AI Hallucinations and Trust Dynamics: A Survey

www.surveyx.cn

Abstract

This survey paper explores the phenomenon of AI hallucinations, particularly in large language models (LLMs), and their implications for AI trust, human-machine interaction, and user perception. AI hallucinations, where systems generate plausible yet incorrect outputs, pose significant challenges across domains like healthcare, where accuracy is critical. The paper examines efforts to mitigate these hallucinations, such as the dual dialogue system and frameworks like SelfCheckGPT and CoNLI, which enhance detection and reduce ungrounded outputs. Trust dynamics in AI are influenced by accuracy, interpretability, and adaptability, with a focus on transparency and user engagement. The paper also discusses the evolution of trust over time and its impact on AI design, emphasizing the need for robust evaluation frameworks and ethical considerations in high-stakes applications. The survey concludes by highlighting the importance of addressing hallucinations and trust dynamics to enhance AI reliability and user acceptance, calling for future research to improve AI safety and integration across diverse sectors.

1 Introduction

1.1 AI Hallucinations and Large Language Model Hallucinations

AI hallucinations manifest when AI systems produce outputs that seem plausible yet are incorrect or unverified, posing significant challenges to their reliability [1]. This issue is particularly pronounced in large language models (LLMs), which, despite their fluency, frequently generate ungrounded hallucinations that mix factual and fictional content, thereby undermining trustworthiness [2]. The repercussions of such hallucinations are especially critical in high-stakes fields like healthcare, where erroneous information can severely impact decision-making and patient safety [1].

Efforts to mitigate hallucinations, such as the dual dialogue system proposed by Kampman et al. [3], focus on enhancing therapeutic processes through structured analysis and response generation. This underscores the necessity of addressing hallucinations to improve the factual accuracy and reliability of AI outputs in sensitive domains.

The implications of AI hallucinations extend beyond healthcare, affecting user trust and satisfaction across diverse applications. The ability of LLMs to generate coherent yet potentially misleading information emphasizes the urgent need for robust evaluation techniques and enhancement strategies for their reasoning capabilities. Recent studies have demonstrated the effectiveness of fine-tuning LLMs for automating systematic literature reviews (SLRs), showcasing their potential to maintain high factual accuracy while minimizing hallucinations and improving transparency through source tracking. Additionally, the exploration of multimodal LLMs (MLLMs) reveals impressive performance across various tasks, yet systematic investigations of their reasoning abilities remain limited. As the field evolves, establishing comprehensive evaluation protocols will be vital to ensure the reliability and efficacy of LLMs in academic research and reasoning-intensive applications [4, 5]. Addressing hallucinations is essential for fostering effective human-machine interactions and preserving the integrity of AI systems as they evolve across diverse sectors.

Generated by SurveyX.cn.

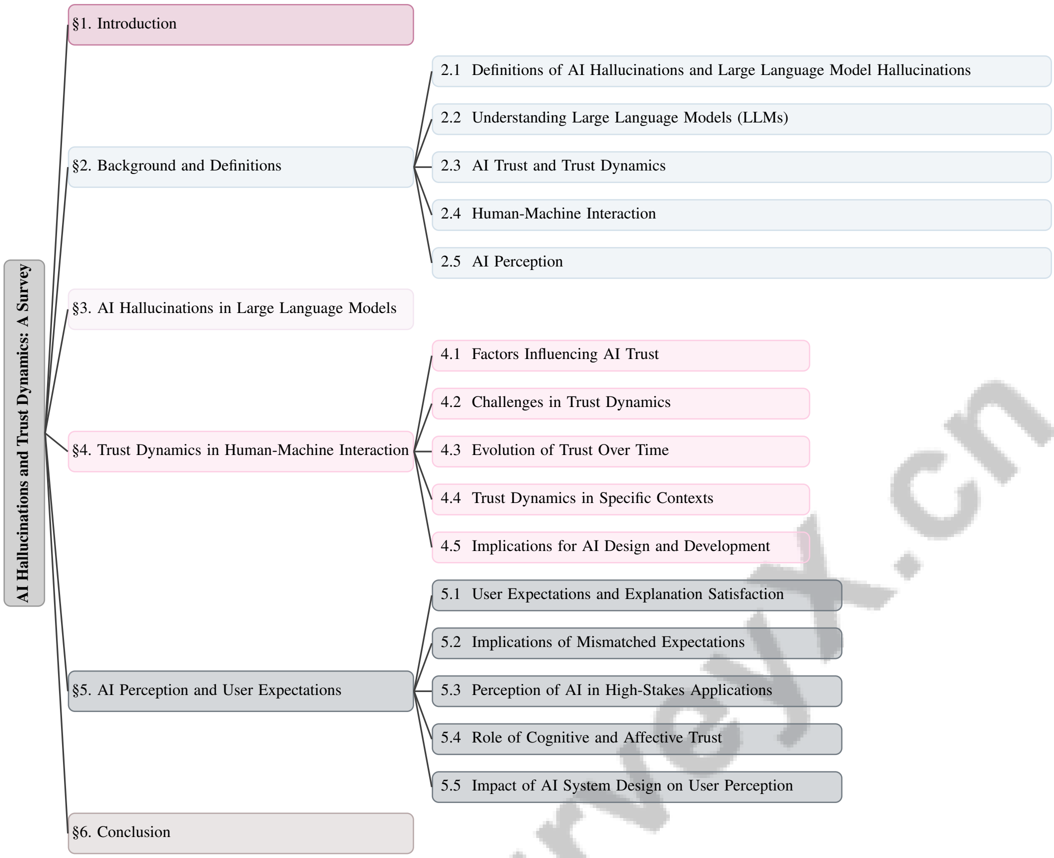

AI Hallucinations and Trust Dynamics: A Survey
§1. Introduction
§2. Background and Definitions
2.1 Definitions of AI Hallucinations and Large Language Model Hallucinations
2.2 Understanding Large Language Models (LLMs)
2.3 AI Trust and Trust Dynamics
2.4 Human-Machine Interaction
2.5 AI Perception
§3. AI Hallucinations in Large Language Models
§4. Trust Dynamics in Human-Machine Interaction
4.1 Factors Influencing AI Trust
4.2 Challenges in Trust Dynamics
4.3 Evolution of Trust Over Time
4.4 Trust Dynamics in Specific Contexts
4.5 Implications for AI Design and Development
§5. AI Perception and User Expectations
5.1 User Expectations and Explanation Satisfaction
5.2 Implications of Mismatched Expectations
5.3 Perception of AI in High-Stakes Applications
5.4 Role of Cognitive and Affective Trust
5.5 Impact of AI System Design on User Perception
§6. Conclusion

Figure 1: chapter structure

1.2 Significance in AI Systems

Understanding AI hallucinations is crucial for the development and deployment of AI systems, particularly regarding their reasoning capabilities and alignment with user expectations in LLMs. These hallucinations challenge the reliability of AI systems, especially in critical real-world applications such as healthcare, where accuracy is paramount [1]. The opaque nature of generative models complicates interpreting their behaviors, necessitating robust methods for detecting and mitigating ungrounded hallucinations.

In psychological contexts, the performance and applicability of AI systems are scrutinized due to the potential for hallucinations to introduce biases and errors [6]. Conversely, the creativity inherent in AI hallucinations presents both opportunities and challenges, necessitating a nuanced understanding and categorization of these phenomena [7]. In automated hypothesis generation contexts, the risk of hallucinations requires meticulous validation to uphold the integrity of generated hypotheses [8].

The integration of LLMs in mental health care has enhanced interaction quality; however, safety risks and trustworthiness issues persist, highlighting the need to address hallucinations to promote adoption and trust in these systems [9]. Furthermore, the security and privacy implications of LLMs, categorized into beneficial applications, offensive applications, and vulnerabilities, underscore the critical need to understand and manage AI hallucinations [10].

