Literature Survey on Hallucination Evaluation in Large Language Models: State-of-the-Art (SOTA)

Yash Khare, Mohan Bhosale, Karthikeyan Sugavanan
khare.y@northeastern.edu,bhosale.m@northeastern.edu,sugavanan.k@northeastern.edu
Khoury College of Computer Sciences, Northeastern University
Boston, MA

Submission Date: 08 December 2024

ABSTRACT

Large Language Models (LLMs) and Multimodal Large Language Models (LMMs) have significantly advanced artificial intelligence by enabling machines to generate and understand human-like text with high accuracy. However, hallucinations—where models produce plausible yet incorrect or nonsensical information—remain a critical challenge, especially in high-stakes applications such as healthcare and education. This literature survey analyzes hallucinations in LLMs and LMMs by reviewing ten key studies from 2020 to 2024. It identifies three main themes: improvements in reasoning capabilities, domain-specific applications, and enhancements in reliability and robustness. The survey evaluates methodologies like Chain-of-Thought prompting and real-time augmentation, and benchmarks such as DAHL and HalluQA that address specialized domains. Additionally, it highlights research gaps including limited interpretability, energy inefficiency, and the need for greater cultural and linguistic diversity. The findings underscore the importance of ethical deployment and the development of more reliable evaluation frameworks. Future research directions include integrating structured and unstructured data, improving energy efficiency, expanding cultural benchmarks, and enhancing model interpretability. This survey aims to guide researchers and practitioners in developing more trustworthy and accurate language models.

Keywords: *Large Language Models, Hallucination Evaluation, Reasoning Capabilities, Domain-Specific Applications, Reliability, Multimodal Models, Mitigation Strategies*

1 INTRODUCTION

1.1 Background

Large Language Models (LLMs) have significantly advanced the field of artificial intelligence by enabling machines to understand and generate human-like text with unprecedented accuracy and coherence [3, 20]. Models such as GPT-3 and its successors have demonstrated remarkable abilities in tasks ranging from automated content creation to sophisticated conversational agents [15, 16]. These advancements have not only set new benchmarks in natural language processing but have also catalyzed extensive research aimed at improving their capabilities and addressing inherent challenges.

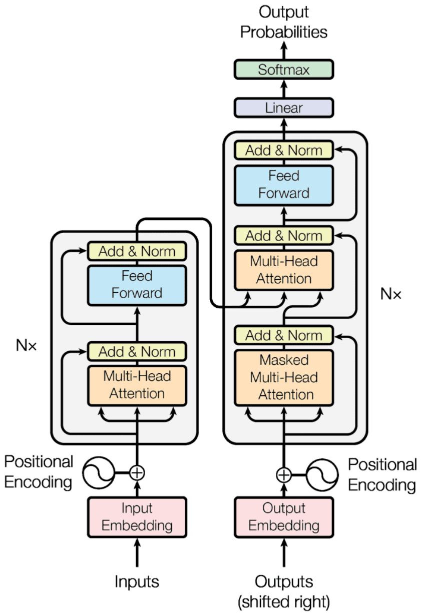

Figure 1: Architecture of a Large Language Model

One of the most critical challenges associated with LLMs is the phenomenon of hallucinations, where models generate information that appears plausible but is factually incorrect or nonsensical [5, 28]. Hallucinations undermine the reliability of LLMs, particularly in applications requiring high accuracy, such as medical diagnosis, legal advice, and academic research [11, 13]. Understanding and mitigating hallucinations is essential for enhancing the trustworthiness and practical utility of LLMs across various domains.

1.2 Objective

The primary objective of this survey is to provide a comprehensive analysis of hallucinations in Large Language Models. This involves exploring the underlying causes, assessing the impact on model performance and user trust, and evaluating the effectiveness of existing mitigation strategies. By systematically reviewing the current literature, this survey aims to:

- **Identify Contributing Factors:** Examine the key factors that lead to hallucinations in LLMs, including training data quality, model architecture, and inference mechanisms [2, 26].
- **Assess Impact:** Evaluate how hallucinations affect the reliability and credibility of LLM-based applications in various real-world scenarios [22, 31].
- **Evaluate Mitigation Techniques:** Review and compare the effectiveness of different approaches aimed at reducing or eliminating hallucinations, such as fine-tuning,

reinforcement learning, and external knowledge integration [10, 24].

