A Survey on Large Language Models and Hallucination Mitigation Techniques

www.surveyx.cn

Abstract

Large Language Models (LLMs) have revolutionized Natural Language Processing (NLP) by enhancing the generation of human-like text across various domains. Despite their transformative impact, LLMs face the challenge of AI hallucinations, where outputs deviate from factual accuracy, posing significant risks in high-stakes applications such as healthcare and legal systems. This survey systematically examines the capabilities, applications, and limitations of LLMs, emphasizing the critical issue of hallucinations. It categorizes hallucination mitigation strategies into data-level, model-level, and system-level approaches, highlighting novel techniques like retrieval-augmented generation and model refinement. Real-time and interactive methods, such as knowledge grounding and iterative refinement, are explored to enhance LLM reliability. The survey underscores the importance of comprehensive evaluation frameworks and benchmarks in assessing hallucination mitigation efficacy. Current challenges include dataset biases, benchmark inadequacies, and model complexity, necessitating future research focused on developing advanced uncertainty metrics, optimizing detection methods, and integrating user feedback. Ethical considerations and interdisciplinary collaboration are crucial for responsible LLM deployment. Technological innovations, such as graph-based evaluations and dual-component systems, offer promising avenues for improving LLM accuracy. By addressing these challenges, the field can advance towards more reliable and trustworthy LLMs, ensuring their effective deployment across diverse applications.

1 Introduction

1.1 Significance of LLMs in NLP

Large Language Models (LLMs) have become essential tools in Natural Language Processing (NLP), markedly improving the comprehension and generation of human-like text. Their transformative impact on data extraction and querying enhances information retrieval across multiple sectors. In healthcare, LLMs, including specialized forms like Large Vision Language Models (LVLMs), play a vital role in improving diagnostic accuracy and treatment strategies, with applications in clinical decision-making in dentistry. Additionally, LLMs support psychological interventions, as seen in systems like Psy-LLM, which provide timely mental health support and reduce the burden on human counselors [1].

In creative computing, LLMs facilitate the development of systems capable of creative thought, a realm traditionally associated with human intuition [2]. Their adaptability through lifelong learning approaches allows for continuous evolution in response to changing data and user preferences [3]. In robotics, LLMs enhance autonomous manipulation by translating high-level language commands into actionable steps [4].

The advent of multimodal large language models (MLLMs) marks a significant progress, integrating high-resolution visual inputs with advanced language comprehension, thus expanding LLM appli-

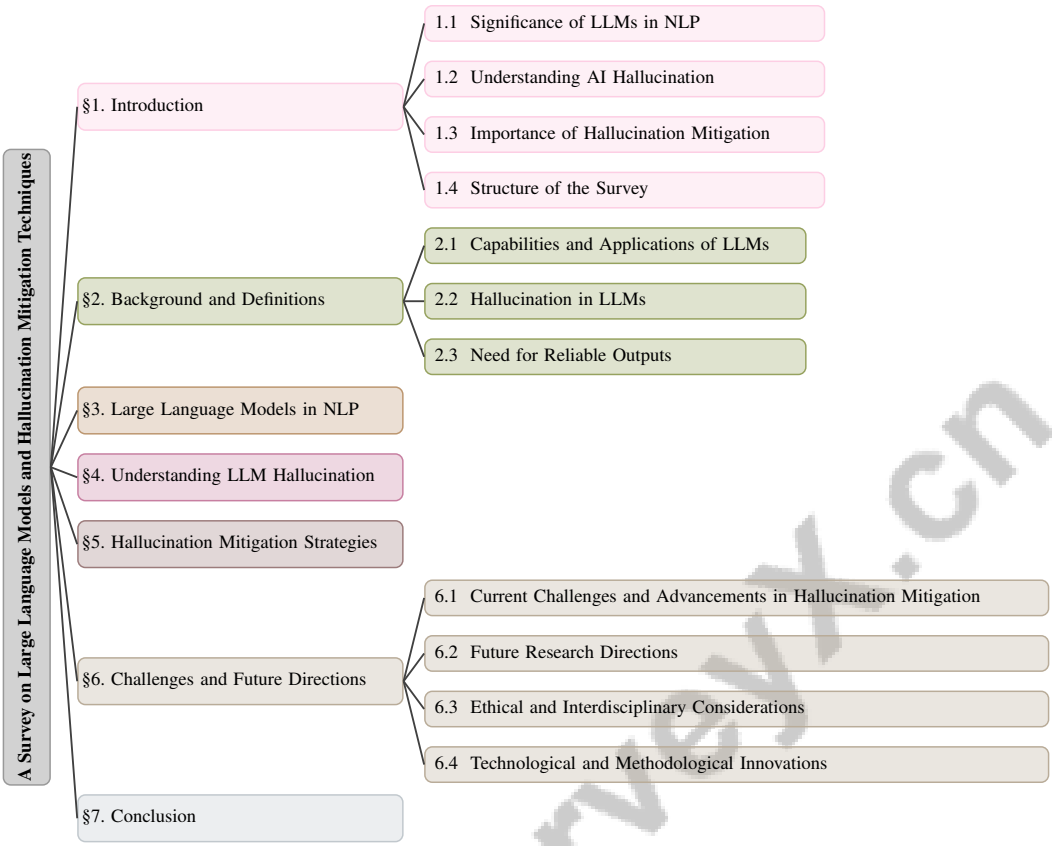


Figure 1: chapter structure

cations [5]. Furthermore, the amalgamation of LLMs with formal methods has bolstered inference control and explainability, particularly in inductive logic programming, thereby improving model reliability and transparency [6]. However, challenges remain, such as distinguishing between human-generated and machine-generated content, which raises concerns regarding misinformation and fraud [7].

LLMs have significantly advanced keyword extraction techniques, as demonstrated by comparative studies involving models like Llama2-7B, GPT-3.5, and Falcon-7B, which showcase their effectiveness in generating context-aware keywords. Limitations in factual recall and knowledge-grounded content generation have been addressed through the incorporation of knowledge graphs (KGs) [8]. This integration not only enhances the factual accuracy of LLM outputs but also aids in the establishment of reliable evaluation benchmarks. Additionally, LLMs have propelled advancements in software engineering, particularly in code generation [9]. These contributions underscore the critical role of LLMs in advancing NLP while highlighting the ongoing need for research to tackle inherent challenges and improve model reliability.