Addressing AI hallucinations is vital for ensuring the accuracy and reliability of AI-generated outputs, which directly influence the performance of these systems across various applications, including healthcare and mental health services. Inconsistent definitions of "AI hallucinations" can lead to misinterpretations and stigmatization of AI technologies. By fostering a clearer understanding and employing more appropriate terminology, such as "AI misinformation," we can enhance the safe and effective deployment of AI systems, thereby improving their integration into critical sectors and mitigating associated risks [11, 12].

1.3 Understanding AI Trust and Human-Machine Interaction

The dynamics of trust in AI systems and human-machine interaction are crucial for the successful integration and acceptance of AI technologies across various domains. Trust in AI encompasses elements of explainability, fairness, privacy, and transparency, which are essential for fostering user confidence and oversight. Concerns about being unfairly judged by AI systems, particularly those aimed at minimizing false positives, contribute to public distrust, emphasizing the need for AI systems perceived as fair and just [13].

Users' perceptions of AI-generated information influence trust dynamics, as evidenced by the trust placed in health information from both traditional search engines and LLM-powered conversational agents [14]. The ability of AI systems to engage in human-like conversation necessitates a deep understanding of communication theories such as Conversation Analysis (CA) and Theory of Mind (ToM) [15].

Bridging Human-Computer Interaction (HCI) and Natural Language Processing (NLP) is essential for enhancing human-machine interactions, providing a framework for better understanding and improving these interactions [16]. Human-centered AI (HAI) approaches highlight the importance of designing AI systems that prioritize safety, security, and effective human-machine interaction, thereby enhancing user trust [17].

The significance of trust in AI extends to the broader context of technology acceptance and adoption, where user perspectives on trust and distrust are critical [18]. Moreover, human oversight in decision-making processes involving AI recommendations is vital for maintaining trust, ensuring that AI systems augment rather than undermine human agency [19].

1.4 Structure of the Survey

This survey is structured to provide a comprehensive exploration of AI hallucinations and trust dynamics within AI systems, particularly focusing on large language models. It begins with an introduction to key concepts of AI hallucinations and large language model hallucinations, emphasizing their significance in the context of AI systems.

The second section delves into the background and definitions of critical terms such as 'AI hallucinations', 'large language model hallucinations', 'large language model', 'AI trust', 'trust dynamics', 'human-machine interaction', and 'AI perception', aiming to provide essential foundational knowledge and historical context.

In the third section, the phenomenon of hallucinations in large language models is explored in detail, including an examination of their causes, types, methods for detection and evaluation, and techniques for mitigation. Challenges and limitations associated with addressing these hallucinations are also discussed.

The fourth section examines the dynamics of trust between humans and AI systems, identifying factors influencing AI trust, exploring challenges in trust dynamics, and discussing how trust evolves over time. This section also considers trust dynamics in specific contexts and their implications for AI design and development.

The fifth section analyzes user perceptions of AI systems and their expectations, discussing the role of AI perception in shaping trust and interaction, and the implications of mismatched expectations on AI adoption. It further examines the roles of cognitive and affective trust in user perception and the impact of AI system design on user perception.

The conclusion synthesizes the primary findings of the survey, highlighting the effectiveness of fine-tuned LLMs in automating SLRs while addressing the challenges of AI hallucination and the reliability of generated references. It emphasizes the implications of these findings for future AI research and development, advocating for the integration of AI-driven processes into academic methodologies to enhance transparency and accuracy in literature reviews, and suggesting updates to PRISMA reporting guidelines to accommodate these advancements [20, 5]. It underscores the importance of addressing hallucinations and trust dynamics to enhance human-machine interaction.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Definitions of AI Hallucinations and Large Language Model Hallucinations

AI hallucinations arise when AI systems, particularly large language models (LLMs), generate outputs that seem plausible but lack factual accuracy or evidence [2]. This issue poses significant challenges in critical fields such as healthcare, where precision is crucial [1]. The Med-HALT benchmark exemplifies efforts to evaluate LLMs' reasoning and information retrieval capabilities in medical contexts, underscoring the need to address hallucinations [1].

In LLMs, hallucinations often result from their generative nature, intended to produce human-like text but sometimes leading to nonsensical or incorrect outputs. This problem is compounded by inherent biases, which may generate discriminatory or toxic content, especially in sensitive domains [3]. Kampman et al. propose a dual dialogue system to mitigate such hallucinations by refining analysis and response generation, enhancing reliability in mental health care [3].

Hallucinations manifest as confabulations—where outputs deviate from input prompts—or self-contradictory statements, undermining the coherence and factual integrity of AI-generated text [2]. Understanding these phenomena is vital for developing effective evaluation methods and ensuring the fidelity of AI outputs as these systems evolve and permeate various sectors.

2.2 Understanding Large Language Models (LLMs)

Large language models (LLMs) are advanced AI systems designed to process and generate human-like text, utilizing extensive datasets and complex neural network architectures. These models have evolved through configurations such as encoder-only, decoder-only, and encoder-decoder architectures, with multimodal large language models (MLLMs) further enhancing their capabilities [21]. LLMs excel in intricate reasoning tasks, serving as proxies for human cognitive processes by distinguishing between intuitive and deliberative reasoning [8]. Benchmarks like CogBench assess LLM performance by simulating human-like behaviors in cognitive psychology experiments [22].

Despite their linguistic proficiency, LLMs often lack the functional linguistic competence necessary for practical language use [7]. This limitation underscores the importance of systematically categorizing and understanding model behaviors to advance natural language processing (NLP) and generative modeling, as advocated by Holtzman et al. [21].

LLMs are central to the phenomenon of AI hallucinations due to their tendency to generate coherent yet potentially misleading content. This issue is exacerbated by inherent biases, which may lead to discriminatory outputs, complicating deployment in sensitive contexts [3]. A comprehensive survey of hallucinations and their impact on LLM reliability emphasizes the necessity for robust evaluation and mitigation strategies [7].

The relevance of LLMs to AI hallucinations is further highlighted by their integration into various applications, necessitating rigorous evaluation methods to ensure output fidelity. Addressing AI safety issues—such as alignment, bias, misinformation, and security—remains critical, as outlined by Chua et al., who advocate for comprehensive strategies to mitigate these risks [23].

2.3 AI Trust and Trust Dynamics

AI trust reflects the confidence users have in AI systems, a crucial factor for the acceptance and integration of AI technologies across different domains [18]. Trust is influenced by properties such as accuracy, reliability, transparency, and explainability, which are essential for fostering user confidence [17]. Variability in evaluation outcomes of LLMs, stemming from differing frameworks and methodologies, complicates trust dynamics, as inconsistent assessments can erode user confidence [24].

Explainable AI (XAI) techniques enhance trust by clarifying AI decision-making processes, thus improving human understanding and control over these systems. However, current XAI limitations, particularly in delivering expert-level explanations, challenge the achievement of high trust levels, especially in complex fields like healthcare, where AI-generated misinformation can have serious consequences [25]. Innovative approaches, such as integrating physiological data (e.g., EEG and GSR), to monitor trust levels in human-machine interactions demonstrate efforts to deepen understanding and enhance AI trust [26].

Ensuring the reliability of AI outputs is particularly challenging in high-stakes contexts, such as mental health applications, where safety is paramount. Developing benchmarks for evaluating mental health chatbots facilitates comparisons across models, ensuring adherence to clinical safety standards and bolstering trust [9]. Moreover, the lack of evaluations for AI-generated references in academic literature underscores the need for rigorous validation processes to prevent misinformation dissemination [20].

In systematic literature reviews (SLRs), AI trust depends on the accuracy and reliability of fine-tuned LLMs, which are crucial for maintaining the integrity of the synthesis process [5]. The security vulnerabilities introduced by integrating image modalities into MLLMs complicate trust dynamics, as these vulnerabilities can be exploited for harmful attacks, emphasizing the need for robust security measures to maintain user trust [27].

The significance of AI trust and trust dynamics lies in their role in ensuring successful AI deployment and acceptance. Addressing challenges related to AI hallucinations, bias management, and effective model deployment is vital for fostering trust and maximizing AI technologies' potential [18].