- **Identify Research Gaps:** Highlight areas where current research is lacking and suggest potential directions for future studies to enhance the factual consistency of LLMs [7, 23].

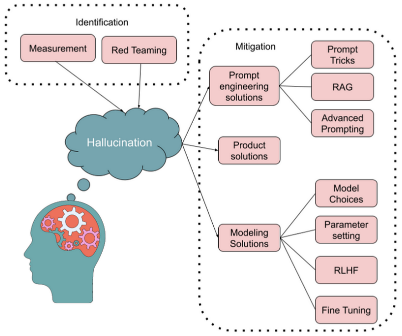

Figure 2: Overview of Hallucination Mitigation Strategies

Through these objectives, the survey seeks to provide valuable insights for researchers, developers, and practitioners working with LLMs, facilitating the development of more reliable and accurate language models.

1.3 Scope

This survey encompasses a wide range of scholarly articles, conference papers, and technical reports published between 2019 and 2024. The sources include reputable databases such as *arXiv*, IEEE Xplore, ACM Digital Library, and SpringerLink. The focus is on studies that investigate the nature and causes of hallucinations in LLMs, explore their implications across different applications, and propose solutions to mitigate these inaccuracies.

The survey prioritizes works that offer empirical evaluations, theoretical analyses, and practical interventions related to hallucination mitigation. It includes comparative studies of various LLM architectures and training methodologies to understand their susceptibility to generating hallucinations [1, 8]. Additionally, the review covers innovative techniques that integrate external knowledge bases and employ advanced training strategies to enhance factual consistency [19, 29].

By concentrating on recent advancements, the survey aims to present an up-to-date overview of the current state of research on hallucinations in LLMs. This ensures that the findings and recommendations are relevant and can inform ongoing and future developments in the field of Large Language Models.

2 METHODOLOGY

2.1 Search Strategy

A thorough strategy was implemented to locate pertinent scientific literature by utilizing a variety of databases and search platforms, including Google Scholar, IEEE Xplore, arXiv, the ACM Digital Library, and SpringerLink.

For each publisher outlined in Table 1, a targeted Google search was executed, focusing on the first page of results using specific keywords. This approach ensured the inclusion of the most relevant and high-quality papers, as these are typically featured on the initial page of search outcomes.

2.2 Keywords

The identification of relevant studies was guided by the following specific keywords:

- “LLM Hallucinations”
- “Factual Consistency in LLM”
- “Mitigating LLM Hallucinations”

The search was restricted to publications from 2020 to 2024, capturing the latest developments in the field of Large Language Models (LLMs) and the associated challenges they present.

2.3 Selection Criteria

The process of selecting papers involved several filtering steps to ensure both relevance and quality:

- **Relevance of First Page Results:** Priority was given to papers appearing on the first page of search results, as these are generally more pertinent and influential.
- **Peer-Reviewed Sources:** Although archived papers (e.g., those from arXiv) were considered to provide a comprehensive overview, preference was given to peer-reviewed publications.
- **Assessment of Contributions:** Each paper was reviewed for its contributions, including the datasets utilized, models developed, and solutions proposed for issues in LLMs, such as hallucinations and factual inaccuracies.
- **Citation Metrics and Publication Quality:** Papers with higher citation counts or those published in reputable journals were given higher priority.

2.4 Distribution of Scientific Papers

To summarize the selected papers, **Table 1: Distribution of Scientific Papers (2020–2024)** categorizes them based on three main keywords—**LLM Hallucinations**, **Factual Consistency in LLM**, and **Mitigating LLM Hallucinations**—each further subdivided into **Multimodal** and **Language** categories.

The distribution reflects the publishers identified during the search process. As depicted in the table, the most relevant papers were generally found on the first page of search results, with the quality of results diminishing on subsequent pages. This underscores the effectiveness of a focused search strategy.

2.5 Rationale for Distribution

The chosen papers encompass a diverse range of publishers, reflecting the current state of research in the specified areas. For example, arXiv contributed a substantial number of papers due to its role in disseminating preprints and early-stage research, whereas IEEE and ACM provided high-quality, peer-reviewed publications. The classification into multimodal and language subcategories facilitates the identification of emerging trends and existing gaps, such as the significant role of multimodal approaches in addressing hallucinations.