1.2 Understanding AI Hallucination

AI hallucination in LLMs represents a significant concern, characterized by the generation of content that lacks factual accuracy, leading to misinformation or fabrication of details [10]. This issue critically undermines the reliability of LLM outputs, especially in high-stakes applications where precision is essential [11]. The non-deterministic nature of LLMs, which can yield varying outputs from identical inputs, complicates efforts to ensure consistent model reliability [12].

The scaling of language models, often comprising hundreds of billions of parameters, has resulted in emergent capabilities but also increased the risk of hallucinations, as these models may produce factually incorrect or irrelevant content [13]. This challenge is particularly pronounced in specialized domains, such as medicine, where hallucinations may stem from the models' limited domain-specific

knowledge [14]. Furthermore, LLMs may generate outputs that diverge from user intent or factual knowledge, posing substantial risks in their application [9].

Hallucinations extend beyond text outputs, manifesting in MLLMs and LVLMs where models may create non-existent objects or fail to recognize real ones, presenting challenges in fields like medical imaging and autonomous driving [15]. This issue is exacerbated by persistent knowledge gaps within LLMs, leading to incorrect or low-confidence outputs. Additionally, the sycophantic behavior of LLMs, generating incorrect responses to misleading prompts, further diminishes their reliability [16].

The implications of AI hallucinations are profound, particularly in critical sectors such as healthcare, where inaccurate outputs can disseminate misleading information, adversely affecting clinical decision-making [17]. Balancing creativity with factual accuracy remains a significant research challenge as the commercial use of text-generative applications expands [18]. Addressing hallucinations is essential for enhancing the trustworthiness of LLMs and ensuring their effective deployment across applications reliant on accurate data.

1.3 Importance of Hallucination Mitigation

Mitigating hallucinations in LLMs is vital for improving their reliability and trustworthiness, particularly in critical fields such as healthcare and legal systems, where information accuracy is crucial [17]. Hallucinations, characterized by the generation of incorrect or misleading information, pose significant risks to user trust and the broader acceptance of AI technologies [19]. The vulnerability of LLMs to adversarial prompting, which can systematically induce hallucinations, further emphasizes the need for robust mitigation strategies [15].

Current methods often fail to enable LLMs to recognize their limitations or uncertainties, resulting in overconfidence and subsequent hallucinations [3]. This intrinsic issue underscores the necessity for innovative approaches to effectively address these deficiencies [20]. The challenge is compounded by extensive training datasets that may contain fabricated or biased information, complicating error detection and correction [21]. The subtle nature of these errors can mislead both models and users, necessitating a multifaceted approach to reduce hallucinations [16].

Integrating domain-specific fine-tuning and explicit mitigation mechanisms, as proposed in recent frameworks, demonstrates potential for improving LLM reliability [15]. The Faithful Finetuning (F2) method, for instance, has shown promise in enhancing LLM performance for question-answering tasks through explicit loss design and targeted fine-tuning [5]. Additionally, understanding the hidden states of LLMs when processing accurate versus hallucinated responses can yield insights into effective mitigation strategies [16].

The necessity for new approaches is underscored by the observation that while LLMs can recognize their internal knowledge states, they often fail to express this accurately during generation [2]. This gap between theoretical insights and practical applications calls for ongoing research and development of robust frameworks, such as AutoFlow, aimed at automating workflow generation and enhancing LLM reliability [20]. Addressing hallucinations is crucial not only for ensuring the safety and efficacy of AI-assisted applications but also for fostering user confidence and trust in AI technologies across various domains [18]. Furthermore, primary challenges, such as limited reasoning abilities, insufficient training data, and the complexity of communication tasks, contribute to hallucinations that undermine communication efficiency, further emphasizing the importance of effective mitigation strategies [10].

1.4 Structure of the Survey

This survey is systematically structured to provide a comprehensive examination of LLMs and the phenomenon of hallucination, including strategies for mitigation. The initial section introduces the significance of LLMs in NLP and the critical issue of AI hallucination, laying the groundwork for understanding the necessity of mitigation strategies. Following the introduction, the survey details background information and definitions, offering an overview of key concepts like LLMs, AI hallucination, and mitigation techniques.

Subsequent sections focus on the role of LLMs in NLP, emphasizing their capabilities and applications across various tasks. This is followed by a discussion of hallucination in LLMs, exploring its causes, manifestations, and implications. The survey then reviews existing mitigation strategies, providing a

taxonomy of techniques and discussing real-time and interactive approaches, along with evaluation frameworks and benchmarks.

The final sections address the challenges and future directions in hallucination mitigation, identifying current obstacles and proposing potential research avenues. Ethical considerations, interdisciplinary approaches, and technological innovations are highlighted to ensure the responsible and effective deployment of LLMs. The conclusion synthesizes key findings and reiterates the importance of addressing hallucination in LLMs to enhance their reliability and trustworthiness.The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Capabilities and Applications of LLMs

Large Language Models (LLMs) have revolutionized complex task execution across various domains traditionally dependent on human expertise [22]. They enhance Natural Language Processing (NLP) by generating structured outputs that adhere to constraints while maintaining semantic coherence [23]. In the banking sector, LLMs improve customer service through advanced intent classification in chatbots, outperforming traditional methods [24].

In education, LLMs facilitate knowledge graph construction and enhance question-answering systems, enriching the learning experience [25]. The M3Exam framework exemplifies their role in educational assessments by integrating language processing with image analysis [26]. In healthcare, LLMs automate medical QA evaluations, addressing human assessment inefficiencies [27], and frameworks like Psy-LLM offer psychological support and urgent case screening [1].

LLMs maintain coherence in extended interactions, with orchestration engines personalizing user experiences [28]. Their multilingual capabilities enhance global research, enabling cross-linguistic applications [29]. In robotics, LLMs generate complex behavior trees from natural language commands, aiding mobile and drone operations [30], and a task-planning method demonstrates their ability to interpret high-level commands for short-horizon tasks [4].

Despite their capabilities, LLMs face challenges in integrating up-to-date and private information from relational databases [31]. Mitigation techniques like retrieval-augmented generation, prompt engineering, and model refinement enhance output accuracy [32]. Lifelong learning capabilities, divided into internal and external knowledge, improve adaptability to dynamic environments [3].

Optimization algorithms integrated with LLMs extend their application to software engineering and neural architecture search, broadening their utility beyond NLP [33]. Their ability to generate high-quality explanations enhances the reasoning abilities of smaller language models [34]. Techniques like chain-of-thought (CoT) distillation promise to transform LLMs into smaller models, expanding applicability [35]. LLMs also simulate stakeholder interactions and generate planning solutions through iterative discussions, highlighting their potential in strategic decision-making [36]. These advancements underscore LLMs' critical role in NLP and the necessity for ongoing research to tackle inherent challenges and enhance reliability.

