© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717980

A Multi-Metric Evaluation Perspective on Hallucination Detection in Low-Resource Governance Documents

PRANJAL GAHLOT[1], RAKSHITHA B S[2]
[1]Scholar, Department of Computer Science & IT, Jain (Deemed to be University), Bangalore, India
[2]Assistant Professor, Department of Computer Science & IT, Jain (Deemed-To-Be-University), Bangalore, India

Abstract- ***The rapid advancement of Large Language Models (LLMs) has significantly improved natural language processing applications across domains such as governance, healthcare, legal analysis, and public information systems. Despite these advancements, LLMs frequently generate hallucinated outputs, where responses appear plausible but contain incorrect or fabricated information. This issue poses serious risks in governance-related applications, where inaccurate information can influence policy interpretation, administrative decision-making, and public trust. Existing studies have proposed several approaches to address hallucinations, including semantic entropy–based detection, benchmark evaluation frameworks, and adversarial testing methods. However, the literature indicates that current solutions remain fragmented and often focus on isolated aspects such as model performance, dataset construction, or benchmark capability rather than comprehensive reliability assessment. This literature review examines recent research on hallucination detection, multilingual and low-resource natural language processing, and evaluation frameworks for LLM reliability. The reviewed studies highlight key challenges, including the lack of multilingual hallucination evaluation, insufficient harm-oriented risk assessment, and limited adversarial robustness testing in governance contexts. Furthermore, existing benchmarks often measure task accuracy rather than factual reliability or societal impact. Based on the analysis of the literature, this review identifies major methodological and contextual gaps and proposes the need for an integrated evaluation framework combining meaning level hallucination detection, harm aware risk modeling, and multilingual robustness assessment. Such an approach could improve the reliability and safety of LLM systems deployed in governance and public service environments.***

Keywords - Large Language Models, Hallucination Detection, Semantic Entropy, Multilingual NLP, Low-Resource Languages, Governance AI, Adversarial Prompting, Benchmark Evaluation, AI Reliability, Natural Language Processing.

I. INTRODUCTION

Recent advancements in Large Language Models (LLMs) have significantly transformed natural language processing applications across multiple domains, including governance, healthcare, legal analysis, and public service systems. These models demonstrate strong capabilities in tasks such as summarization, question answering, translation, and policy analysis. However, a major challenge that continues to affect the reliability of LLMs is the phenomenon of hallucination, where models generate factually incorrect or fabricated information while appearing confident in their responses. Studies such as Farquhar et al. (2024) highlight that hallucinations often occur at the meaning level, making them difficult to detect using traditional token-level confidence measures. Similarly, research by Huang et al. (2025) and Ji et al. (2023) shows that hallucinations remain a persistent issue across natural language generation tasks, raising concerns regarding the factual reliability of AI-generated content.

This problem becomes more critical in governance and public decision-support systems, where inaccurate information may influence policy interpretation, legal decisions, or public trust. Additionally, the issue is amplified in multilingual and low-resource language environments, where limited datasets and weaker model grounding increase the likelihood of erroneous outputs. Studies on low-resource NLP, such as Pakray et al. (2025) and Zhang et al. (2024), demonstrate that transfer learning and multilingual transformers improve performance but do not adequately address hallucination robustness. Furthermore, benchmarking frameworks such as LexGLUE (Chalkidis et al., 2022) and MEGA (Ahuja et al., 2023) primarily

evaluate task performance rather than factual reliability or misinformation risk.

The literature reviewed in this study consists primarily of peer reviewed journal articles and conference papers published between 2022 and 2026, focusing on three major research themes: hallucination detection in LLMs, multilingual and low-resource NLP systems, and evaluation frameworks for AI reliability. These studies employ a range of methodologies including systematic literature reviews, benchmark development, adversarial testing, and experimental model evaluations. Despite differences in approach, the sources share a common goal of improving LLM reliability, evaluation methods, and real-world deployment safety.

While existing research has proposed several techniques to detect hallucinations, such as semantic entropy (Farquhar et al., 2024) and QA-based verification frameworks (Liu et al., 2026), significant gaps remain. Current studies rarely integrate multilingual evaluation, harm-oriented risk assessment, and adversarial robustness testing within governance contexts. Moreover, most benchmarks measure model capability rather than the severity or societal impact of hallucinated outputs, which limits their usefulness in critical decision-making environments. Therefore, the central research question guiding this review is: How can hallucination detection and evaluation frameworks be improved to ensure reliable and safe deployment of multilingual LLM systems in governance applications?