Table 1: Distribution of Scientific Papers

#	Publisher	LLM Hallucinations		Factual Consistency in LLM		Mitigating LLM Hallucinations		Total
		Multimodal	**Language**	**Multimodal**	**Language**	**Multimodal**	**Language**	
1	IEEE	4	3	0	1	2	3	13
2	arXiv	10	10	7	9	10	10	56
3	ACM	2	4	3	3	1	2	15
4	Springer	0	3	2	1	1	2	9
5	Science Direct	2	1	2	2	2	1	10
Total		**18**	**21**	**14**	**16**	**16**	**18**	**103**

3 LITERATURE REVIEW

In this section, we present a detailed survey of current research on hallucination evaluation in Large Language Models (LLMs) and Multimodal Large Language Models (LMMs). We begin with a thematic analysis (Section 3.1), identifying key motifs in the progression of benchmarks—ranging from enhanced reasoning capabilities to domain-specialized assessments and reliability improvements. We then provide a comparative analysis (Section 3.2) that delves deeper into the methods, dataset types, and evaluation challenges that define the landscape of hallucination research. Tables 2 and 3 offer structured summaries of representative benchmarks, highlighting their thematic contributions and comparative attributes, respectively.

3.1 Thematic Analysis

The thematic analysis reveals the multifaceted nature of hallucination evaluation benchmarks, underscoring three major themes: (1) advancements in reasoning capabilities, (2) domain-specific applications, and (3) efforts to improve reliability and robustness.

3.1.1 Reasoning Capabilities. A central thread in hallucination evaluation is the progressive refinement of LLMs' reasoning capabilities. Techniques such as Chain-of-Thought (CoT) prompting allow models to perform multi-step logical inferences, thus enabling more reliable and contextually coherent outputs. Benchmarks like ChainPoll and SAC3 push these boundaries by introducing reasoning consistency checks—cross-model response consistency in SAC3 and adherence/correctness metrics in ChainPoll—thereby facilitating more nuanced assessments of reasoning quality [6, 27].

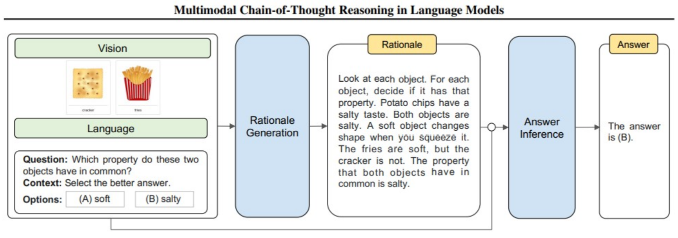
Multimodal Chain-of-Thought Reasoning in Language Models
Vision
cracker
fries
Language
Question: Which property do these two objects have in common?
Context: Select the better answer.
Options: (A) soft (B) salty
Rationale Generation
Rationale
Look at each object. For each object, decide if it has that property. Potato chips have a salty taste. Both objects are salty. A soft object changes shape when you squeeze it. The fries are soft, but the cracker is not. The property that both objects have in common is salty.
Answer Inference
Answer
The answer is (B).

Figure 3: Chain-of-Thought Prompting in LLMs

The emphasis on reasoning extends beyond text-based modalities. For example, ViBe focuses on temporal reasoning within text-to-video generation tasks, ensuring alignment of generated visual content with temporal logic in input prompts [17]. This evolution from simple factual checks toward complex, multi-step cognitive operations signals a gradual shift toward evaluating and ensuring the internal consistency, logical flow, and fidelity of reasoning processes within LLMs and LMMs.

3.1.2 Domain-Specific Applications. Benchmarks have moved beyond general-purpose tests to meet the specialized demands of fields such as biomedicine, cultural studies, and vision-language integration. Domain-specific benchmarks, like DAHL, dissect factual conflicts in the biomedical sphere by meticulously evaluating model outputs at the atomic level [18]. HalluQA targets Chinese linguistic and cultural contexts, ensuring that hallucination detection is not limited to Western language models and extending the evaluation to reflect cultural diversity [4]. Similarly, Reefknot focuses on multimodal relation reasoning, examining how well LMMs link visual and textual cues without inventing unsupported relationships [30].