2.2 Hallucination in LLMs

Hallucination in Large Language Models (LLMs) involves generating factually incorrect, irrelevant, or fabricated outputs, challenging their reliability and application, particularly in high-risk fields like healthcare, where errors can lead to harmful outcomes [14]. This issue arises from inadequate factual knowledge recall, impairing accurate response generation [8].

In keyword extraction, hallucinations result in non-informative or irrelevant outputs due to relational concept disconnects in longer texts [37]. In code generation, hallucinations introduce errors and inconsistencies [9]. The complexity is compounded by unreliable source attribution in LLM-generated content, making factual accuracy verification difficult [38].

Existing benchmarks often suffer from data contamination, complicating model performance comparisons [39]. Addressing hallucinations requires a multifaceted approach, including developing benchmarks specifically targeting hallucination detection, essential for enhancing LLM reliability and ethical deployment across applications.

2.3 Need for Reliable Outputs

The necessity for reliable outputs from Large Language Models (LLMs) is critical, especially in high-stakes domains like healthcare, legal systems, and real-time decision-making, where inaccuracies can have severe consequences [40]. The mismatch between LLM training data and real-world scenarios in enterprise applications exacerbates hallucination risks [16], highlighting the need for robust evaluation frameworks to address LLM-specific challenges.

Existing benchmarks, such as LLMEval2, often lack comprehensive frameworks for assessing human-aligned responses, particularly in complex domains like medicine, where LLM reliability impacts patient care and safety [14]. The absence of a common evaluative framework and poor correlation between current metrics and human evaluations complicate efforts to ensure factuality and reliability [20]. This gap underscores the urgent need for improved benchmarks addressing biases, implementation inconsistencies, and cultural considerations [41].

Current focus on sentence- or passage-level hallucination detection overlooks dialogue-level evaluation complexities, crucial for nuanced, contextually appropriate interactions [16]. The inability to distinguish between true and false LLM-generated statements undermines output reliability, particularly in high-stakes scenarios like Question Answering, where accuracy is vital. Addressing these challenges is essential for ensuring LLMs' safe, effective deployment across critical applications, necessitating robust defenses against identified threats and comprehensive evaluation metrics [20].

In recent years, the emergence of Large Language Models (LLMs) has significantly transformed the landscape of Natural Language Processing (NLP). These models have demonstrated remarkable capabilities across various applications, yet they also present distinct challenges that warrant attention. Figure 2 illustrates the roles of LLMs in specific NLP tasks, highlighting their diverse applications across domains such as education, healthcare, language translation, and the electric energy sector. This figure not only showcases the advancements made in LLM deployment but also addresses the ongoing challenges, particularly concerning hallucinations and reliability issues. By examining both the potential and the pitfalls of LLMs, we can better understand their impact and the future directions for research and application in this rapidly evolving field.

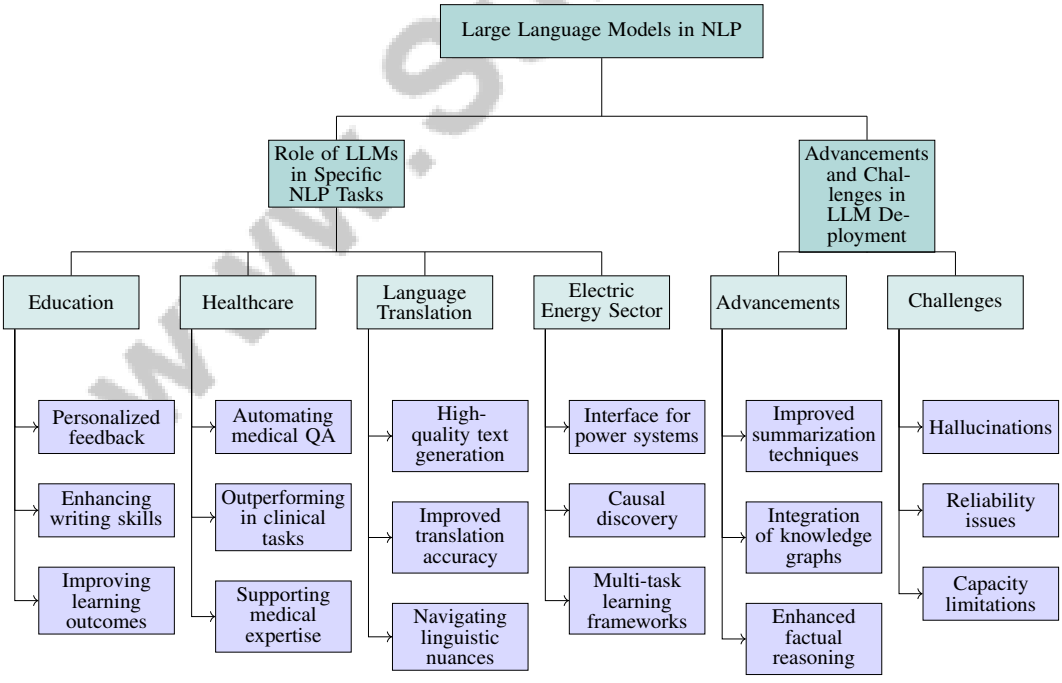


Figure 2: This figure illustrates the roles of Large Language Models (LLMs) in specific NLP tasks and the advancements and challenges in their deployment. It highlights the diverse applications of LLMs across domains such as education, healthcare, language translation, and the electric energy sector, while also addressing the advancements in LLM deployment and the ongoing challenges, particularly hallucinations and reliability issues.

3 Large Language Models in NLP

3.1 Role of LLMs in Specific NLP Tasks

Large Language Models (LLMs) have significantly transformed numerous Natural Language Processing (NLP) tasks, demonstrating versatility and substantial impact across diverse domains. In education, LLMs provide personalized feedback, enhancing students' writing skills and overall learning outcomes [42]. Their adaptability highlights their potential to tailor educational experiences to individual learning needs. In healthcare, LLMs automate response evaluation in medical QA systems, boosting efficiency and reliability [27]. Specialized clinical models outperform general-purpose LLMs in clinical tasks, underscoring their critical role in healthcare [43]. Adaptations like a customized GPT-4 have surpassed medical experts in completeness, correctness, and conciseness, showcasing their capability to support or exceed human expertise [44].