This literature review argues that although substantial progress has been made in hallucination detection and multilingual NLP research, there is still no unified framework that integrates meaning-level hallucination detection, harm-aware evaluation, and multilingual robustness for governance AI systems. By analyzing existing studies across these domains, this review identifies key methodological and contextual gaps that hinder the development of reliable AI-assisted governance tools.

The remainder of this literature review is organized as follows. First, it examines existing research on LLM hallucination detection and classification frameworks. Second, it reviews studies on multilingual and low-resource NLP approaches and their implications for AI reliability. Third, it discusses evaluation benchmarks and adversarial testing methods used to assess LLM performance. Finally, the review synthesizes the findings to highlight major research gaps and outline potential directions for future research.

In summary, understanding the limitations of current hallucination detection methods and evaluation frameworks is essential for developing trustworthy multilingual AI systems, particularly in governance contexts where accuracy, transparency, and accountability are critical.

1. Hallucination in Large Language Models

Hallucination in large language models has emerged as a critical research challenge in natural language generation systems. Hallucinations occur when models generate outputs that appear coherent but contain fabricated or incorrect information. Early work in this area focused on identifying the causes and classifications of hallucinations across different natural language generation tasks. Ji et al. (2023) provided one of the earliest comprehensive surveys on hallucination in natural language generation, proposing a taxonomy that distinguishes between intrinsic hallucinations, which arise from inconsistencies within the input data, and extrinsic hallucinations, which occur when the model introduces unsupported factual information. This framework established a foundational understanding of hallucination behavior and guided subsequent research in the field.

More recent studies have expanded this taxonomy to address challenges specific to large language models. Huang et al. (2025) refined hallucination classifications by distinguishing between factuality hallucinations and faithfulness hallucinations, reflecting the complexities of modern transformer-based models. Their review highlights that hallucinations often arise not only from incorrect knowledge but also from failures in instruction following or contextual understanding. While these studies provide a conceptual foundation for understanding hallucinations, they largely focus on

theoretical classification rather than practical mitigation strategies. Consequently, further research is required to develop reliable detection frameworks that can operate effectively across different domains and languages.

Building on these taxonomies, recent work has begun to explore methods for detecting hallucinations at the semantic level rather than relying solely on token-level probability metrics. Farquhar et al. (2024) introduced the concept of semantic entropy, which measures uncertainty by clustering semantically equivalent model outputs using natural language inference techniques. This method demonstrated improved performance in detecting model confabulations compared to traditional confidence-based approaches. However, the evaluation of this approach was limited to question-answering datasets, leaving its applicability in governance-related applications largely unexplored.

II. DETECTION AND EVALUATION FRAMEWORKS FOR HALLUCINATION

Beyond theoretical classification, researchers have proposed several frameworks to detect and evaluate hallucinations in LLM-generated outputs. One notable approach is the QA-based verification framework proposed by Liu et al. (2026) for evaluating factual consistency in summarization tasks. This method decomposes generated summaries into atomic claims and verifies them through question-answering models against the source document. By validating individual claims, the framework enables transparent identification of factual inconsistencies without requiring reference summaries. Although this approach has shown promising results in detecting unfaithful summaries, its application remains largely limited to summarization datasets.

Another significant development in hallucination evaluation is the introduction of harm-oriented assessment frameworks. Asgari et al. (2025) proposed a clinical safety framework that categorizes hallucinations based on their potential impact on patient safety. Their approach combines automated detection methods with expert annotations to classify errors according to severity levels. This work highlights the importance of evaluating hallucinations not only in terms of frequency but also in terms of the potential harm they may cause in real-world settings. While this framework provides valuable insights for healthcare applications, similar harm-oriented evaluation models have not yet been widely applied to governance or public policy contexts.

In addition to evaluation frameworks, researchers have also examined the vulnerability of LLMs to adversarial manipulation. Omar et al. (2025) conducted adversarial stress testing on clinical LLMs by inserting fabricated information into prompts. Their results revealed hallucination rates of up to 83 percent when models were exposed to adversarial prompts. This finding underscores the importance of robust testing frameworks to ensure that LLM systems remain reliable when deployed in high-stakes environments.