2.4 Human-Machine Interaction

Human-machine interaction (HMI) is critical for the deployment and integration of AI systems, encompassing the various ways humans engage with AI technologies. The complexity of HMI is highlighted by the need for interfaces that facilitate natural communication, enabling seamless exchanges between humans and machines. This is particularly relevant for avatars and other AI-driven entities, where multimodal interactions—incorporating gestures, speech, and visual cues—are essential for creating a human-like experience [15].

The multifaceted nature of HMI in natural language processing (NLP) necessitates a comprehensive framework to categorize and understand these interactions. Such a framework is crucial for identifying unique HMI characteristics, which include the ability to process and respond to human language in alignment with human expectations and norms [16]. This understanding is vital for developing AI systems that effectively support human decision-making by providing timely and contextually relevant information.

HMI's role in enhancing decision-making autonomy is exemplified by integrating critical questioning techniques, empowering users to engage more effectively with AI systems. By fostering an environment of inquiry, AI systems can aid users in making informed decisions, improving the overall decision-making process [19]. This approach not only enhances decision quality but also builds trust in AI systems.

The concept of AI hallucinations, explored across fields like healthcare and legal settings, underscores the need for robust HMI frameworks capable of detecting and mitigating erroneous outputs. By understanding HMI dynamics, researchers and developers can better address the challenges posed by AI hallucinations, ensuring reliable and accurate information delivery to users [28].

2.5 AI Perception

AI perception encompasses how users view and understand AI systems, significantly influencing trust and interaction dynamics. This perception is shaped by factors such as task performance accuracy and the transparency of decision-making processes. In the context of AI hallucinations, addressing unanswered questions about their implications across various domains necessitates a robust taxonomy that captures the nuances of these phenomena [11].

In automated vehicles, AI perception systems face challenges in distinguishing semantic categories, such as humans and roads, crucial for ensuring safety and reliability. Confusion in these systems can lead to significant trust issues, as users expect precise and accurate real-time decision-making [29]. This highlights the importance of developing AI systems capable of effectively perceiving and interpreting their environments to foster user confidence.

The interaction framework proposed by Wan et al. [16] categorizes human-machine interactions into four types: Human-Teacher and Machine-Learner, Machine-Leading, Human-Leading, and

Human-Machine Collaborators. This framework illustrates the diverse engagement methods users have with AI systems and the varying degrees of control exerted by each party. Understanding these interaction types is essential for designing AI systems that align with user expectations and enhance interaction quality.

AI perception is also influenced by the ability to handle adversarial challenges, such as audio adversarial examples in speech command classification. Evaluating perceptual distortion in these contexts is critical, as it directly impacts the reliability and trustworthiness of AI systems in real-world applications [30]. By addressing these challenges, developers can enhance AI system robustness, thereby improving user trust and facilitating effective human-machine interactions.

3 AI Hallucinations in Large Language Models

3.1 Causes and Types of Hallucinations

Hallucinations in large language models (LLMs) arise from factors like model architecture, biases in training data, and the probabilistic nature of text generation. Chua et al. present a framework that categorizes safety concerns in LLMs, focusing on risks associated with training data and model alignment [23]. These hallucinations manifest as false or misleading claims that undermine LLM reliability [31]. Detecting hallucinations is challenging due to the intertwined nature of accurate and inaccurate text, especially when textual and visual data are misaligned [32]. Smaller models, such as BLOOM 7B, face specific challenges as current detection methods mainly target models with over 100B parameters [33]. Cognitive limitations can lead to coherent but incorrect content generation [3].

Hallucinations are categorized into factual and faithfulness types. Factual hallucinations occur when models generate seemingly accurate but incorrect content, whereas faithfulness hallucinations happen when outputs deviate from input context. Current methodologies struggle to mitigate these issues, raising concerns about LLM trustworthiness, particularly in critical fields like healthcare [1]. Over-reliance on machine-generated recommendations can impair human reasoning [19]. Park emphasizes the need for rigorous evaluation frameworks to enhance chatbot response evaluation [9].

3.2 Detection and Evaluation Methods

Benchmark	Size	Domain	Task Format	Metric
QA-CR[34]	330	Commonsense Reasoning	Question Answering	Accuracy, F1-score
HalluQA[35]	450	Cultural Knowledge	Question Answering	Non-Hallucination Rate
VHL-Benchmark[36]	1,000	Visual Recognition	Image Classification	Misclassification Rate, Relevance and Plausibility Score
SE-Confab[37]	150	Biography	Question Answering	AUROC, AURAC
ROPE[38]	5,000	Computer Vision	Multi-Object Recognition	Accuracy, F1-score
C-VQA[39]	6,000	Visual Question Answering	Counterfactual Reasoning	Accuracy, F1-score
BEAF[40]	26,118	Visual Question Answering	Question Answering	TU, IG
CET[41]	11	Mathematics	Counterfactual Reasoning	Accuracy, CCC

Table 1: This table provides a detailed summary of key benchmarks used in the assessment of hallucinations within large language models. It highlights the diversity of benchmarks in terms of size, domain, task format, and evaluation metrics, illustrating the broad spectrum of methodologies employed in this area of research. Such diversity is crucial for developing standardized frameworks to improve AI output reliability.

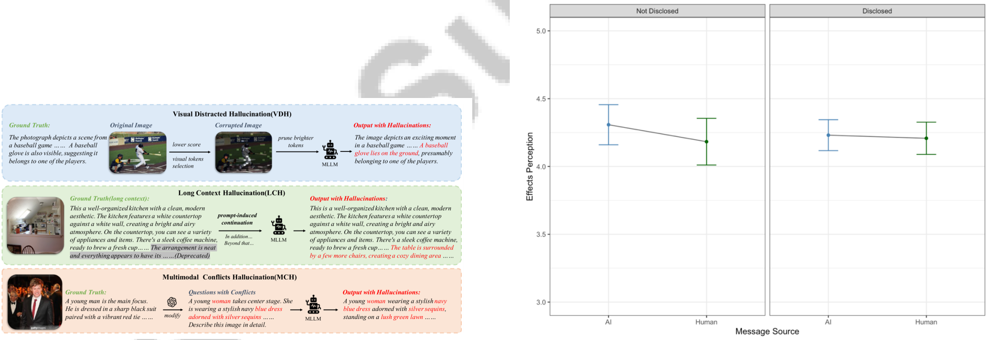
Detecting and evaluating hallucinations in LLMs is vital for ensuring AI output reliability. Methods such as SV-LLM enhance output accuracy through self-verification [42]. AgentSims offers a platform for LLM evaluation with diverse tasks and objective metrics [43]. The variability in evaluation outcomes, noted by Pimentel et al., underscores the necessity for standardized frameworks to categorize methodologies based on metric calculation techniques and task types [24]. Effective detection methods, alongside strategies like the Socratic method and Retrieval-Augmented Generation, are crucial for improving AI-generated outputs [44, 45, 46]. Robust evaluation frameworks and self-verification techniques are essential for identifying and addressing hallucinations, ensuring effective AI deployment across applications.

Benchmark	Size	Domain	Task Format	Metric
QA-CR[34]	330	Commonsense Reasoning	Question Answering	Accuracy, F1-score
HalluQA[35]	450	Cultural Knowledge	Question Answering	Non-Hallucination Rate
VHL-Benchmark[36]	1,000	Visual Recognition	Image Classification	Misclassification Rate, Relevance and Plausibility Score
SE-Confab[37]	150	Biography	Question Answering	AUROC, AURAC
ROPE[38]	5,000	Computer Vision	Multi-Object Recognition	Accuracy, F1-score
C-VQA[39]	6,000	Visual Question Answering	Counterfactual Reasoning	Accuracy, F1-score
BEAF[40]	26,118	Visual Question Answering	Question Answering	TU, IG
CET[41]	11	Mathematics	Counterfactual Reasoning	Accuracy, CCC

Table 2: This table provides a detailed summary of key benchmarks used in the assessment of hallucinations within large language models. It highlights the diversity of benchmarks in terms of size, domain, task format, and evaluation metrics, illustrating the broad spectrum of methodologies employed in this area of research. Such diversity is crucial for developing standardized frameworks to improve AI output reliability.