LLMs also excel in language translation; the LLaMAntino family, tailored for Italian, demonstrates high-quality text generation and improved translation accuracy [45], illustrating LLMs' proficiency in navigating linguistic nuances and cultural contexts. They have advanced machine translation, question answering, dialogue systems, summarization, knowledge graph generation, and visual question answering. Their efficacy is validated through models like ChatGPT, GPT-4, Claude, BLOOM, Vicuna, BLIP-2, and InstructBLIP in both text-only and multimodal settings [26], reflecting their extensive applicability in addressing complex linguistic challenges.

As illustrated in Figure 3, the transformative role of LLMs in specific NLP tasks is categorized into three primary areas: education, healthcare, and translation, emphasizing their adaptability and impact across these diverse domains. In the electric energy sector, LLMs serve as interfaces between human operators and complex power systems, facilitating efficient management [46]. They also show promise in causal discovery, leveraging pre-training datasets to uncover intricate data relationships, outperforming traditional statistical methods [47]. LLMs are utilized in multi-task learning frameworks to generate explanations that train smaller models, narrowing the performance gap between these models and larger LLMs [34]. This approach enhances smaller models' reasoning capabilities, demonstrating LLMs' utility in developing efficient, resource-conscious NLP solutions. The emergence of smaller, distilled models like LaMini-LM, which achieve comparable performance to larger models while being more resource-efficient, exemplifies LLMs' adaptability to computational constraints [48]. These advancements underscore LLMs' transformative influence in specific NLP tasks, driving innovation and enhancing language processing applications' efficiency and effectiveness.

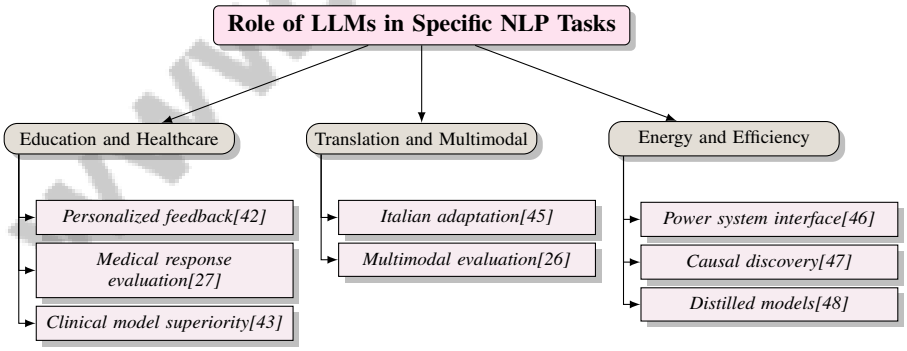


Figure 3: This figure illustrates the transformative role of Large Language Models (LLMs) in specific NLP tasks, highlighting their applications in education, healthcare, translation, multimodal settings, and energy efficiency. The figure categorizes the tasks into three primary areas, showcasing LLMs' adaptability and impact across diverse domains.

3.2 Advancements and Challenges in LLM Deployment

The deployment of Large Language Models (LLMs) has driven significant advancements in Natural Language Processing (NLP), but it also presents challenges, particularly related to hallucinations. Evaluations of models like GPT-4, GPT-3.5, BARD, GooglePaLM 2, and LLaMA-2 across varied

difficulty levels using datasets such as TruthfulQA, LSAT Reasoning, and Mathematical Reasoning reveal progress in model capabilities while highlighting persistent reliability issues [49]. A primary challenge is hallucination, where models generate factually incorrect or irrelevant information, exacerbated by extensive training corpora complicating data transfer and post-training processes [38]. A taxonomy for categorizing hallucination types has been introduced, providing a structured approach to understanding these challenges, emphasizing the distinct difficulties LLMs face compared to traditional natural language generation models [50].

Despite advancements in summarization techniques improving LLM outputs' efficiency and coherence, hallucinations remain a significant barrier to reliable deployment, as recent studies indicate [12]. The BLOOM 7B model exemplifies these advancements while also illustrating capacity limitations leading to hallucinations [17]. To mitigate these issues, integrating knowledge graphs into LLMs has been proposed to enhance factual reasoning capabilities, offering a promising framework for reducing hallucinations and improving model reliability [8]. Despite these advancements, LLM deployment continues to face challenges in effectively managing and mitigating hallucinations, essential for their reliable application across various contexts.

4 Understanding LLM Hallucination

An in-depth examination of hallucinations in Large Language Models (LLMs) highlights their tendency to produce confident yet erroneous responses due to biases in training data and misinterpretations of ambiguous prompts. These inaccuracies pose significant risks in sensitive domains like medical record summarization and legal advice, where even minor errors can have severe consequences. Addressing these complexities is crucial for the safe integration of LLMs into practical applications, as evidenced by over thirty-two mitigation techniques categorized by dataset utilization and feedback mechanisms [51, 52, 32]. This framework aims to analyze hallucination manifestations and guide efforts to mitigate their impact on LLM reliability.

4.1 Causes and Types of Hallucination in LLMs

Hallucinations in LLMs arise from both intrinsic and extrinsic factors. Intrinsically, LLMs often fail to recall and apply relevant knowledge accurately, leading to errors [8]. The substantial computational resources required for training, typically accessible only to large organizations, further limit insights into the mechanisms of emergent abilities [13]. The interplay between memory and hallucination suggests that LLMs with explicit memory mechanisms may reduce hallucination occurrences [18].

Extrinsically, biases in training datasets skew outputs towards misinformation, and the inability of LLMs to validate real-time information exacerbates inaccuracies in dynamic contexts. A systematic classification of hallucinations aids in understanding and identifying complex multi-tasking scenarios, supporting targeted mitigation strategies that enhance LLM reliability in critical applications such as medical summarization and legal advice [51, 53, 32]. Addressing these intrinsic and extrinsic factors is essential for ongoing research aimed at resolving ethical concerns and performance limitations in LLMs.

4.2 Manifestations and Challenges in Detecting Hallucinations

Hallucinations in LLMs manifest as outputs diverging from factual accuracy, posing significant challenges for detection and mitigation. These can include incorrect facts, irrelevant information, or entirely fabricated content. The models' inability to reliably attribute sources complicates verification, leading to potential misinformation [38, 39]. In keyword extraction, hallucinations may yield non-informative or irrelevant keywords, undermining accurate representation [37]. Similarly, in code generation, hallucinations can introduce errors, affecting output correctness [9]. The contamination of benchmarks with inadvertently leaked data further complicates fair performance assessment [39].