III. MULTILINGUAL AND LOW-RESOURCE NATURAL LANGUAGE PROCESSING

Another critical dimension of LLM reliability concerns the performance of models in multilingual and low-resource language environments. While many NLP systems are trained primarily on high-resource languages such as English, the majority of the world's languages remain underrepresented in training datasets. Pakray et al. (2025) highlight that limited data availability continues to be a major barrier to effective NLP development in low-resource languages. Their survey emphasizes the role of transfer learning and multilingual transformer models, such as mBERT and XLM-R, in improving performance across languages with limited datasets.

Research on multilingual sentiment analysis and classification tasks further illustrates the challenges associated with low-resource languages. Roy (2024) proposed an ensemble transformer approach for bilingual sentiment analysis in Kannada and Malayalam. The results demonstrated that combining multiple transformer models can improve classification accuracy in underrepresented languages. However, these studies primarily focus on task performance rather than the reliability of generated information. Consequently, little attention has been given to the relationship between multilingual modeling and hallucination robustness.

Similarly, research in low-resource machine translation has focused on improving translation quality through techniques such as data augmentation and transfer learning. Zhang et al. (2024) conducted a comprehensive survey of neural machine translation approaches for low-resource languages, concluding that transformer-based architectures significantly outperform earlier statistical machine translation methods. Nevertheless, evaluation metrics in machine translation studies typically rely on measures such as BLEU scores, which capture linguistic similarity but do not adequately assess factual accuracy or hallucination risk.

IV. BENCHMARKING AND EVALUATION OF LLM SYSTEMS

Benchmarking plays an essential role in evaluating the capabilities of large language models. Several benchmark datasets have been developed to measure model performance across different tasks and domains. For example, Chalkidis et al. (2022) introduced LexGLUE, a benchmark suite designed specifically for legal natural language processing tasks. The benchmark includes multiple datasets covering tasks such as legal document classification and case outcome prediction. Although LexGLUE provides a standardized evaluation framework for legal NLP research, its evaluation metrics primarily focus on accuracy and F1 scores rather than factual reliability.

More recent studies have expanded benchmarking efforts to include multilingual evaluation frameworks. Ahuja et al. (2023) proposed the MEGA benchmark, which evaluates language models across seventy languages using sixteen different NLP datasets. This framework provides valuable insights into the cross-lingual capabilities of large language models. However, the benchmark still primarily measures task performance rather than the reliability or trustworthiness of generated outputs.

Similarly, Bang et al. (2023) conducted a large-scale evaluation of ChatGPT across twenty-three NLP datasets to assess reasoning and language understanding capabilities. Their findings suggest that benchmark scores can provide a useful proxy for model capability, but they may not accurately reflect real-world reliability. These results highlight the limitations of existing evaluation frameworks and suggest that additional metrics are required to assess the factual integrity and safety of LLM-generated information.

V. EMERGING MULTILINGUAL LARGE LANGUAGE MODELS

The development of multilingual instruction-tuned models has further expanded the capabilities of LLM systems. Üstün et al. (2024) introduced Aya, an open-access multilingual language model trained on instruction datasets covering over one hundred languages. The study demonstrated that large-scale instruction tuning can significantly improve cross-lingual generalization and task performance. Such models have the potential to support multilingual governance systems and public service applications.
Despite these advancements, the evaluation of multilingual LLMs continues to focus primarily on performance metrics rather than hallucination risk or harm severity. As a result, there remains a significant gap in understanding how these models behave when deployed in high-stakes environments such as governance, healthcare, or legal systems.

Research Gap

A critical analysis of the existing literature on Large Language Models (LLMs), hallucination detection, multilingual NLP, and evaluation frameworks reveals several unresolved issues that limit the reliability and real-world deployment of these systems. Although previous studies have explored different aspects of LLM performance, they tend to address these aspects in isolation. This fragmentation creates significant research gaps, particularly in areas where reliability, multilingual capability, and risk assessment intersect. The following subsections describe the major research gaps identified from the literature.

1. Lack of Integrated Hallucination Detection Frameworks

One of the most prominent gaps in the literature is the absence of comprehensive frameworks that integrate multiple hallucination detection approaches. Existing studies mainly propose individual techniques such as semantic entropy based detection, question-answering verification frameworks, or

taxonomy-based classification methods. While these approaches contribute valuable insights, they are typically designed for specific tasks such as summarization, question answering, or dialogue generation.