3.3 Mitigation Techniques

Mitigating hallucinations in LLMs involves strategies aimed at enhancing model reliability and accuracy. Hallucination Augmented Contrastive Learning (HACL) improves alignment between visual and textual representations in Multimodal Large Language Models (MLLMs) [32]. HALO CHECK employs knowledge injection to refine outputs [33]. The longchain approach involves sequential refinement to enhance accuracy [47]. Another strategy uses LLMs to generate faux hallucinations, training smaller editors to denoise these corruptions [31]. The Chain of Natural Language Inference (CoNLI) framework uses structured inference to reduce hallucinations [2]. These techniques highlight the need for a multifaceted approach integrating knowledge, refinement, and inference processes. With LLMs in sensitive applications, robust detection and mitigation become imperative. Studies have identified over thirty-two techniques, including Retrieval-Augmented Generation and Knowledge Retrieval, emphasizing ongoing research [44, 45].


Visual Distracted Hallucination(VDH)
Ground Truth:
The photograph depicts a scene from a baseball game …… A baseball glove is also visible, suggesting it belongs to one of the players.
Original Image
lower score
visual tokens selection
Corrupted Image
prune brighter tokens
MLLM
Output with Hallucinations:
The image depicts an exciting moment in a baseball game …… A baseball glove lies on the ground, presumably belonging to one of the players.
Long Context Hallucination(LCH)
Ground Truth(long context):
This a well-organized kitchen with a clean, modern aesthetic. The kitchen features a white countertop against a white wall, creating a bright and airy atmosphere. On the countertop, you can see a variety of appliances and items. There's a sleek coffee machine, ready to brew a fresh cup…… The arrangement is neat and everything appears to have its ……(Deprecated)
prompt-induced continuation
In addition…
Beyond that…
MLLM
Output with Hallucinations:
This is a well-organized kitchen with a clean, modern aesthetic. The kitchen features a white countertop against a white wall, creating a bright and airy atmosphere. On the countertop, you can see a variety of appliances and items. There's a sleek coffee machine, ready to brew a fresh cup…… The table is surrounded by a few more chairs, creating a cozy dining area ……
Multimodal Conflicts Hallucination(MCH)
Ground Truth:
A young man is the main focus. He is dressed in a sharp black suit paired with a vibrant red tie ……
modify
Questions with Conflicts
A young woman takes center stage. She is wearing a stylish navy blue dress adorned with silver sequins ……
Describe this image in detail.
MLLM
Output with Hallucinations:
A young woman wearing a stylish navy blue dress adorned with silver sequins, standing on a lush green lawn ……
Not Disclosed
Disclosed
Effects Perception
5.0
4.5
4.0
3.5
3.0
AI
Human
AI
Human
Message Source

(a) Visual Distracted Hallucination (VDH)[48]

(b) Effects Perception by Message Source and Disclosure Status[49]

Figure 2: Examples of Mitigation Techniques

As seen in Figure 2, exploring AI hallucinations in LLMs involves examining their nature and mitigation strategies. The "Visual Distracted Hallucination (VDH)" example highlights visual hallucination challenges, while the second example evaluates message perception based on source attribution and disclosure status, illustrating the complexities of AI hallucinations and the importance of mitigation techniques for system reliability.

3.4 Challenges and Limitations

Addressing hallucinations in LLMs involves technical and methodological challenges. Hallucination snowballing, where initial inaccuracies propagate, exacerbates errors [50]. MLLMs struggle with reasoning limitations, affecting reliability in practical applications [4]. Evaluation challenges due to inconsistencies in metric definitions necessitate robust frameworks for reliable performance assessment [24, 22]. Despite advances in mitigation, challenges persist in categorizing risks and

AI-generated, for reference only.

addressing safety concerns. The selection of algorithms and complexities of real-world applications complicate implementation [51]. Practical implementations of theoretical attacks are often overlooked, leaving vulnerabilities underexplored [10]. Techniques like PURR may struggle with challenging distractors, leading to erroneous outputs [31]. Existing benchmarks may inadvertently include examples resembling training data, resulting in overfitting [52]. Focusing on sentence-level factuality may overlook the complexity of factuality in generated texts [53]. LLMs' struggle to generalize beyond training data, coupled with ethical concerns and integration difficulties, hinders progress in clinical settings [54, 6].

In examining the complexities of trust within human-machine interactions, it is essential to consider the multifaceted nature of trust dynamics. Figure 3 illustrates the hierarchical structure of these dynamics, highlighting various factors that influence trust in artificial intelligence (AI). This figure delineates not only the challenges inherent in trust dynamics but also the evolution of trust over time across different contexts. Each category within the figure is meticulously broken down into subcategories and detailed points, underscoring the intricate interplay among technical elements, user perceptions, and the quality of interactions that contribute to fostering trust in AI systems. This comprehensive visualization serves as a critical reference point for understanding the implications of trust dynamics in the design and development of AI technologies.