Detection is hindered by the inherent variability of LLM outputs, which can differ with identical inputs, making consistent detection criteria challenging [12]. Additionally, the lack of standardized benchmarks for hallucination detection limits effective methodology development [14]. Establishing comprehensive benchmarks and evaluation frameworks is essential for measuring the occurrence and consequences of hallucinations. Recent advancements emphasize the need for systematic categorization of hallucination instances and targeted interventions, including dynamic methods

that adapt to real-time assessments of model performance. Techniques like Retrieval-Augmented Generation and Knowledge Retrieval can enhance understanding of hallucinations, promoting safer LLM deployment in critical applications where errors can have significant repercussions [51, 54, 32]. These frameworks should account for hallucination manifestations and provide reliable metrics for evaluating factual accuracy. Integrating real-time validation and source attribution capabilities into LLMs could further enhance reliability and reduce hallucination prevalence.

4.3 Implications for LLM Applications

The implications of hallucinations in LLMs are substantial, affecting their deployment across various applications. Hallucinations, characterized by factually incorrect or fabricated outputs, challenge the reliability and trustworthiness of LLMs in critical domains. For instance, while LLMs demonstrate potential in automating ontology learning tasks, reliability issues necessitate addressing hallucinations to guide future research effectively [55]. LLMs' ability to enhance interpretability through natural language explanations is promising; however, hallucinations undermine these benefits, highlighting the need for robust mitigation strategies [56]. In scientific writing, AI's strengths in improving writing quality are well-documented, yet hallucinations can lead to misinformation, compromising scientific integrity [57].

Moreover, LLMs have shown improvements in participatory urban planning frameworks, enhancing stakeholder satisfaction. Nevertheless, the reliability of their outputs is crucial for maintaining trust and effective decision-making [36]. The development of a two-layer wider LLM network has improved text quality evaluation, yet the persistent issue of hallucinations calls for continuous model refinement [58]. The cognitive capabilities of LLMs, developed during pretraining and fine-tuning, underscore the complexity of their learning processes. Addressing hallucinations requires a comprehensive understanding of these phases to enhance the factual accuracy and expressive capabilities of LLM outputs [59]. While LLMs offer significant advancements across applications, the challenge of hallucinations remains a critical barrier to their reliable use, necessitating ongoing research and development to enhance robustness and trustworthiness.

5 Hallucination Mitigation Strategies

5.1 Taxonomy, Categorization, and Novel Mitigation Techniques

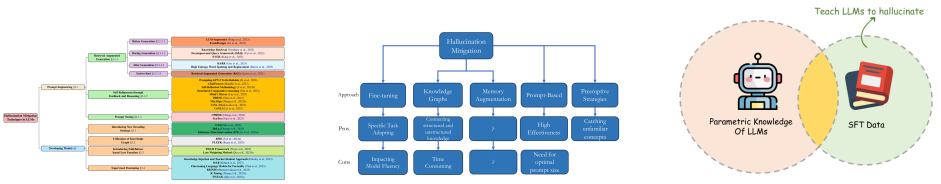
The taxonomy of hallucination mitigation techniques in Large Language Models (LLMs) is structured into data-level, model-level, and system-level approaches, each addressing specific challenges such as misleading content generation in critical applications like medical summarization and legal advice. This classification elucidates over thirty-two techniques, including Retrieval-Augmented Generation (RAG) and Knowledge Retrieval, that aim to enhance LLM output reliability and factual accuracy [51, 53, 32].

Data-level strategies improve model robustness through input augmentation. Techniques like retrieval-augmented generation integrate external knowledge sources, enhancing context relevance to mitigate hallucinations [60]. Knowledge-tuning, which incorporates structured medical knowledge bases, exemplifies effective data-level interventions for factual accuracy [14]. The LLM-TAKE framework reduces hallucinations and improves keyword relevance through structured quality checks [37].

Model-level strategies refine LLMs' internal mechanisms. The HALOCHECK framework quantifies hallucination severity by focusing on entailment relationships, representing an innovative approach [17]. Geometry-inspired methods, like scaling the readout vector, significantly enhance performance over existing model editing techniques [18]. Hybrid approaches illustrate the integration of sophisticated inference methods to detect and mitigate hallucinations [10].

System-level approaches develop innovative frameworks to enhance model performance. A novel taxonomy categorizes threats by their sources and types, offering a structured approach to addressing hallucinations [20]. LLM-Pruner employs structural pruning for model efficiency while reducing hallucinations [38]. Integrating knowledge graphs into LLMs, categorized into pre-training, during-training, and post-training enhancements, underscores the adaptability of system-level strategies [8].

By categorizing hallucination manifestations and outlining targeted mitigation strategies, this taxonomy enhances LLM application across fields like business process management, medical summarization, and complex decision-making tasks, promoting safe deployment in real-world scenarios [53, 32]. These strategies bolster LLM reliability and foster robust framework development.



(a) Prompts and Techniques for Hallucination Mitigation in Large Language Models[32]

(b) Hallucination Mitigation Approaches[53]

(c) Teaching Language Models to hallucinate: A Venn Diagram Explains the Intersection of Parametric Knowledge and SFT Data[50]

Figure 4: Examples of Taxonomy, Categorization, and Novel Mitigation Techniques

As illustrated in Figure 4, hallucination in LLMs poses significant challenges, prompting the development of various mitigation strategies. These are systematically categorized and explored through illustrative examples, offering a comprehensive taxonomy and novel techniques to enhance LLM output reliability and accuracy [32, 53, 50].

5.2 Real-Time and Interactive Approaches

Method Name	Knowledge Grounding	Iterative Refinement	Interactive Mechanisms
SF[61]	External Knowledge Independence	Refining Evaluation	Self-evaluation Technique
MLE[28]	Temporal Graph Database	Iterative Reflection Process	User Interactions
FTSE[62]	Verified Sources	Repeated Adjustments	User Interaction
DeCo[15]	Preceding Layers	Dynamic Modulation	User Interaction
MHM[63]	Knowledge Enrichment	Fine-tuning Lms	User Interaction
GCS[18]	Context-grounding	Model Editing	Memory Mechanisms
MM[35]	Verified Sources	Repeated Adjustments	Real-time Collaboration
PFSCDM[64]	Logical Reasoning Model	Iteratively Refine Text	User-friendly Tool

Table 1: This table presents a comparative analysis of various real-time and interactive methods aimed at mitigating hallucinations in large language models (LLMs). It highlights the key aspects of knowledge grounding, iterative refinement, and interactive mechanisms employed by each method. The table serves as a comprehensive overview of current strategies enhancing LLM reliability and applicability in dynamic environments.