As a result, there is limited research examining how these methods can be combined to provide robust, domain-independent hallucination detection systems. Furthermore, most frameworks operate at a single level of analysis, such as token level probability or semantic similarity, without integrating multiple verification mechanisms. This creates a gap in developing multi-layered evaluation systems capable of detecting hallucinations more reliably across different applications.

2. Insufficient Multilingual Evaluation of Hallucination Behavior

Another major research gap concerns the limited exploration of hallucination behavior in multilingual and low resource language settings. Although multilingual models such as transformer-based architectures have improved performance across multiple languages, the majority of hallucination detection research has been conducted using Englishlanguage datasets.

This creates a significant limitation, particularly for languages with limited training data. In multilingual environments, differences in linguistic structure, cultural context, and dataset availability may influence how hallucinations occur and how effectively they can be detected. However, current studies rarely evaluate whether hallucination detection methods developed for English can generalize effectively to low resource languages such as Kannada, Malayalam, or other regional languages. Consequently, there is a need for research that examines cross lingual hallucination patterns and detection strategies, particularly in multilingual systems designed for global or governance-related applications.

3. Limited Harm-Oriented Evaluation of Hallucinations

Most existing research evaluates hallucinations based on accuracy metrics, factual consistency scores, or benchmark performance indicators. While these metrics provide useful technical insights, they do not fully capture the real-world consequences of hallucinated information.

For instance, hallucinated outputs in domains such as healthcare, law, or public policy may lead to severe consequences if incorrect information influences decision making processes. Although some studies have introduced risk based evaluation frameworks, such approaches remain relatively rare and are typically limited to specific domains.

This indicates a gap in the development of harm-oriented evaluation frameworks that assess not only whether hallucinations occur but also the severity and potential societal impact of those hallucinations. Addressing this gap would allow researchers to prioritize mitigation strategies based on real-world risk levels.

4. Weak Connection Between Benchmark Performance and Reliability

Benchmark datasets are widely used to measure the capabilities of language models across different tasks. However, a growing body of literature indicates that high benchmark scores do not necessarily translate to reliable real-world performance.

Many benchmarks focus on task completion metrics such as accuracy, BLEU scores, or F1 scores. These metrics evaluate whether a model produces outputs that match reference answers but do not measure whether the generated information is factually correct or trustworthy. Consequently, models that perform well on benchmarks may still generate hallucinated or misleading outputs in practical applications.

This reveals a significant gap in the development of evaluation metrics that directly measure factual reliability and trustworthiness rather than solely focusing on task accuracy.

5. Limited Research in Governance and Public Policy Contexts

Another important research gap concerns the lack of domain-specific evaluation of LLM reliability in governance related applications. Many studies focus on domains such as machine translation,

summarization, or healthcare, where the datasets and evaluation metrics are relatively well established.

However, the application of LLMs in governance, policy analysis, and public administration introduces unique challenges. These systems may generate information that influences policy interpretation, public communication, or administrative decision-making. In such contexts, hallucinated information could have serious implications for public trust, policy implementation, and institutional accountability.

Despite these risks, relatively few studies have examined how hallucination detection frameworks perform in governance related tasks or how LLMs can be evaluated for policy-sensitive applications.

6. Vulnerability of LLMs to Adversarial Prompt Manipulation

Recent research has demonstrated that LLMs can be highly vulnerable to adversarial prompt manipulation, where malicious or misleading information is intentionally introduced into prompts to influence model outputs. Adversarial testing studies reveal that even advanced models may generate fabricated information when exposed to carefully crafted prompts.

However, current research primarily focuses on identifying these vulnerabilities rather than developing robust defense mechanisms or evaluation frameworks to mitigate adversarial hallucinations. Moreover, adversarial robustness testing has not yet been widely integrated into standard LLM evaluation pipelines.

This indicates a need for future research that systematically examines how adversarial prompts influence hallucination behavior and how detection mechanisms can be strengthened to resist such attacks.

Overall Research Gap

In summary, the literature reveals several interconnected gaps that limit the reliability of large language models in practical applications. Existing research provides valuable contributions in areas such as hallucination detection techniques, multilingual modeling, and benchmark evaluation. However, these areas remain largely disconnected.