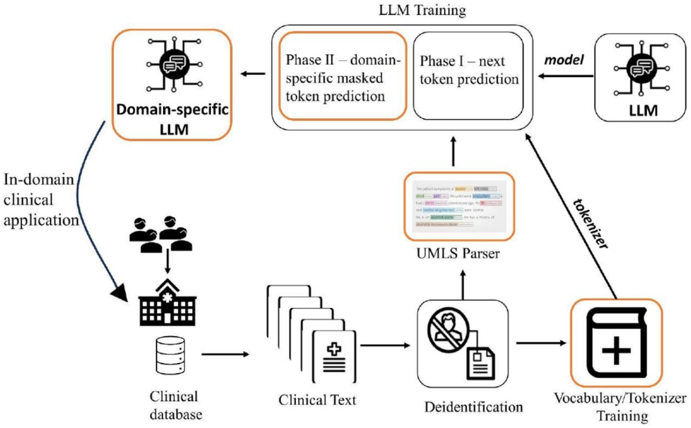
LLM Training
Domain-specific LLM
Phase II – domain-specific masked token prediction
Phase I – next token prediction
model
LLM
In-domain clinical application
UMLS Parser
tokenizer
Clinical database
Clinical Text
Deidentification
Vocabulary/Tokenizer Training

Figure 4: Domain-Specific Evaluation in Biomedicine

These domain-oriented benchmarks fill critical gaps, pushing models to operate reliably in high-stakes and specialized environments. They also highlight how the evaluation of hallucinations must be sensitive to domain conventions, terminologies, and the nuanced types of errors that arise outside of general, open-domain tasks.

3.1.3 Reliability Improvements. Reliability remains a central concern. Benchmarks increasingly incorporate innovative mechanisms to detect, mitigate, and reduce hallucinations. For instance,

Table 2: Benchmark Analysis for Evaluating Hallucinations in Large Language Models

Benchmark Name	Reasoning	Domain-Specific	Reliability	Unique Contributions
DAHL, 2024	Fact-conflict analysis	Biomedical	In-depth evaluation over multiple-choice methods	• 8,573 curated biomedical questions • DAHL Score for atomic accuracy • Minimal accuracy gain beyond 7-8B parameters
Reefknot, 2024	Relation hallucination	Visual Genome	Comprehensive over object/attribute benchmarks	• 20,000+ real-world samples • Defined relation hallucinations • VG-based corpus • 9.75% reduction via mitigation
ViBe, 2024	Five hallucination types	Text-to-Video	Addresses video generation gaps	• Five defined hallucination types • 3,782 annotated videos • Ensemble classifier baseline • TimeSFormer + CNN performance
SAC3, 2023	Question and model-level	Black-box Evaluations	Improved over self-consistency methods	• Two hallucination types • Equivalent question perturbation • Cross-model consistency checking • Superior detection performance
FACTOR, 2024	True vs. incorrect facts	Wiki, News, Expert	Controlled and domain-specific facts	• Factual corpus transformation • Score improvement with size and retrieval • Better factuality reflection • Scalable across domains
HalluQA, 2023	Imitative and factual errors	Chinese-Specific Domains	Addresses Chinese LLM evaluations	• 450 adversarial questions • Two hallucination types • GPT-4 automated evaluation • Analysis of types and causes
ChainPoll, 2023	Incorrect claims	Open-domain QA	Higher AUROC than existing metrics	• High-performing detection method • RealHall benchmark datasets • Adherence and Correctness metrics • Cost-effective and transparent
FreshQA, 2023	Current knowledge	Real-time QA Tasks	Two-mode correctness and hallucination	• Dynamic QA benchmark • FreshPrompt method • Concise answers reduce hallucination • Impact of evidence retrieval
AlphaIntellect, 2024	Chain-of-Thought	Open-domain QA	Improved binary classification	• Participated in SemEval-2024 Task 6
HQH, 2024	Hallucination quality	Vision-Language QA	Consistency in benchmarks	• HQM framework • Test-retest reliability • Criterion validity and coverage • High-quality LVLM benchmark • Evaluates 10+ LVLMs
RealTime QA, 2024	Current events	Open-domain QA	Challenges static QA datasets	• Weekly dynamic QA platform • Extended real-time evaluations • Importance of up-to-date retrieval • Identifies outdated answer tendencies