Real-time and interactive approaches are crucial for mitigating hallucinations in LLMs, ensuring reliability in dynamic environments. Table 1 provides a comparative overview of real-time and interactive approaches employed to mitigate hallucinations in large language models, emphasizing the roles of knowledge grounding, iterative refinement, and interactive mechanisms. Knowledge grounding, where outputs are anchored to verified sources, enhances factuality and context relevance [61]. The Multi-LLM Orchestration Engine exemplifies this by integrating user interactions to build conversation graphs, improving real-time application factuality [28].

Iterative refinement processes are vital for interactive mitigation. Fine-tuning LLMs by computing semantic entropy and adjusting training labels based on uncertainty thresholds enhances accuracy and reduces hallucination likelihood [62]. Adaptive retrieval processes correct potential hallucinations by activating upon detecting inconsistencies [65].

The DeCo method dynamically selects MLLM layers to integrate their knowledge into the final output, correcting hallucinations during inference [15]. The MHM method involves fine-tuning models to retrieve accurate information and suppress incorrect predictions, supporting real-time mitigation [63].

Interactive methods like SELF-FAMILIARITY evaluate model familiarity with input concepts, preventing hallucinations [61]. Generation Constraint Scaling adjusts the readout vector in memory-augmented LLMs to constrain output generation, reducing hallucinations [18].

Real-time collaboration among LLMs enhances abstention decision accuracy by leveraging multiple model feedback, improving output reliability [66]. The Mind's Mirror method uses diverse Chain-of-Thoughts (CoTs) and self-evaluation outputs for training smaller models, exemplifying self-evaluation integration [35].

Triggering self-contradictions and detecting them using logical reasoning refines generated text while preserving fluency, providing an interactive mechanism for hallucination mitigation [64]. These approaches significantly enhance LLM reliability and applicability across dynamic and high-stakes environments.

5.3 Evaluation Frameworks and Benchmarks

Benchmark	Size	Domain	Task Format	Metric
LLM-EB[39]	1,000,000	Evaluation Benchmarking	Question Answering	Accuracy, F1-score
LLM-Bench[41]	23	Language Understanding	Multiple Choice Questions (mcqs)	Accuracy, F1-score
HALLUCODE[9]	5,663	Software Engineering	Code Generation	Accuracy of Hallucination Existence Recognition, Valid Rate
HalluDial[67]	146,856	Dialogue Systems	Hallucination Detection	F1-score, Accuracy
LLM-LR[68]	24,802	Natural Language Processing	Sentiment Analysis	Accuracy, F1-score
LLM-as-a-Judge[42]	100	Education	Text Evaluation	Grammaticality, Fluency
CLSB[44]	199,000	Clinical Text Summarization	Summarization	BLEU, ROUGE-L
LLMs4OL[55]	3,000,000	Biomedicine	Term Typing	MAP@1, F1-score

Table 2: Table presents a comprehensive overview of various benchmarks used in evaluating large language models (LLMs) across different domains. The table details the benchmark names, their sizes, the specific domains they cover, the task formats they employ, and the metrics used for performance evaluation. This information is crucial for understanding the scope and applicability of each benchmark in assessing hallucination mitigation strategies in LLMs.

Comprehensive evaluation frameworks and benchmarks are essential for assessing hallucination mitigation strategies in LLMs. The benchmark by [39] emphasizes avoiding data contamination and employs diverse evaluation methods, crucial for reliable LLM performance assessments. This benchmark provides guidelines that enhance evaluation integrity and LLM output reliability.

A unified evaluation framework proposed by [41] critically assesses existing benchmarks, improving accuracy and comprehensiveness. It addresses current inadequacies and suggests enhancements for holistic evaluations, particularly in hallucination detection and mitigation.

Metrics are pivotal in evaluating LLM effectiveness in identifying hallucinations. Appropriate metric selection, reflecting theoretical and practical considerations, ensures evaluations capture performance nuances in real-world scenarios [9]. These metrics provide a quantitative basis for comparing mitigation strategies, informing more effective approaches.

Table 2 provides a detailed overview of the key benchmarks employed in evaluating large language models, highlighting their relevance in the context of hallucination mitigation and performance assessment. These frameworks and benchmarks are crucial for advancing LLM hallucination mitigation. They equip researchers and practitioners with tools to rigorously assess performance, ensuring safe deployment across applications. Future research should refine automated synthesis tools by developing domain-specific benchmarks tailored to unique challenges, such as systematic reviews and keyword extraction. Studies indicate advanced methodologies can significantly reduce hallucination and improve synthesis reliability in academic research. Exploring LLMs like GPT-3.5 and Llama2-7B in domain-driven keyword extraction can optimize information retrieval and categorization. Addressing these needs can establish robust frameworks ensuring methodological transparency and reliability across diverse research domains [69, 70].

6 Challenges and Future Directions

6.1 Current Challenges and Advancements in Hallucination Mitigation

The complexity and resource demands of Large Language Models (LLMs) pose significant challenges for hallucination mitigation. A critical issue is the dependency on specific datasets, which can

introduce biases and limit generalizability across communication tasks [10]. Existing benchmarks often focus on binary classification, failing to capture the nuances necessary for effective content detection, especially in collaborative contexts [7]. In the medical field, benchmarks frequently neglect the integration of structured medical knowledge, crucial for mitigating hallucinations in specialized applications [14]. Incorporating knowledge graphs into LLMs can introduce noise and increase computational demands [8]. Benchmark leakage further undermines LLM evaluations, necessitating stringent guidelines to prevent data contamination [39]. Model compression techniques like pruning aim to enhance efficiency but often degrade performance at high pruning rates, complicating efforts to maintain model integrity [38]. Current benchmarks primarily assess functional correctness, overlooking the critical issue of hallucinations, which can lead to unreliable outputs in tasks such as code generation [9]. Despite these challenges, advancements have been made, such as strategies reducing sycophantic behavior to enhance output reliability [16] and the DeCo method, which dynamically corrects hallucinations without extensive retraining, showing promise in adapting to various Multimodal Large Language Models (MLLMs) [15]. Future research should focus on developing standardized evaluation metrics and generalizable mitigation techniques, essential for accurately assessing LLM capabilities and addressing current benchmark inadequacies [41]. Optimizing LLM architectures and aligning them with human values are crucial for enhancing output reliability and trustworthiness.