There is therefore a strong need for research that develops an integrated framework combining multilingual evaluation, semantic level hallucination detection, adversarial robustness testing, and harm-oriented risk assessment. Such a framework would enable more comprehensive evaluation of LLM reliability and support the safe deployment of AI systems in high-stakes domains such as governance, healthcare, and public administration.

Figure 1. Research gaps in multilingual hallucination detection and governance AI evaluation.

System Architecture of the Proposed Framework

The proposed system architecture is designed as a multi layered evaluation framework for detecting and assessing hallucinations generated by Large Language Models (LLMs) in multilingual governance related environments. The architecture integrates semantic verification, multilingual robustness evaluation, adversarial testing, and harm oriented assessment into a unified pipeline to improve the reliability and trustworthiness of AI-generated outputs.

The architecture begins with the input layer, where governance documents, policy texts, legal records, multilingual datasets, or user queries are provided to the system. These inputs may include low resource languages such as Kannada and Malayalam in addition to high resource languages like English. The input data is processed by a multilingual large language model (LLM) responsible for generating responses such as summaries, translations, policy interpretations, or question-answering outputs.

The generated response is then forwarded to the semantic verification layer, which evaluates factual consistency between the generated output and the original source document. This layer uses semantic

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717980

similarity analysis and contradiction detection techniques to identify meaning-level inconsistencies that may not be detectable through traditional token-level confidence methods.

Following semantic verification, the response is analyzed by the hallucination detection module, which identifies fabricated information, unsupported claims, contextual deviations, and misleading outputs. The detected hallucinations are further processed by the hallucination classification layer, where they are categorized into multiple types including factual hallucinations, legal hallucinations, contextual hallucinations, procedural hallucinations, and multilingual translation-based hallucinations.

After classification, the framework applies a harm severity assessment layer that evaluates the potential societal and governance-related impact of the hallucinated information. This module assigns severity levels such as low, medium, high, or critical depending on the possible consequences of the generated misinformation in governance and public administration contexts.

To improve robustness evaluation, the architecture also incorporates an adversarial testing module that intentionally introduces manipulated prompts and misleading contextual inputs to assess the stability and reliability of the LLM under adversarial conditions. This stage helps identify vulnerabilities in multilingual governance AI systems and measures resistance against prompt based hallucination attacks. Finally, outputs from all evaluation layers are integrated into a final reliability scoring module, which computes an overall trustworthiness score for the generated response. This score reflects semantic consistency, multilingual robustness, adversarial resistance, and harm severity. The proposed architecture therefore enables comprehensive evaluation of LLM reliability and supports the safe deployment of multilingual AI systems in governance, legal, and public service applications.

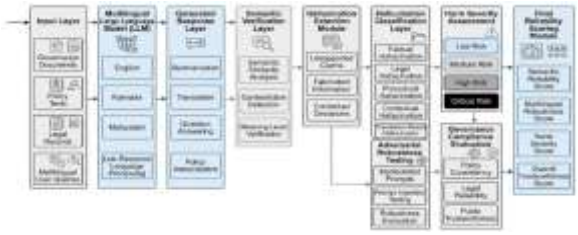

Figure 2: Proposed multi-metric hallucination evaluation architecture for multilingual governance AI systems.

Working Mechanism of the Proposed Framework

The proposed framework operates through a multi stage evaluation pipeline designed to detect, classify, and assess hallucinations generated by Large Language Models (LLMs) in multilingual governance related environments. Initially, governance documents, policy texts, or user queries are provided as input to a multilingual LLM capable of processing low resource languages such as Kannada and Malayalam alongside high-resource languages like English. The model generates responses in the form of summaries, policy interpretations, question-answering outputs, or translated content. Since hallucinated information may appear semantically correct while being factually inaccurate, the generated response is passed through a semantic verification layer that compares the response with the original source document using semantic similarity and contradiction detection techniques.

Following semantic verification, the framework applies a hallucination detection module that identifies inconsistencies, fabricated claims, unsupported facts, or contextual deviations in the generated output. The detected hallucinations are then categorized into multiple classes, including factual hallucinations, contextual hallucinations, procedural hallucinations, legal hallucinations, and multilingual translation-based hallucinations. This classification process enables the system to analyze not only whether hallucinations occur but also the nature of the generated misinformation.