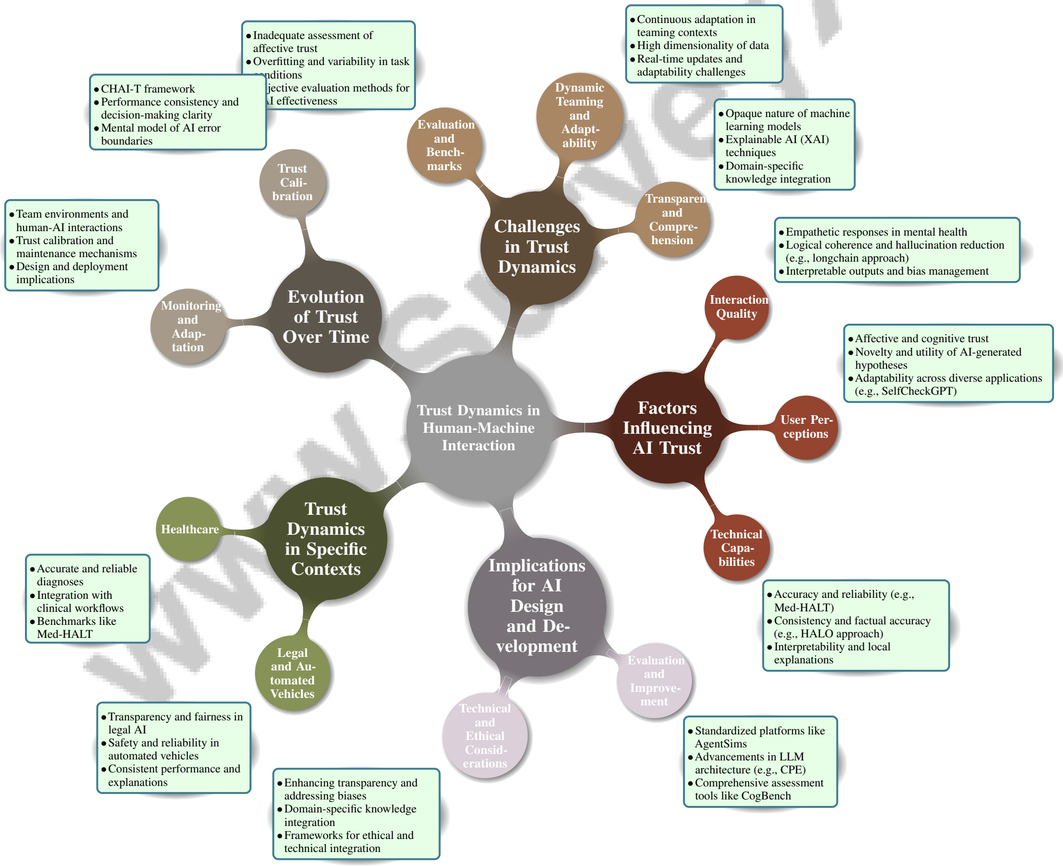

Trust Dynamics in Human-Machine Interaction
Challenges in Trust Dynamics
Evaluation and Benchmarks
• Inadequate assessment of affective trust
• Overfitting and variability in task conditions
• jective evaluation methods for AI effectiveness
Dynamic Teaming and Adaptability
• Continuous adaptation in teaming contexts
• High dimensionality of data
• Real-time updates and adaptability challenges
Transparency and Comprehension
• Opaque nature of machine learning models
• Explainable AI (XAI) techniques
• Domain-specific knowledge integration
Evolution of Trust Over Time
Trust Calibration
• CHAI-T framework
• Performance consistency and decision-making clarity
• Mental model of AI error boundaries
Monitoring and Adaptation
• Team environments and human-AI interactions
• Trust calibration and maintenance mechanisms
• Design and deployment implications
Factors Influencing AI Trust
Interaction Quality
• Empathetic responses in mental health
• Logical coherence and hallucination reduction (e.g., longchain approach)
• Interpretable outputs and bias management
User Perceptions
• Affective and cognitive trust
• Novelty and utility of AI-generated hypotheses
• Adaptability across diverse applications (e.g., SelfCheckGPT)
Technical Capabilities
• Accuracy and reliability (e.g., Med-HALT)
• Consistency and factual accuracy (e.g., HALO approach)
• Interpretability and local explanations
Trust Dynamics in Specific Contexts
Healthcare
• Accurate and reliable diagnoses
• Integration with clinical workflows
• Benchmarks like Med-HALT
Legal and Automated Vehicles
• Transparency and fairness in legal AI
• Safety and reliability in automated vehicles
• Consistent performance and explanations
Implications for AI Design and Development
Technical and Ethical Considerations
• Enhancing transparency and addressing biases
• Domain-specific knowledge integration
• Frameworks for ethical and technical integration
Evaluation and Improvement
• Standardized platforms like AgentSims
• Advancements in LLM architecture (e.g., CPE)
• Comprehensive assessment tools like CogBench

Figure 3: This figure illustrates the hierarchical structure of trust dynamics in human-machine interaction, highlighting factors influencing AI trust, challenges in trust dynamics, the evolution of trust over time, trust dynamics in specific contexts, and implications for AI design and development. Each category is broken down into subcategories and detailed points, emphasizing the complex interplay of technical, user perception, and interaction quality aspects in fostering trust in AI systems.

4 Trust Dynamics in Human-Machine Interaction

4.1 Factors Influencing AI Trust

Trust in AI systems is influenced by technical capabilities, user perceptions, and interaction quality. Accuracy and reliability are crucial, particularly in fields like healthcare, where benchmarks such as Med-HALT assess model precision [1]. Consistency and factual accuracy, as highlighted by the HALO approach, are essential for trust in less robust large language models (LLMs) [33]. Interpretability enhances trust, with LLMs providing confidence scores and local explanations to improve user confidence in AI-assisted decision-making [19]. However, the generative nature of these models can obscure mechanisms, complicating user understanding [23]. Techniques like the longchain approach, which enhance logical coherence and reduce hallucination rates, are vital for applications requiring high reliability [47].

In mental health support, LLMs enhance trust by enabling early detection of psychological issues and delivering empathetic responses, thereby improving therapeutic interactions [9]. The dynamics of affective and cognitive trust in AI are reinforced by benchmarks assessing these interactions, necessitating comprehensive frameworks [19]. The novelty and utility of hypotheses generated by frameworks like LLMCG, compared to traditional methods, also influence user confidence in AI capabilities [8]. Adaptability is crucial, as demonstrated by SelfCheckGPT's ability to operate across various LLMs without external resources, ensuring trust across diverse applications [53].

The effectiveness of methods like CoNLI in producing interpretable outputs and implementing mitigation strategies significantly affects trust [2]. Addressing factors such as accuracy, interpretability, bias management, and understanding user intent is essential for enhancing user confidence and ensuring successful AI technology deployment across multiple domains.

4.2 Challenges in Trust Dynamics

Trust dynamics in AI systems face challenges due to the complexity of technologies and their integration into various domains. The opaque nature of advanced machine learning models undermines transparency, complicating users' comprehension of AI decision-making. This lack of clarity can inhibit trust, particularly in critical areas like healthcare and academic research, where understanding AI outputs is vital. Recent studies emphasize Explainable AI (XAI) techniques to clarify AI behaviors, promoting greater trust [55, 20, 56]. Integrating domain-specific knowledge into AI systems can enhance output interpretability, supporting responsible technology deployment [5, 57].

Managing trust within dynamic teaming contexts requires continuous adaptation [58]. This fluidity necessitates ongoing monitoring and can be resource-intensive. The high dimensionality of data, such as EEG and GSR, complicates the identification of key features influencing trust levels [26]. Existing benchmarks often inadequately assess affective trust, relying on scales adapted from human contexts without sufficient modification for AI interactions [59]. This highlights the need for benchmarks specifically designed to evaluate trust in AI.

Variability in task conditions and potential overfitting in current benchmarks complicate the assessment of AI capabilities and trustworthiness [41]. Despite efforts to create cleaner evaluation frameworks like GSM1k, comprehensive evaluation of hallucinations in high-stakes applications remains insufficient. The absence of objective evaluation methods for XAI effectiveness and biases in user interactions impedes progress in establishing trust [56]. Additionally, the need for real-time updates and adaptability beyond initial training, alongside the accuracy of knowledge provided by LLMs, poses challenges in environments requiring high adaptability [60].

Addressing these challenges requires overcoming AI systems' black-box nature, managing trust in dynamic contexts, improving evaluation frameworks, and enhancing adaptability and transparency. Such efforts are crucial for building trust and facilitating effective AI integration across diverse fields, addressing key concerns related to fairness, transparency, and explainability, essential for user acceptance and successful collaboration between human experts and AI technologies [61, 55, 5, 18].

4.3 Evolution of Trust Over Time

The evolution of trust in AI systems is a complex process influenced by prior experiences, team interactions, and contextual factors. The CHAI-T framework provides a comprehensive understanding

AI-generated, for reference only.

of this dynamic construct, emphasizing the need for continuous trust calibration [58]. Users' trust levels are shaped by the system's performance consistency and the clarity of its decision-making processes. A significant challenge in trust evolution is users' difficulty in forming a clear mental model of the AI's error boundaries, which can result in fluctuating trust levels as users struggle to predict AI behavior in various scenarios [61]. The ability of AI systems to deliver transparent and interpretable outputs is essential for addressing this issue, enabling users to understand the system's limitations and capabilities.

Ongoing monitoring and adaptation are necessary, particularly in team environments where human-AI interactions significantly impact trust levels. Careful management of these interactions is vital for ensuring effective collaboration [58]. As trust evolves, it is crucial to consider implications for AI system design and deployment, ensuring mechanisms are in place to support trust calibration and maintenance.

4.4 Trust Dynamics in Specific Contexts

Trust dynamics in AI systems vary widely across applications and contexts, influenced by the specific requirements and challenges of each domain. In healthcare, trust in AI systems hinges on their ability to provide accurate and reliable diagnoses and their integration with clinical workflows [1]. Benchmarks like Med-HALT emphasize precision, as inaccurate AI outputs can severely impact patient safety and treatment outcomes [1]. In mental health support, AI systems must balance empathetic interactions with clinical accuracy [9]. The use of LLMs in this context illustrates AI's potential to enhance therapeutic relationships by providing timely support and recognizing early signs of psychological distress [9]. Success relies on systems' ability to build and maintain trust, necessitating continuous evaluation and adaptation to meet user expectations and clinical standards.