Table 3: Comparative Analysis Table

Benchmark Name	Methods Used	Dataset Type	Challenges Addressed	Evaluation Metrics	Baseline Models Evaluated	Year
DAHL	Deconstructing responses into atomic units and evaluating fact-conflicting hallucinations	Biomedical domain-specific dataset with 8,573 questions across 29 categories	Hallucination in long-form text generation within the biomedical domain	DAHL Score (average accuracy of atomic units)	GPT-4, BioGPT	2024 (submitted on 14 Nov 2024)
Reefknot	Systematic definition of relation hallucinations, comparative evaluation across three tasks, and confidence-based mitigation strategy	Real-world scenario dataset with over 20,000 samples derived from Visual Genome (VG) scene graph dataset	Relation hallucinations in Multimodal Large Language Models (MLLMs)	Hallucination rate reduction	Current MLLMs	2024 (submitted on 18 Aug 2024)
ViBe	Classification of hallucinations into five categories, ensemble classifier configurations	Large-scale dataset of 3,782 hallucinated videos annotated by humans	Hallucinations in Text-to-Video (T2V) models	Accuracy and F1 score	10 open-source T2V models	2024 (submitted on 16 Nov 2024)
SAC3	Sampling-based method using semantically equivalent question perturbation and cross-model response consistency checking	Multiple question-answering and open-domain generation benchmarks	Hallucinations in language models, including question-level and model-level hallucinations	Detection performance for non-factual and factual statements	Gpt-3.5-turbo, Falcon-7b, Guanaco-33b	2023 (submitted on 3 Nov 2023)
FACTOR	Automatic transformation of factual corpus into benchmark, evaluating LM's propensity to generate true facts vs. similar incorrect statements	Three benchmarks: Wiki-FACTOR, News-FACTOR, and Expert-FACTOR	Factuality evaluation of language models in specific domains	Benchmark scores, comparison with perplexity	GPT-2, GPT-Neo, OPT	2024 (last revised on 4 Feb 2024)
HalluQA	Adversarial question design, automated evaluation using GPT-4	450 meticulously designed adversarial questions spanning multiple domains, considering Chinese historical culture, customs, and social phenomena	Hallucinations in Chinese large language models	Non-hallucination rates	24 large language models including ERNIE-Bot, Baichuan2, ChatGLM, Qwen, SparkDesk	2023 (submitted on 5 Oct 2023, last revised on 25 Oct 2023)
ChainPoll	Innovative hallucination detection method, introducing RealHall benchmark dataset	RealHall: refined collection of four datasets challenging for modern LLMs and relevant to real-world scenarios	Hallucinations in Large Language Model (LLM) outputs	AUROC, Adherence, Correctness	Various LLMs (GPT-3.5-turbo, GPT-4)	2023 (submitted on 22 Oct 2023)
FreshQA	Dynamic question answering platform with weekly question announcements and evaluations	Dynamic QA benchmark with diverse question types, including fast-changing world knowledge and questions with false premises	Answering questions about up-to-date information, challenging static assumptions in open-domain QA	Real-time evaluation results	GPT-3, T5	2024 (last revised on 28 Feb 2024)
AlphaIntellect (SemEval)	Fine-tuned binary classifiers	Open-domain QA	Detecting hallucinations in general tasks	Accuracy	Not specified	2024
HQH	Hallucination benchmark Quality Measurement framework (HQM)	Vision-language general QA	Hallucinations in Large Vision-Language Models (LVLMs)	Test-retest reliability, parallel-forms reliability, criterion validity, coverage of hallucination types	Over 10 representative LVLMs, including GPT-4o and Gemini-1.5-Pro	2024 (submitted on 24 Jun 2024, last revised on 9 Oct 2024)
RealTime QA	Dynamic question answering platform with weekly question announcements and evaluations	Questions about current world events and novel information	Answering questions about up-to-date information, challenging static assumptions in open-domain QA	Real-time evaluation results	GPT-3, T5	2024 (last revised on 28 Feb 2024)

Reefknot introduces a confidence-based mitigation strategy that quantitatively lowers the hallucination rate [30]. FreshQA leverages real-time augmentation via search engines, ensuring factual consistency and minimizing "hallucinatory" content when models confront fast-changing information [21]. HQH introduces a rigorous framework (HQM) to systematically evaluate the quality and reliability of hallucination benchmarks themselves, applying psychometric testing principles to ensure reproducible and valid evaluations [25].

By integrating reliability-focused strategies, these benchmarks refine not only the detection of hallucinations but also the trustworthiness of the tools used to measure them. As shown in Table 2, each benchmark brings distinct reliability improvements, offering a growing toolkit of methods to curb and understand hallucinations in increasingly complex tasks.

3.2 Comparative Analysis

While the thematic analysis provides a high-level overview of the strategic directions in hallucination benchmarks, a comparative lens helps clarify specific methodologies, dataset designs, and the challenges these benchmarks aim to resolve.