After hallucination classification, the framework performs a harm-oriented risk assessment to evaluate the potential societal and governance impact of the hallucinated output. In this stage, hallucinations are

assigned severity levels such as low, medium, high, or critical depending on their possible consequences in governance and public administration contexts. For example, minor linguistic inconsistencies may be categorized as low risk, whereas fabricated legal or policy-related information may be classified as critical risk due to its potential influence on administrative decision-making and public trust.

To further improve robustness evaluation, the framework incorporates an adversarial testing mechanism in which manipulated or misleading prompts are intentionally introduced to examine the stability of the language model under adversarial conditions. This stage helps identify vulnerabilities in multilingual governance systems and measures the resistance of the model against prompt based hallucination attacks. Finally, outputs from semantic verification, multilingual robustness evaluation, harm assessment, and adversarial testing are integrated to generate an overall reliability score representing the trustworthiness and safety of the generated response. This multi-metric evaluation approach enables comprehensive assessment of LLM reliability and supports safer deployment of AI systems in governance and public service applications.

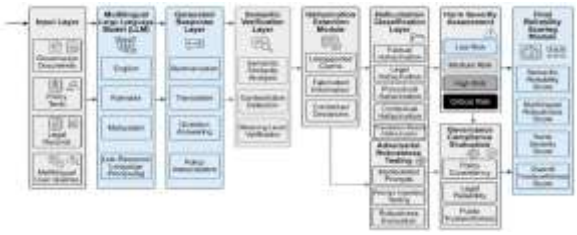

Figure 2: Proposed multi-metric hallucination evaluation architecture for multilingual governance AI systems.

Working Mechanism of the Proposed Framework

The proposed framework operates through a multi stage evaluation pipeline designed to detect, classify, and assess hallucinations generated by Large Language Models (LLMs) in multilingual governance related environments. Initially, governance documents, policy texts, or user queries are provided as input to a multilingual LLM capable of processing low resource languages such as Kannada and Malayalam alongside high-resource languages like English. The model generates responses in the form of summaries, policy interpretations, question-answering outputs, or translated content. Since hallucinated information may appear semantically correct while being factually inaccurate, the generated response is passed through a semantic verification layer that compares the response with the original source document using semantic similarity and contradiction detection techniques.

Following semantic verification, the framework applies a hallucination detection module that identifies inconsistencies, fabricated claims, unsupported facts, or contextual deviations in the generated output. The detected hallucinations are then categorized into multiple classes, including factual hallucinations, contextual hallucinations, procedural hallucinations, legal hallucinations, and multilingual translation-based hallucinations. This classification process enables the system to analyze not only whether hallucinations occur but also the nature of the generated misinformation.

After hallucination classification, the framework performs a harm-oriented risk assessment to evaluate the potential societal and governance impact of the hallucinated output. In this stage, hallucinations are assigned severity levels such as low, medium, high, or critical depending on their possible consequences in governance and public administration contexts. For example, minor linguistic inconsistencies may be categorized as low risk, whereas fabricated legal or policy-related information may be classified as critical risk due to its potential influence on administrative decision-making and public trust.

To further improve robustness evaluation, the framework incorporates an adversarial testing mechanism in which manipulated or misleading prompts are intentionally introduced to examine the stability of the language model under adversarial conditions. This stage helps identify vulnerabilities in multilingual governance systems and measures the resistance of the model against prompt based hallucination attacks. Finally, outputs from semantic verification, multilingual robustness evaluation, harm assessment, and adversarial testing are integrated to generate an overall reliability score representing the trustworthiness and safety of the generated response.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717980

This multi-metric evaluation approach enables comprehensive assessment of LLM reliability and supports safer deployment of AI systems in governance and public service applications.

Figure 3. Workflow of the proposed multi-metric hallucination evaluation framework.

VI. CONCLUSION

Brief Summary of Findings

The reviewed literature highlights significant progress in the development and evaluation of Large Language Models (LLMs), particularly in areas such as hallucination detection, multilingual natural language processing, and benchmarking frameworks. Studies focusing on hallucination behavior have provided valuable conceptual foundations by classifying hallucinations and proposing detection mechanisms. For instance, research on hallucination taxonomy has helped identify the underlying causes of inaccurate model outputs, while semantic-based detection methods have demonstrated promising results in identifying unreliable responses. Additionally, evaluation frameworks developed for domains such as healthcare and summarization have emphasized the importance of assessing factual consistency and reliability rather than relying solely on traditional performance metrics.