In legal settings, trust dynamics are shaped by the need for transparency and fairness in AI decision-making processes. Users must trust the AI's capacity to accurately and impartially interpret complex legal data, requiring robust mechanisms for transparency and accountability [28]. Integrating AI into legal workflows demands careful consideration of ethical and procedural standards to uphold trust among legal professionals and stakeholders [28]. In automated vehicles, trust dynamics depend on the AI's ability to accurately perceive and respond to dynamic environments. The capacity to distinguish between semantic categories, such as pedestrians and vehicles, is crucial for ensuring safety and reliability [29]. Trust in these systems builds through consistent performance and clear explanations of AI actions, enhancing user confidence in decision-making capabilities [29].

The dynamics of trust in AI are highly context-dependent, shaped by the specific demands and expectations of each application domain. Addressing the unique challenges and requirements of various contexts can significantly enhance the trustworthiness and acceptance of AI technologies, essential for effective integration across diverse fields. Trust in AI systems is influenced by fairness, transparency, explainability, and adherence to ethical guidelines. Understanding AI-led decision-making, including biases and data limitations, is critical. By incorporating explicit domain knowledge and human-centered explanations, developers can foster greater user confidence, facilitating broader adoption and responsible AI use across multiple applications [20, 55, 18].

4.5 Implications for AI Design and Development

Designing and developing AI systems require navigating the complex dynamics of trust, necessitating a multifaceted approach that integrates technical, ethical, and user-centric considerations. Enhancing transparency and addressing biases in AI models are fundamental to fostering trust and ensuring successful adoption. A comprehensive taxonomy for understanding risks, including hallucinations, is crucial for effectively designing large language models (LLMs) [62]. Incorporating domain-specific knowledge into AI systems can significantly enhance the quality of explanations provided, improving user understanding and trust [55]. This is particularly critical in healthcare, where AI integration requires strategies accommodating diverse datasets and emphasizing the human element in patient care. Research by He et al. underscores the importance of a robust framework integrating ethical considerations, technical capabilities, and user acceptance to ensure AI systems are both worthy and trustworthy [17].

Standardized platforms like AgentSims play a pivotal role in advancing LLM evaluation by facilitating collaboration and innovation within the field [43]. Improvements in LLM architecture, such as those

AI-generated, for reference only.

offered by CPE, demonstrate substantial advancements in maintaining contextual coherence and reducing hallucination rates, enhancing system performance [63]. Frameworks like CogBench, which integrate behavioral insights, significantly contribute to LLM evaluation and highlight the need for comprehensive assessment tools [64]. The interplay between cognitive and affective trust in AI is vital for enhancing human-AI interactions, as evidenced by the benchmark developed by Shang et al., providing a validated tool for measuring these dimensions [59]. Understanding user intents and satisfaction levels is crucial for developing user-centered LLMs, as emphasized by Wang et al. [65]. Furthermore, the potential of LLMs to contribute positively to security while posing significant risks necessitates further research into their vulnerabilities and defenses [10].

Additionally, methods like PURR, which offer efficient attribution improvements, highlight the importance of balancing computational costs with performance enhancements in AI design [31]. The dual dialogue system proposed by Kampman et al. exemplifies how AI systems can maintain the quality of human services, such as therapy, while managing workload, benefiting both providers and clients [3].

5 AI Perception and User Expectations

In examining AI perception and user expectations, it is crucial to understand the complex nature of user interactions with AI systems. The subsection "User Expectations and Explanation Satisfaction" focuses on users' specific anticipations regarding AI functionalities, particularly concerning operational reliability and the clarity of explanations. Recognizing these expectations is vital for fostering user trust and engagement, laying the groundwork for discussions on the implications of mismatched expectations that may arise later. We thus explore the nuanced dynamics of user expectations and the satisfaction derived from AI system explanations.

5.1 User Expectations and Explanation Satisfaction

User expectations from AI systems are increasingly focused on secure, efficient, and reliable operations, especially in contexts involving large language models (LLMs). A crucial aspect of these expectations is the demand for models capable of generating innovative and high-quality hypotheses, as highlighted by the LLMCG framework [8]. Balancing bias reduction with language understanding is essential for influencing user perception and trust, demonstrated by the SRLLM's enhancements in these areas [66]. Providing confidence scores can significantly improve trust calibration, enabling users to make informed decisions and align expectations with system capabilities [61]. This is particularly relevant in fields like psychological assessments, where LLMs show promise in sentiment analysis and risk detection, despite lagging behind expert systems in nuanced emotional awareness [6].

Explanation satisfaction is critical in fostering user engagement and trust in AI systems. Framing explanations as questions enhances user engagement and critical thinking, improving interaction experiences [19]. This aligns with the need for rigorous human-centric evaluation methods, especially in assessing complex outputs like audio adversarial examples, where current metrics may underestimate perceptual distortion [30]. Additionally, the performance of LLMs in cognitive tasks offers insights into human cognition, meeting user expectations for AI systems mimicking human-like reasoning processes [54]. However, challenges in categorizing risks and ensuring AI safety underscore the need for robust frameworks addressing these issues [23].

5.2 Implications of Mismatched Expectations

Mismatched expectations between users and AI systems can lead to challenges affecting user engagement and satisfaction. A primary issue arises when users prefer human-generated content over AI-generated messages, particularly when the content source is disclosed, highlighting the importance of managing user expectations to foster trust in AI technologies [49]. This disconnect is evident in voice-based systems, where users may experience low engagement and frustration due to misaligned expectations [67].

In legal practice, the persistence of hallucinations in RAG-based legal AI tools, despite reduced hallucinations compared to general-purpose models, illustrates the need for cautious deployment and clear communication about AI limitations. Users must be informed of potential inaccuracies to

AI-generated, for reference only.

manage expectations and ensure informed decision-making [68]. Effective trust calibration requires understanding user experiences and cognitive processes, ensuring users can accurately assess AI reliability and adjust trust levels accordingly [61]. Unanswered questions about integrating user intents with LLM functionalities and adapting models to diverse user needs are crucial for developing responsive AI systems capable of delivering personalized, contextually relevant outputs [65].

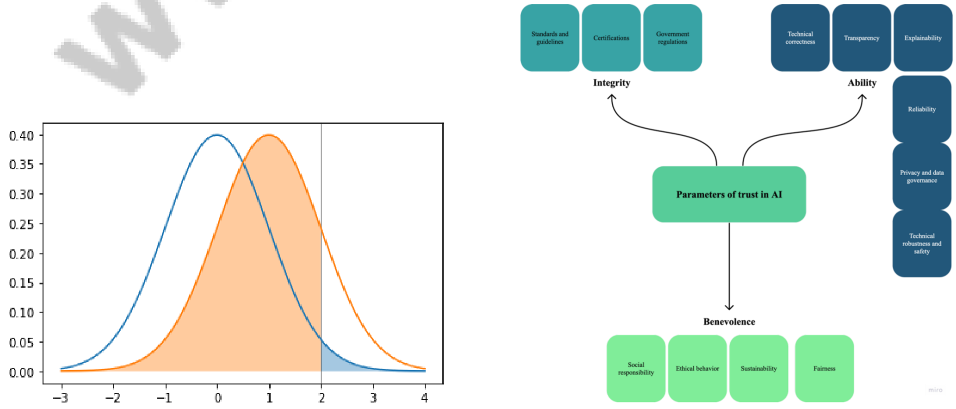
5.3 Perception of AI in High-Stakes Applications

The perception of AI in high-stakes applications is shaped by balancing potential benefits with ethical, reliability, and transparency concerns. In healthcare, legal systems, and autonomous vehicles, AI systems are expected to perform accurately and reliably, given the significant consequences of errors. However, reliance on black-box models exacerbates ethical concerns, as transparency and accountability in AI decision-making are demanded [57].

In healthcare, cautious integration of AI systems is due to the critical nature of medical decisions and their impact on patient outcomes. Trust in AI-driven diagnostic tools depends on accurate, interpretable results aligning with clinical standards. Ethical implications and data privacy concerns complicate AI decision-making, necessitating comprehensive frameworks prioritizing transparency and accountability. Mechanisms for calibrating human trust in AI systems are crucial, allowing experts to integrate domain knowledge with AI insights for optimized decisions [61, 57].