6.2 Future Research Directions

Future research should prioritize the development of advanced uncertainty metrics applicable across diverse architectures, improving output reliability [10]. Optimizing extraction algorithms and refining models with domain-specific datasets, integrating human expertise in keyword extraction, can enhance accuracy and relevance [37]. Incorporating user feedback mechanisms and alternative model architectures can continually improve dataset quality and model performance. Refining detection methods and investigating novel mitigation techniques are crucial for addressing hallucinations in various contexts. Applying NLI-based re-rankers to additional NLP tasks and exploring efficient integration strategies for NLI in decoding algorithms represent promising avenues [12]. In MLLMs, combining DeCo with other strategies and expanding its application across broader models could enhance effectiveness [14]. Further research should focus on developing standardized frameworks for evaluating both quantitative and qualitative intelligence in LLMs, considering implications for super-human AI capabilities [22]. Understanding sycophantic tendencies and extending analysis to other languages and contexts will yield deeper insights into model behavior and mitigation strategies [7]. Addressing data privacy concerns when accessing databases and integrating real-world questions requiring multiple database accesses is another critical direction [31]. Enhancing LLM reasoning capabilities, developing robust hallucination mitigation techniques, and addressing security challenges in LLM-powered communication systems are essential for advancing the field [10]. Improving recovery processes and exploring methods to enhance performance at higher pruning rates will advance model efficiency [38]. Optimizing collaboration among LLMs to enhance knowledge retrieval and abstention capabilities presents significant opportunities for improving model reliability [66]. Developing robust evaluation frameworks that minimize data contamination risks and enhance LLM assessment reliability is crucial for the field's advancement [39]. Pursuing these research directions can significantly improve the reliability and trustworthiness of LLMs, facilitating successful deployment in diverse applications.

6.3 Ethical and Interdisciplinary Considerations

Deploying LLMs in practical applications requires a robust ethical framework to manage potential risks and ensure responsible usage. Ethical considerations are critical in high-stakes domains like healthcare, where careful management of LLM implementations is essential to prevent harm and ensure precision [27]. The tendency of LLMs to generate misinformation or hallucinations underscores the need for continuous risk assessments and user education, emphasizing ethical guidelines prioritizing accuracy and reliability [41]. Biases in training data can lead to incorrect outputs and ethical dilemmas, necessitating adherence to guidelines emphasizing privacy, confidentiality, and excluding personally identifiable information. Maintaining human oversight in LLM evaluations is essential to uphold ethical standards [27]. Interdisciplinary collaboration is vital to address complex challenges posed by LLMs. Integrating insights from computer science, ethics, and policy-making can enhance security and reliability [71]. Collaboration fosters standardized benchmarks for evaluating

natural language generation systems, promoting joint efforts to tackle hallucination effectively [72]. Ethical deployment of closed-source models like GPT-4 requires transparency and interdisciplinary cooperation to ensure thorough ethical considerations [73]. Comprehensive ethical frameworks and interdisciplinary strategies are essential for responsible LLM deployment. Addressing unique ethical challenges such as hallucination, verifiable accountability, and censorship complexities requires tailored guidelines and dynamic auditing systems reflecting specific contexts. Collaboration among stakeholders from academia, industry, government, and civil society is crucial to navigating ethical, legal, and societal implications, ensuring transformative potential is harnessed while mitigating risks like bias and misinformation [57, 37, 74]. By prioritizing ethical considerations and fostering collaboration, the field can progress towards more reliable and trustworthy AI systems, enhancing positive societal impact.

6.4 Technological and Methodological Innovations

Technological and methodological innovations are pivotal in enhancing hallucination mitigation in LLMs. Significant advancements include methods improving contextual understanding in causal discovery tasks, potentially leading to more accurate outputs. Investigating closed-source models may excel in handling complex causal relationships [47]. ReCaption, with its dual-component structure combining caption rewriting and fine-tuning, effectively reduces fine-grained object hallucinations, improving LLM outputs in visual contexts [75]. Integrating dual-component systems into LLM frameworks can significantly enhance performance in tasks requiring precise visual and textual alignment. GraphEval offers a systematic approach to evaluating LLM outputs, identifying inconsistencies and enhancing explainability. This knowledge graph-based method provides a robust framework for assessing factual accuracy, contributing to reliable and transparent AI systems [76]. The SELF-FAMILIARITY approach proactively prevents hallucinations and maintains consistent performance across instruction styles and model types, highlighting adaptability in mitigation strategies [61]. Future research should focus on improving synergy between retrievers and LLMs, potentially through joint training or alternative architectures, leading to cohesive systems enhancing hallucination mitigation effectiveness [77]. Pursuing these technological and methodological innovations can advance towards reliable and trustworthy LLMs, delivering accurate and contextually relevant outputs across diverse applications.

7 Conclusion

This survey examines the critical challenge of hallucinations in Large Language Models (LLMs) and underscores the necessity for effective mitigation strategies to enhance their reliability across various applications. Notable progress has been achieved in understanding and addressing hallucinations, with innovative techniques such as CONFIDENCE-DRIVEN INFERENCE enhancing sample size and coverage, thereby ensuring robust statistical validity. The significance of contextual understanding in refining model outputs is highlighted, with effective methodologies identified for hallucination detection and mitigation. Post-ChatGPT developments have been pivotal, providing a foundation for ongoing research into hallucination management. A newly proposed framework for identifying and resolving self-contradictions in LLM outputs demonstrates promising results in maintaining text quality while improving accuracy, showcasing the potential for increased dependability in LLM-generated content. These findings collectively stress the imperative for sustained research efforts to develop advanced hallucination mitigation strategies, ensuring the safe and efficient deployment of LLMs in essential sectors.

References

[1] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. Psy-llm: Scaling up global mental health psychological services with ai-based large language models, 2023.

[2] Jeremy Straub and Zach Johnson. Initial development and evaluation of the creative artificial intelligence through recurring developments and determinations (cairdd) system, 2024.

[3] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey, 2024.

[4] Haokun Liu, Yaonan Zhu, Kenji Kato, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. Llm-based human-robot collaboration framework for manipulation tasks, 2023.

[5] Qi She, Junwen Pan, Xin Wan, Rui Zhang, Dawei Lu, and Kai Huang. Mammothmoda: Multi-modal large language model, 2024.

[6] João Pedro Gandarela, Danilo S. Carvalho, and André Freitas. Inductive learning of logical theories with llms: An expressivity-graded analysis, 2025.

[7] Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, Aleksandr Gordeev, Vladimir Dokholyan, and Maksim Kuprashevich. Gigacheck: Detecting llm-generated content, 2024.

[8] Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling, 2024.

[9] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*, 2024.