3.2.1 Methods. A broad range of methods is employed to detect and evaluate hallucinations. From simple scaling of model sizes to more sophisticated approaches like semantically equivalent prompt perturbation (SAC3) and search-augmented prompting (FreshQA), these methodologies reveal that there is no one-size-fits-all solution. Some approaches rely on fine-tuned models tailored to niche domains—e.g., BioGPT for biomedical text—while others evaluate general-purpose models like GPT-4 [12, 15]. Increasingly, benchmarks are also experimenting with cross-model comparisons, ensemble classifiers (ViBe), and dynamic evaluation platforms (RealTime QA) that track performance over time and across evolving scenarios [9, 17].

3.2.2 Dataset Types. Datasets underpin the entire hallucination evaluation ecosystem, and their construction and domain focus vary considerably. General-purpose datasets (e.g., FACTOR) enable broad comparisons of LLM factuality across Wiki, news, and expert domains, providing insights into scaling behaviors and retrieval augmentation [14]. Domain-specific datasets, such as those from DAHL or Reefknot, target specialized knowledge domains like biomedicine or multimodal relations [18, 30]. This trade-off highlights a key tension: broad generalization versus deep, domain-sensitive insights.

Other benchmarks, like HalluQA or HQH, integrate cultural, linguistic, or modality-specific elements into their datasets [4, 25]. This diversity ensures that evaluations can capture nuances that are often overlooked in more generic test sets. The choice of dataset profoundly influences what is measured and what constitutes "hallucination" in a given context.

3.2.3 Challenges. The complexity of hallucination evaluation surfaces several challenges. Scalability is a persistent issue, with large-scale datasets (ViBe, HQH) demanding efficient computational resources and annotation strategies [17, 25]. Interpretability is equally pressing—many benchmarks strive to identify not just whether a hallucination occurs, but why and how it manifests. The lack of transparency in some large language models, the cost of human annotation, and the difficulty of benchmarking rapidly evolving world knowledge (RealTime QA, FreshQA) create obstacles that the field is only beginning to address [9, 21].

Benchmarks like SAC3 attempt to enhance interpretability through cross-model and cross-prompt consistency checks, while FreshQA's real-time augmentation strategies exemplify attempts to keep pace with evolving factual landscapes [21, 27]. Moreover, concerns about energy consumption and environmental impacts also loom large, raising questions about how to sustainably conduct large-scale evaluations without trading off rigor or comprehensiveness.

4 CRITICAL ANALYSIS

4.1 Evaluation of Research Quality

The methodologies analyzed are robust and scalable, leveraging advanced techniques such as semantically equivalent prompt perturbations and confidence-based mitigation strategies. However, they are often limited by high computational costs, which can hinder widespread adoption and scalability. Benchmarks like SAC3 and FreshQA demonstrate strong detection and mitigation capabilities, yet they require significant computational resources for training and evaluation [21, 27]. Additionally, the reliance on large datasets and complex model architectures may pose barriers for smaller research groups or organizations with limited resources.

4.2 Insights and Discussion

The two tables (Table 2 for thematic analysis and Table 3 for comparative analysis) collectively provide a structured view of the current benchmark landscape. By examining them side-by-side, we derive several key insights:

4.2.1 Increasing Complexity and Specialization. The progression from simple factual checks to complex, multi-step reasoning and multimodal evaluations indicates that the field is maturing. Early benchmarks focused primarily on detecting factual inconsistencies, while recent contributions delve into logical coherence, temporal reasoning, and cultural awareness. This increasing complexity reflects the growing demands for LLMs to perform more sophisticated tasks reliably.

4.2.2 Growing Importance of Domain Context. As specialized benchmarks (e.g., DAHL, Reefknot, HalluQA) show, hallucination evaluation cannot be one-size-fits-all. Different domains present unique challenges—biomedical knowledge must be precise, cultural contexts must be respected, and multimodal tasks must ensure alignment between text and visuals. This domain sensitivity is essential for deploying LLMs responsibly in varied real-world settings.

4.2.3 Methodological Innovation and Diversity. The data reveals a wide range of evaluation strategies, including semantically equivalent prompt perturbations, real-time retrieval augmentation, cross-model comparisons, ensemble classifiers (ViBe), and psychometrically grounded assessments (HQH). This methodological heterogeneity signals that researchers are actively experimenting with new ways to break open the "black box" of LLM reasoning, making the evaluation process more transparent and reliable.