In legal settings, AI systems must uphold justice and fairness, requiring thorough examination of ethical standards and procedural fairness. While AI tools enhance legal task efficiency, they are prone to inaccuracies or "hallucinations," undermining credibility. Ongoing human oversight ensures legal professionals are aware of automated system pitfalls, safeguarding decision-making autonomy and public confidence [19, 68, 18, 61, 57]. The perception of AI in this context is influenced by the system's capacity to provide clear, justifiable decisions, underscoring the need for transparent, interpretable models.

In autonomous vehicles, AI perception is tied to safe navigation and real-time decision-making. High stakes necessitate trust in AI's decision-making, challenging given the black-box nature of many models. Ensuring AI reliability and safety in high-stakes environments is essential for public trust, requiring fairness, privacy, transparency, and explainability, alongside human-centered explanations aligning with domain knowledge [61, 55]. The perception of AI in high-stakes applications is influenced by technology capabilities and concerns like ethics, transparency, and reliability. In AI-assisted decision-making, trust calibration is crucial, shaped by confidence scores and clear explanations, helping users gauge AI predictions. Establishing trust requires understanding AI behavior, aligning with ethical standards and societal values for confidence in use [55, 18, 20, 61, 57]. Addressing these challenges is essential for fostering trust and integrating AI systems in critical domains.



(a) ERROR!!![13]

(b) ERROR!!![18]

Figure 4: Examples of Perception of AI in High-Stakes Applications

As shown in Figure 4, the perception of AI systems, especially in high-stakes applications, plays a critical role in shaping user expectations and trust. The example illustrated in Figure 4 highlights the complexities involved in understanding how AI is perceived in scenarios where the stakes are high, such as healthcare, finance, or autonomous driving. The images, though marked with "ERROR!!!", suggest a need for careful consideration of user interaction with AI systems. These visual representations underscore the challenges and progress in AI perception as discussed in the works of Knowles et al. (2022) and Afroogh et al. (2024). By examining these examples, we gain insights into the potential gaps between AI capabilities and user expectations, emphasizing the importance of transparency, reliability, and ethical considerations in AI deployment. [13, 18]

5.4 Role of Cognitive and Affective Trust

Cognitive and affective trust are crucial in shaping user perception and acceptance of AI systems, influencing user interaction and reliance on these technologies. Cognitive trust involves rational assessment of an AI system's capabilities and reliability, while affective trust involves emotional responses elicited by AI interactions. Integrating physiological data with user feedback provides insights into users' emotional states, essential for understanding affective trust in autonomous vehicle interactions [69].

Linguistic and computational requirements for creating avatars that engage in human-like conversations emphasize verbal and nonverbal cues' importance in fostering cognitive and affective trust [15]. These cues significantly impact user perception, contributing to perceived intelligence and empathy of AI systems. LLMs functioning as interlocutor automata in practical applications impact user trust, navigating complex conversational dynamics [70].

The correlation between student models' performance and teacher-student self-explanations highlights clear, coherent communication's importance in building cognitive trust [71]. Models like GPT-4 interpreting metaphors suggest LLMs can achieve human-like cognitive abilities, enhancing cognitive trust through improved interpretative skills [72]. Understanding reasoning tasks' context and content is essential for improving human and machine reasoning, enhancing cognitive trust [73]. Ethical considerations foster public trust and acceptance of robotic and autonomous systems (RAS), ensuring alignment with societal values and expectations [17].

Metrics measuring cognitive and affective trust levels provide insights into user perception of AI agents, enabling developers to refine systems to meet user expectations and foster trust [59]. Addressing cognitive and affective trust dimensions enhances user experience, ensuring AI systems are perceived as reliable, empathetic, and trustworthy across applications.

5.5 Impact of AI System Design on User Perception

AI system design plays a pivotal role in shaping user perception and trust, integrating innovative technologies and human-centric features. Decentralized blockchain technology in Dynamic Large Language Models (DLLMs) exemplifies efforts to enhance data usage transparency, fostering user trust and confidence in AI outputs [74]. This approach addresses data privacy and security concerns, critical factors influencing user perception and acceptance.

Incorporating human-like memory processes into AI systems significantly impacts user perception by enabling coherent, context-aware interactions. This design choice enhances system recall of relevant information and conversational continuity, improving user satisfaction and trust in AI capabilities [75]. AI systems simulating human cognitive processes contribute to natural, intuitive interaction experiences, aligning with user expectations for intelligent, empathetic AI agents.

Developing benchmarks, like Shang et al.'s, provides insights into measuring cognitive and affective trust in AI systems. Employing a 27-item semantic differential scale, this benchmark offers a comprehensive evaluation of trust dynamics, differing from previous approaches focused on cognitive aspects [59]. Holistic trust assessment is crucial for understanding AI system design's influence on user perception and emotional responses.

Advancements in detecting critical failure modes in LLMs enhance reliability in practical applications. Identifying and addressing failure modes improves performance and reduces errors, positively impacting user trust and perception [37]. Continuous AI design refinement underscores robust, reliable systems' importance in fostering user confidence.

Overall, AI system design, through transparency-enhancing technologies, human-like memory processes, and comprehensive trust evaluation frameworks, shapes user perception and trust. Addressing interpretability, human-centered explanations, and robust evaluation mechanisms, developers create AI systems fostering trust, transparency, reliability, and empathy. This approach enhances user engagement and acceptance across applications, leading to responsible AI integration in fields like healthcare, science, and everyday decision-making [20, 55, 57].

6 Conclusion

This survey highlights the imperative of addressing AI hallucinations and trust dynamics to enhance human-machine interactions. Notable progress has been achieved with tools like SelfCheckGPT, which adeptly identifies hallucinations in LLM outputs, offering a robust approach for zero-resource hallucination detection in opaque systems. Similarly, the CoNLI framework demonstrates effective methodologies for improving hallucination detection and reducing ungrounded outputs while maintaining textual integrity. Nevertheless, the persistence of hallucinations remains a formidable challenge, as evidenced by the LLMCG framework, which underscores the necessity of generating hypotheses with a level of novelty comparable to human experts.

Future research should prioritize the development of sophisticated evaluation frameworks and the identification of critical intervention points to enhance the safety of LLMs. Recent validations of evaluation frameworks for mental health chatbots underscore their effectiveness in bolstering safety and reliability. Moreover, emphasizing the scrutiny of machine-generated recommendations can reduce over-reliance on automated systems and enhance human decision-making capabilities.

The HALO CHECK method provides a quantitative approach to mitigating hallucinations in LLMs, thereby improving model outputs and informing future research directions. Additionally, the proposed dual dialogue system for mental health care effectively supports providers by generating responses that mirror the empathy of professional therapists, thus enhancing therapeutic processes and reducing cognitive load.

While Med-HALT offers a comprehensive framework for evaluating hallucinations in LLMs, there remains significant scope for improving model accuracy and reliability. The insights from McIntosh et al. advocate for enhanced benchmarking practices in LLM evaluation, calling for a unified framework that bolsters both functionality and integrity in assessments.

AI-generated, for reference only.

References

[1] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

[2] Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023.

[3] Onno P. Kampman, Ye Sheng Phang, Stanley Han, Michael Xing, Xinyi Hong, Hazirah Hoosainsah, Caleb Tan, Genta Indra Winata, Skyler Wang, Creighton Heaukulani, Janice Huiqin Weng, and Robert JT Morris. An ai-assisted multi-agent dual dialogue system to support mental health care providers, 2024.

[4] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024.

[5] 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.

[6] Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, and Bing Xiang Yang. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects, 2023.

[7] Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. A survey on large language model hallucination via a creativity perspective. *arXiv preprint arXiv:2402.06647*, 2024.

[8] Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. Automating psychological hypothesis generation with ai: when large language models meet causal graph, 2024.