Another major theme emerging from the literature is the growing importance of multilingual and low-resource language processing. Research shows that many languages remain underrepresented in NLP datasets, leading to disparities in model performance. Studies on multilingual transformers and transfer learning demonstrate that cross-lingual approaches can significantly improve performance in low resource settings. However, these studies primarily focus on improving task accuracy and rarely address issues related to hallucination or factual reliability in multilingual contexts.

A third key theme concerns the benchmarking and evaluation of LLM capabilities. Existing benchmarks have been widely used to measure reasoning ability, language understanding, and task performance. While these benchmarks provide important insights into model capabilities, they often fail to measure the reliability or safety of generated outputs. As a result, there remains a disconnect between benchmark performance and real-world trustworthiness, especially in high-stakes domains such as governance, healthcare, and legal systems.

VII. LIMITATIONS AND FUTURE DIRECTIONS

Despite the significant contributions of existing studies, several limitations remain evident in the current body of literature. One major limitation is that many studies examine hallucination detection methods in isolated experimental settings, such as question-answering or summarization tasks. This narrow focus limits the generalizability of findings to other real-world applications where LLMs interact with complex and diverse data sources.

Another limitation is the lack of multilingual evaluation for hallucination detection techniques. While multilingual NLP research has improved performance across different languages, most hallucination detection frameworks have been developed and tested primarily in English-language environments. This creates a research gap in understanding how hallucinations manifest in low-resource languages and whether existing detection approaches remain effective across linguistic contexts.

Furthermore, current benchmarking frameworks largely emphasize model accuracy and task performance rather than the societal impact or risk associated with incorrect outputs. As LLMs are increasingly integrated into governance systems, policy analysis tools, and public information platforms, the absence of harm-oriented evaluation metrics becomes a critical concern.

Future research should therefore focus on developing integrated evaluation frameworks that combine

hallucination detection, multilingual robustness testing, and risk-based assessment. Such frameworks would enable researchers and practitioners to evaluate LLM reliability more comprehensively. Additionally, expanding datasets and benchmarks to include low-resource languages and governance-related scenarios would improve the practical applicability of LLM research.

VIII. IMPLICATIONS OF THE THESIS AND FINAL ANALYSIS

Taken together, the reviewed studies demonstrate that while LLM technology has achieved remarkable progress in language understanding and generation, challenges related to reliability, factual accuracy, and multilingual robustness remain unresolved. The findings suggest that existing research approaches often address individual aspects of the problem rather than developing comprehensive solutions.

This literature review contributes to the field by synthesizing research from multiple domains, including hallucination detection, multilingual NLP, and benchmark evaluation. By identifying the intersections and gaps between these research areas, the analysis highlights the need for a holistic framework for evaluating LLM reliability, particularly in governance related applications.

The implications of this analysis are significant for researchers and practitioners working in artificial intelligence and natural language processing. Scholars can use these findings to design new evaluation methodologies that incorporate meaning level verification, multilingual assessment, and harm-based risk modeling. Such approaches will be essential for ensuring that LLM systems can be deployed safely and responsibly in critical decision-making environments.

In conclusion, addressing the identified research gaps will not only improve the technical reliability of large language models but will also enhance trust, accountability, and transparency in AI-driven systems used in public and institutional context.

IX. DECLARATION OF AI ASSISTANCE

AI tools were used only for grammar improvement, formatting, and structural guidance during the preparation of this manuscript. All research analysis, interpretations, and conclusions are the original work of the author.

REFERENCES

[1] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” Nature, vol. 630, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.

[2] Z. Ji et al., “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, no. 12, Art. no. 248, 2023, doi: 10.1145/3571730.

[3] L. Huang et al., “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” ACM Transactions on Information Systems, vol. 43, no. 2, Art. no. 42, 2025, doi: 10.1145/3703155.

[4] S. Liu et al., “A hallucination detection and mitigation framework for faithful text summarization using large language models,” Scientific Reports, vol. 16, Art. no. 1374, 2026, doi: 10.1038/s41598-025-31075-1.

