memorization: On quality and specificity of copyright compliance, 2024.

[74] Baolong Bi, Shenghua Liu, Yiwei Wang, Lingrui Mei, Junfeng Fang, Hongcheng Gao, Shiyu Ni, and Xueqi Cheng. Is factuality enhancement a free lunch for llms? better factuality can lead to worse context-faithfulness, 2024.

[75] Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models, 2024.

[76] Lam Nguyen Tung, Steven Cho, Xiaoning Du, Neelofar Neelofar, Valerio Terragni, Stefano Ruberto, and Aldeida Aleti. Automated trustworthiness oracle generation for machine learning text classifiers, 2024.

[77] Todor Ivanov and Valeri Penchev. Ai benchmarks and datasets for llm evaluation, 2024.

[78] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question decomposition improves the faithfulness of model-generated reasoning, 2023.

[79] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball, 2023.

[80] Yuhao Wang, Zhiyuan Zhu, Heyang Liu, Yusheng Liao, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Drawing the line: Enhancing trustworthiness of mllms through the power of refusal, 2024.

[81] Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination?, 2024.

[82] Yugeng Liu, Tianshuo Cong, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. Robustness over time: Understanding adversarial examples' effectiveness on longitudinal versions of large language models, 2024.

[83] Bishwamittra Ghosh, Sarah Hasan, Naheed Anjum Arafat, and Arijit Khan. Logical consistency of large language models in fact-checking. *arXiv preprint arXiv:2412.16100*, 2024.

[84] Qingru Zhang, Xiaodong Yu, Chandan Singh, Xiaodong Liu, Liyuan Liu, Jianfeng Gao, Tuo Zhao, Dan Roth, and Hao Cheng. Model tells itself where to attend: Faithfulness meets automatic attention steering, 2024.

[85] Lei Yu, Virginie Do, Karen Hambardzumyan, and Nicola Cancedda. Robust llm safeguarding via refusal feature adversarial training, 2024.

[86] Neo Christopher Chung, George Dyer, and Lennart Brocki. Challenges of large language models for mental health counseling, 2023.

[87] Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. An audit on the perspectives and challenges of hallucinations in nlp, 2024.

[88] Song Wang, Xun Wang, Jie Mei, Yujia Xie, Sean Muarray, Zhang Li, Lingfeng Wu, Si-Qing Chen, and Wayne Xiong. Developing a reliable, general-purpose hallucination detection and mitigation service: Insights and lessons learned, 2024.

[89] Aswin RRV, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. Chaos with keywords: Exposing large language models sycophantic hallucination to misleading keywords and evaluating defense strategies, 2024.

[90] Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, Hang Su, Yinpeng Dong, and Jun Zhu. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models, 2024.

[91] Lu Wang, Max Song, Rezvaneh Rezapour, Bum Chul Kwon, and Jina Huh-Yoo. People's perceptions toward bias and related concepts in large language models: A systematic review, 2024.

[92] Gwenyth Isobel Meadows, Nicholas Wai Long Lau, Eva Adelina Susanto, Chi Lok Yu, and Aditya Paul. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models, 2024.

[93] Mahjabin Nahar, Haeseung Seo, Eun-Ju Lee, Aiping Xiong, and Dongwon Lee. Fakes of varying shades: How warning affects human perception and engagement regarding llm hallucinations, 2024.

[94] Elijah Berberette, Jack Hutchins, and Amir Sadovnik. Redefining "hallucination" in llms: Towards a psychology-informed framework for mitigating misinformation, 2024.

[95] Shahar Avin, Haydn Belfield, Miles Brundage, Gretchen Krueger, Jasmine Wang, Adrian Weller, Markus Anderljung, Igor Krawczuk, David Krueger, Jonathan Lebensold, Tegan Maharaj, and Noa Zilberman. Filling gaps in trustworthy development of ai, 2021.

[96] Sivan Schwartz, Avi Yaeli, and Segev Shlomov. Enhancing trust in llm-based ai automation agents: New considerations and future challenges, 2023.

[97] Krishna Ronanki, Beatriz Cabrero-Daniel, Jennifer Horkoff, and Christian Berger. Re-centric recommendations for the development of trustworthy(er) autonomous systems, 2024.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.