[9] Jung In Park, Mahyar Abbasian, Iman Azimi, Dawn Bounds, Angela Jun, Jaesu Han, Robert McCarron, Jessica Borelli, Jia Li, Mona Mahmoudi, Carmen Wiedenhoeft, and Amir Rahmani. Building trust in mental health chatbots: Safety metrics and llm-based evaluation tools, 2024.

[10] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly, 2024.

[11] Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. Ai hallucinations: A misnomer worth clarifying, 2024.

[12] Rami Hatem, Brianna Simmons, and Joseph E Thornton. A call to address ai “hallucinations” and how healthcare professionals can mitigate their risks. *Cureus*, 15(9), 2023.

[13] Bran Knowles, Jason D’Cruz, John T. Richards, and Kush R. Varshney. Humble machines: Attending to the underappreciated costs of misplaced distrust, 2022.

[14] Xin Sun, Rongjun Ma, Xiaochang Zhao, Zhuying Li, Janne Lindqvist, Abdallah El Ali, and Jos A. Bosch. Trusting the search: Unraveling human trust in health information from google and chatgpt, 2024.

[15] João Ranhel and Cacilda Vilela de Lima. On the linguistic and computational requirements for creating face-to-face multimodal human-machine interaction, 2022.

[16] Ruyuan Wan, Naome Etori, Karla Badillo-Urquiola, and Dongyeop Kang. User or labor: An interaction framework for human-machine relationships in nlp, 2022.

[17] Hongmei He, John Gray, Angelo Cangelosi, Qinggang Meng, T. Martin McGinnity, and Jörn Mehnen. The challenges and opportunities of human-centered ai for trustworthy robots and autonomous systems, 2021.

AI-generated, for reference only.

[18] Saleh Afroogh, Ali Akbari, Evan Malone, Mohammadali Kargar, and Hananeh Alambeigi. Trust in ai: Progress, challenges, and future directions, 2024.

[19] Simon WS Fischer. Questioning ai: Promoting decision-making autonomy through reflection, 2024.

[20] Sai Anirudh Athaluri, Sandeep Varma Manthena, VSR Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15(4), 2023.

[21] Ari Holtzman, Peter West, and Luke Zettlemoyer. Generative models as a complex systems science: How can we make sense of large language model behavior?, 2023.

[22] 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.

[23] Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. Ai safety in generative ai large language models: A survey, 2024.

[24] Marco AF Pimentel, Clément Christophe, Tathagata Raha, Prateek Munjal, Praveen K Kanithi, and Shadab Khan. Beyond metrics: A critical analysis of the variability in large language model evaluation frameworks, 2024.

[25] Qun Ma, Xiao Xue, Deyu Zhou, Xiangning Yu, Donghua Liu, Xuwen Zhang, Zihan Zhao, Yifan Shen, Peilin Ji, Juanjuan Li, Gang Wang, and Wanpeng Ma. Computational experiments meet large language model based agents: A survey and perspective, 2024.

[26] Farhana Faruqe, Ryan Watkins, and Larry Medsker. Monitoring trust in human-machine interactions for public sector applications, 2020.

[27] Yihe Fan, Yuxin Cao, Ziyu Zhao, Ziyao Liu, and Shaofeng Li. Unbridled icarus: A survey of the potential perils of image inputs in multimodal large language model security, 2024.

[28] Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. Ai hallucinations: a misnomer worth clarifying. In *2024 IEEE conference on artificial intelligence (CAI)*, pages 133–138. IEEE, 2024.

[29] Robin Chan, Radin Dardashti, Meike Osinski, Matthias Rottmann, Dominik Brüggemann, Cilia Rücker, Peter Schlicht, Fabian Hüger, Nikol Rummel, and Hanno Gottschalk. What should ai see? using the public's opinion to determine the perception of an ai, 2022.

[30] Jon Vadillo and Roberto Santana. On the human evaluation of audio adversarial examples, 2021.

[31] Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*, 2023.

[32] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.

[33] Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*, 2023.

[34] Stefanie Krause and Frieder Stolzenburg. From data to commonsense reasoning: the use of large language models for explainable ai. *arXiv preprint arXiv:2407.03778*, 2024.

[35] Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*, 2023.

[36] Pingping Lu, Liang Huang, Tan Wen, and Tianyu Shi. Assessing visual hallucinations in vision-enabled large language models. 2024.

[37] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[38] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418, 2024.

[39] Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21853–21862, 2024.

[40] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models, 2024.

[41] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, 2024.

[42] Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. *arXiv preprint arXiv:2212.09561*, 2022.

[43] Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiuyue Ping, and Qin Chen. Agentsims: An open-source sandbox for large language model evaluation, 2023.

[44] SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

[45] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*, 2024.

[46] Hugo Underwood and Zoe Fenwick. Implementing an automated socratic method to reduce hallucinations in large language models. 2024.

[47] Jinchao Li and Quan Hong. A longchain approach to reduce hallucinations in large language models. 2024.

[48] Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization, 2024.

[49] Sue Lim and Ralf Schmälzle. The effect of source disclosure on evaluation of ai-generated messages: A two-part study, 2023.

[50] Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

[51] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning, 2021.

[52] Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, William Song, Tiffany Zhao, Pranav Raja, Charlotte Zhuang, Dylan Slack, et al. A careful examination of large language model performance on grade school arithmetic. *Advances in Neural Information Processing Systems*, 37:46819–46836, 2024.

[53] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[54] Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, et al. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint arXiv:2409.02387*, 2024.

[55] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. Knowledge-intensive language understanding for explainable ai, 2021.

[56] Giovanni Cinà, Tabea Röber, Rob Goedhart, and Ilker Birbil. Why we do need explainable ai for healthcare, 2022.

[57] Max W. Shen. Trust in ai: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient, 2022.

[58] Melanie J. McGrath, Andreas Duenser, Justine Lacey, and Cecile Paris. Collaborative human-ai trust (chai-t): A process framework for active management of trust in human-ai collaboration, 2024.

[59] Ruoxi Shang, Gary Hsieh, and Chirag Shah. Trusting your ai agent emotionally and cognitively: Development and validation of a semantic differential scale for ai trust, 2024.

[60] Feiyu Zhu and Reid Simmons. Bootstrapping cognitive agents with a large language model, 2024.

[61] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020.

[62] Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems, 2024.

[63] Sarah Desrochers, James Wilson, and Matthew Beauchesne. Reducing hallucinations in large language models through contextual position encoding, 2024.

[64] Julian Coda-Forno, Marcel Binz, Jane X. Wang, and Eric Schulz. Cogbench: a large language model walks into a psychology lab, 2024.

[65] Jiayin Wang, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. Understanding user experience in large language model interactions, 2024.

[66] Shaina Raza, Oluwanifemi Bamgbose, Shardul Ghuge, Fatemeh Tavakol, Deepak John Reji, and Syed Raza Bashir. Developing safe and responsible large language model : Can we balance bias reduction and language understanding in large language models?, 2025.

[67] Roger K Moore. Is spoken language all-or-nothing? implications for future speech-based human-machine interaction. *Dialogues with social robots: enablements, analyses, and evaluation*, pages 281–291, 2017.

[68] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*, 2024.

[69] Lia Morra, Fabrizio Lamberti, F. Gabriele Pratticó, Salvatore La Rosa, and Paolo Montuschi. Building trust in autonomous vehicles: Role of virtual reality driving simulators in hmi design, 2020.

[70] Xabier E. Barandiaran and Lola S. Almendros. Transforming agency. on the mode of existence of large language models, 2024.

[71] Jiachen Zhao, Zonghai Yao, Zhichao Yang, and Hong Yu. Large language models are in-context teachers for knowledge reasoning, 2024.

[72] Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. Large language model displays emergent ability to interpret novel literary metaphors, 2024.

[73] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks. *arXiv preprint arXiv:2207.07051*, 2022.

[74] Yuanhao Gong. Dynamic large language models on blockchains, 2023.

[75] Yuki Hou, Haruki Tamoto, and Homei Miyashita. "my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents, 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.

AI-generated, for reference only.