[10] Yinqiu Liu, Guangyuan Liu, Ruichen Zhang, Dusit Niyato, Zehui Xiong, Dong In Kim, Kaibin Huang, and Hongyang Du. Hallucination-aware optimization for large language model-empowered communications, 2024.

[11] 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.

[12] Arvind Krishna Sridhar and Erik Visser. Improved beam search for hallucination mitigation in abstractive summarization, 2023.

[13] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[14] Haochun Wang, Sendong Zhao, Zewen Qiang, Zijian Li, Nuwa Xi, Yanrui Du, MuZhen Cai, Haoqiang Guo, Yuhan Chen, Haoming Xu, Bing Qin, and Ting Liu. Knowledge-tuning large language models with structured medical knowledge bases for reliable response generation in chinese, 2023.

[15] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024.

[16] 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.

[17] 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, 2023.

[18] Georgios Kollias, Payel Das, and Subhajit Chaudhury. Generation constraint scaling can mitigate hallucination, 2024.

[19] Vipula Rawte, Aman Chadha, Amit Sheth, and Amitava Das. Tutorial proposal: Hallucination in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries*, pages 68–72, 2024.

[20] 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.

[21] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

[22] Nils Körber, Silvan Wehrli, and Christopher Irrgang. How to measure the intelligence of large language models?, 2024.

[23] Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. "we need structured output": Towards user-centered constraints on large language model output, 2024.

[24] Bibiána Lajčinová, Patrik Valábek, and Michal Spišiak. Intent classification for bank chatbots through llm fine-tuning, 2024.

[25] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models, 2023.

[26] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, 2023.

[27] Jack Krolik, Herprit Mahal, Feroz Ahmad, Gaurav Trivedi, and Bahador Saket. Towards leveraging large language models for automated medical qa evaluation, 2024.

[28] Sumedh Rasal. A multi-llm orchestration engine for personalized, context-rich assistance, 2024.

[29] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. A survey on large language models with multilingualism: Recent advances and new frontiers, 2025.

[30] Artem Lykov and Dzmitry Tsetserukou. Llm-brain: Ai-driven fast generation of robot behaviour tree based on large language model, 2023.

[31] Zongyue Qin, Chen Luo, Zhengyang Wang, Haoming Jiang, and Yizhou Sun. Relational database augmented large language model, 2024.

[32] 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.

[33] Sen Huang, Kaixiang Yang, Sheng Qi, and Rui Wang. When large language model meets optimization, 2024.

[34] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. Explanations from large language models make small reasoners better, 2022.

[35] Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. Mind's mirror: Distilling self-evaluation capability and comprehensive thinking from large language models, 2024.

[36] Zhilun Zhou, Yuming Lin, and Yong Li. Large language model empowered participatory urban planning, 2024.

[37] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Llm-take: Theme aware keyword extraction using large language models, 2023.

[38] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

[39] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

[40] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models, 2024.

[41] 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.

[42] Seungyoon Kim and Seungone Kim. Can language models evaluate human written text? case study on korean student writing for education, 2024.

[43] Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models?, 2023.

[44] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. Adapted large language models can outperform medical experts in clinical text summarization, 2024.

[45] Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. Llamantino: Llama 2 models for effective text generation in italian language, 2023.

[46] Subir Majumder, Lin Dong, Fatemeh Doudi, Yuting Cai, Chao Tian, Dileep Kalathi, Kevin Ding, Anupam A. Thatte, Na Li, and Le Xie. Exploring the capabilities and limitations of large language models in the electric energy sector, 2024.

[47] Tao Feng, Lizhen Qu, Niket Tandon, Zhuang Li, Xiaoxi Kang, and Gholamreza Haffari. From pre-training corpora to large language models: What factors influence llm performance in causal discovery tasks?, 2024.

[48] Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. Lamini-lm: A diverse herd of distilled models from large-scale instructions, 2024.

[49] Aniket Kumar Singh, Suman Devkota, Bishal Lamichhane, Uttam Dhakal, and Chandra Dhakal. The confidence-competence gap in large language models: A cognitive study, 2023.

[50] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[51] 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.

[52] Hazel Kim, Adel Bibi, Philip Torr, and Yarin Gal. Detecting llm hallucination through layer-wise information deficiency: Analysis of unanswerable questions and ambiguous prompts, 2024.

[53] Alessandro Bruno, Pier Luigi Mazzeo, Aladine Chetouani, Marouane Tliba, and Mohamed Amine Kerkouri. Insights into classifying and mitigating llms' hallucinations. *arXiv preprint arXiv:2311.08117*, 2023.

[54] Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. Constructing benchmarks and interventions for combating hallucinations in llms, 2024.

[55] Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. Llms4ol: Large language models for ontology learning, 2023.

[56] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models, 2024.

[57] Ahmed S. BaHammam, Khaled Trabelsi, Seithikurippu R. Pandi-Perumal, and Hiatham Jahrami. Adapting to the impact of ai in scientific writing: Balancing benefits and drawbacks while developing policies and regulations, 2023.

[58] Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*, 2023.

[59] Yuzi Yan, Jialian Li, Yipin Zhang, and Dong Yan. Exploring the llm journey from cognition to expression with linear representations, 2024.

[60] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[61] Junyu Luo, Cao Xiao, and Fenglong Ma. Zero-resource hallucination prevention for large language models. *arXiv preprint arXiv:2309.02654*, 2023.

[62] Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, Kunal Handa, and Yarin Gal. Fine-tuning large language models to appropriately abstain with semantic entropy, 2024.

[63] Lei Yu, Meng Cao, Jackie Chi Kit Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations. *arXiv preprint arXiv:2403.18167*, 2024.

[64] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*, 2023.

[65] Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models, 2024.

[66] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

[67] Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation, 2024.

[68] Md. Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. Do large language models speak all languages equally? a comparative study in low-resource settings, 2024.

[69] 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.

[70] Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. Comparative study of domain driven terms extraction using large language models, 2024.

[71] Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. Securing large language models: Threats, vulnerabilities and responsible practices, 2024.

[72] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[73] Sanjana Ramprasad, Elisa Ferracane, and Zachary C Lipton. Analyzing llm behavior in dialogue summarization: Unveiling circumstantial hallucination trends. *arXiv preprint arXiv:2406.03487*, 2024.

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

[75] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer, 2024.

[76] Hannah Sansford, Nicholas Richardson, Hermina Petric Maretic, and Juba Nait Saada. Grapheval: A knowledge-graph based llm hallucination evaluation framework, 2024.

[77] Patrice Béchard and Orlando Marquez Ayala. Reducing hallucination in structured outputs via retrieval-augmented generation, 2024.

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.