4.2.4 Emphasis on Reliability and Trustworthiness. Many benchmarks are now going beyond identifying hallucinations to mitigating them and improving the reliability of evaluations. Techniques like confidence scoring, real-time information retrieval, and robust test-retest methodologies illustrate a growing focus on trustworthiness. Improved reliability metrics and framework designs are helping ensure that future LLM evaluations are both valid and reproducible.

4.2.5 Persistent Challenges and Future Directions. Despite these advances, challenges remain. Scalability, annotation costs, dynamic world knowledge, and interpretability are unresolved issues that the data in these tables brings into sharp relief. Addressing these will likely involve collaborative efforts across AI subfields, from efficient model training to human-in-the-loop annotation strategies and ethical frameworks that guide model deployment.

4.3 Identification of Gaps

Despite the advancements, several research gaps persist:

- **Limited Interpretability of LLMs**: Understanding why and how hallucinations occur remains a challenge. Current benchmarks focus on detection and mitigation without fully unraveling the underlying causes of hallucinations.
- **Energy Inefficiency During Training**: The high computational demands of training and evaluating large models contribute to significant energy consumption, raising sustainability concerns.
- **Cultural and Linguistic Diversity**: While benchmarks like HalluQA address Chinese linguistic contexts, there is a need for more diverse benchmarks that cover a wider range of languages and cultural contexts.
- **Dynamic Knowledge Integration**: Existing benchmarks struggle to keep up with the rapidly changing world knowledge, necessitating more dynamic and real-time evaluation frameworks.

4.4 Implications

LLMs have profound practical implications across various sectors, including healthcare, education, and automation. In healthcare, accurate and reliable information is critical, and hallucinations can lead to misinformation with potentially severe consequences [12]. In education, LLMs can serve as tutors or information providers, where inaccuracies can mislead learners. Automation tasks relying on LLMs for decision-making or content generation must ensure high factual accuracy to maintain trust and efficacy. The findings of this survey call for ethical considerations in deploying LLMs, emphasizing the need for transparency, accountability, and continuous monitoring to mitigate the risks associated with hallucinations.

4.5 Limitations

This survey primarily focuses on recent publications, which may overlook foundational studies that have significantly contributed to the field. Additionally, the computational challenges associated with large-scale evaluations restrict the scalability of some benchmarks. The reliance on specific datasets and model architectures may limit the generalizability of the findings. Future research should aim to address these limitations by incorporating a broader range of studies and exploring more efficient evaluation methodologies.

5 CONCLUSION

5.1 Summary of Findings

This survey highlights key advancements in LLMs, addressing reasoning, reliability, and domain-specific capabilities. Thematic analysis underscores the evolution from basic factual checks to sophisticated reasoning and multimodal evaluations, while comparative analysis elucidates the diverse methodologies and dataset types employed in current research. Benchmarks like DAHL, Reefknot, and HQH exemplify the tailored approaches necessary for different domains and reliability enhancements. Overall, the field is progressing towards more nuanced and context-aware evaluation frameworks that aim to enhance the trustworthiness and applicability of LLMs across various real-world scenarios.

5.2 Future Directions

Future work should focus on:

- **Hybrid Models Integrating Structured and Unstructured Data**: Developing models that can seamlessly combine structured databases with unstructured text to enhance factual accuracy and reduce hallucinations.
- **Enhancing Energy Efficiency in LLM Training and Inference**: Innovating more energy-efficient algorithms and model architectures to mitigate the environmental impact of large-scale LLM deployments.
- **Expanding Cultural and Linguistic Benchmarks**: Creating more diverse benchmarks that encompass a broader range of languages and cultural contexts to ensure global applicability and fairness.
- **Real-Time and Adaptive Evaluation Frameworks**: Designing evaluation systems that can adapt to evolving knowledge bases and provide real-time assessments of LLM performance.
- **Improving Interpretability and Explainability**: Enhancing methods to understand and explain the internal mechanisms of LLMs, facilitating better identification and mitigation of hallucinations.

6 GROUP REFLECTION

6.1 Learning Experience

The project enhanced our understanding of literature review methodologies and advancements in Large Language Models (LLMs). Through comprehensive analysis and synthesis of various benchmarks and methodologies, we gained deeper insights

Table 4: Work Distribution

Team Member	Tasks
Yash Khare	(1) Conducted literature search and analysis (2) Documented findings in the report & presentation
Mohan Bhosale	(1) Conducted literature search and analysis (2) Documented findings in the report & presentation
Karthikeyan Sugavanan	(1) Documented findings in the report & presentation

into the complexities and challenges associated with hallucination evaluation in LLMs and Multimodal Large Language Models (LMMs).

6.2 Challenges

Key challenges included identifying relevant papers amidst a rapidly growing body of literature and synthesizing diverse findings into coherent themes. Additionally, ensuring the consistency and accuracy of information across multiple sources required meticulous attention to detail and collaborative effort.

6.3 Work Distribution

Each group member contributed equally to literature analysis, writing, and presentation preparation. The distribution of tasks ensured that all aspects of the project were thoroughly covered and that each member developed a comprehensive understanding of the subject matter.

REFERENCES

[1] S. Banerjee, A. Agarwal, and S. Singla. Llms will always hallucinate, and we need to live with this, 2024.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

[4] Q. Cheng, T. Sun, W. Zhang, S. Wang, X. Liu, M. Zhang, J. He, M. Huang, Z. Yin, K. Chen, and X. Qiu. Evaluating hallucinations in chinese large language models, 2023.

[5] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

[6] R. Friel and A. Sanyal. Chainpoll: A high efficacy method for llm hallucination detection, 2023.

[7] Z. Guo, P. Wang, Y. Wang, and S. Yu. Improving small language models on pubmedqa via generative data augmentation, 2023.

[8] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, Nov. 2024.

[9] J. Kasai, K. Sakaguchi, Y. Takahashi, R. L. Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, and K. Inui. Realtime qa: What's the answer right now?, 2024.

[10] R. Krishnan, P. Khanna, and O. Tickoo. Enhancing trust in large language models with uncertainty-aware fine-tuning, 2024.

[11] J. Li, J. Chen, R. Ren, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models, 2024.

[12] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), Sept. 2022.

[13] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization, 2020.

[14] D. Muhlgay, O. Ram, I. Magar, Y. Levine, N. Ratner, Y. Belinkov, O. Abend, K. Leyton-Brown, A. Shashua, and Y. Shoham. Generating benchmarks for factuality evaluation of language models, 2024.

[15] OpenAI. Gpt-4 technical report, 2024.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[17] V. Rawte, S. Jain, A. Sinha, G. Kaushik, A. Bansal, P. R. Vishwanath, S. R. Jain, A. N. Reganti, V. Jain, A. Chadha, A. P. Sheth, and A. Das. Vibe: A text-to-video benchmark for evaluating hallucination in large multimodal models, 2024.

[18] J. Seo, J. Lim, D. Jang, and H. Shin. Dahl: Domain-specific automated hallucination evaluation of long-form text through a benchmark dataset in biomedicine, 2024.

[19] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize from human feedback, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[21] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, and T. Luong. Freshllms: Refreshing large language models with search engine augmentation, 2023.

[22] X. Wang, J. Pan, L. Ding, and C. Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding, 2024.

[23] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. Georgiev, R. J. Das, and P. Nakov. Factuality of large language models: A survey, 2024.

[24] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge conflicts for llms: A survey, 2024.

[25] B. Yan, J. Zhang, Z. Yuan, S. Shan, and X. Chen. Evaluating the quality of hallucination benchmarks for large vision-language models, 2024.

[26] Z. Yuan, Y. Shang, Y. Zhou, Z. Dong, Z. Zhou, C. Xue, B. Wu, Z. Li, Q. Gu, Y. J. Lee, Y. Yan, B. Chen, G. Sun, and K. Keutzer. Llm inference unveiled: Survey and roofline model insights, 2024.

[27] J. Zhang, Z. Li, K. Das, B. A. Malin, and S. Kumar. Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency, 2024.

[28] L. Zhao, K. Nguyen, and H. III. Hallucination detection for grounded instruction generation. pages 4044–4053, 01 2023.

[29] D. Zheng, M. Lapata, and J. Z. Pan. Large language models as reliable knowledge bases?, 2024.

[30] K. Zheng, J. Chen, Y. Yan, X. Zou, and X. Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models, 2024.

[31] Y. Zhou, H. Fan, S. Gao, Y. Yang, X. Zhang, J. Li, and Y. Guo. Retrieval and localization with observation constraints, 2021.