[5] E. Asgari et al., “A framework to assess clinical safety and hallucination rates of large language models for medical text summarisation,” npj Digital Medicine, vol. 8, Art. no. 274, 2025, doi: 10.1038/s41746-025-01670-7.

[6] M. Omar et al., “Large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support,” Communications Medicine, vol. 5, Art. no. 330, 2025, doi: 10.1038/s43856-025-01021-3.

[7] R. Massenon et al., “‘My AI is lying to me’: User-reported LLM hallucinations in AI mobile apps reviews,” Scientific Reports, vol. 15, Art. no. 30397, 2025, doi: 10.1038/s41598-025-15416-8.

[8] A. Alansari and H. Luqman, "LLM hallucination: A comprehensive survey," arXiv preprint, arXiv:2510.06265, 2025.

[9] P. Roy, "Deep ensemble network for sentiment analysis in bi-lingual low-resource languages," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 1, Art. no. 8, 2024, doi: 10.1145/3600229.

[10] P. Pakray et al., "Natural language processing applications for low-resource languages," Natural Language Processing, vol. 31, pp. 183–197, 2025, doi: 10.1017/nlp.2024.33.

[11] I. Chalkidis et al., "LexGLUE: A benchmark dataset for legal language understanding in English," in Proc. ACL, 2022, pp. 4310–4330.

[12] F. Ariai et al., "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges," ACM Computing Surveys, vol. 58, no. 6, Art. no. 163, 2025, doi: 10.1145/3777009.

[13] D. Yadav et al., "Cross-lingual named entity recognition for low-resource languages," in Proc. ACL Workshop on Multilingual Representation Learning, 2024, pp. 167–174.

[14] S. Maddu and V. Sanapala, "A survey on NLP tasks, resources, and techniques for low-resource Telugu-English code-mixed text," ACM Transactions on Asian and Low-Resource Language Information Processing, 2024, doi: 10.1145/3695766.

[15] M. Omar et al., "Large language models are highly vulnerable to adversarial hallucination," medRxiv preprint, doi: 10.1101/2025.03.18.25324184, 2025.

[16] J. Zhang et al., "Neural machine translation for low-resource languages: A survey," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 6, Art. no. 80, 2024, doi: 10.1145/3665244.

[17] D. Sulistyo et al., "Pivoted low-resource multilingual translation with named entity optimization," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 24, no. 5, 2025, doi: 10.1145/3727876.

[18] B. Wanjawa et al., "KenSwQuAD—A question answering dataset for Swahili low-resource language," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 4, Art. no. 113, 2023, doi: 10.1145/3578553.

[19] M. Munaf et al., "Low-resource summarization using pre-trained language models," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 10, Art. no. 141, 2024, doi: 10.1145/3675780.

[20] E. Ramdinmawii and S. Nath, "Resource building and classification of Mizo folk songs," Natural Language Processing, vol. 31, pp. 655–673, 2024, doi: 10.1017/nlp.2024.23.

[21] A. Üstün et al., "Aya model: An instruction finetuned open-access multilingual language model," in Proc. ACL, 2024, pp. 15894–15939.

[22] BigScience Workshop, "BLOOM: A 176B-parameter open-access multilingual language model," arXiv preprint, arXiv:2211.05100, 2023.

[23] Y. Bang et al., "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in Proc. IJCNLP-AACL, 2023, pp. 675–718.

[24] [K. Ahuja et al., "MEGA: Multilingual evaluation of generative AI," in Proc. EMNLP, 2023, pp. 4232–4267.

[25] Anonymous, "Towards inclusive NLP: Evaluating LLMs on low-resource Indo-Iranian languages," NeurIPS submission, 2025.

[26] J. Nay, "Natural language processing and machine learning for law and policy texts," SSRN Electronic Journal, 2018, doi: 10.2139/ssrn.3438276.

[27] D. Premasiri et al., "Survey on legal information extraction: Current status and open challenges," Knowledge and Information

Systems, vol. 67, pp. 11287–11358, 2025, doi: 10.1007/s10115-025-02600-5.

[28] U. Khalid, "Natural language processing in legal document analysis," Multidisciplinary Research in Computing Information Systems, vol. 4, no. 2, pp. 87–97, 2024.

[29] I. Trancoso et al., "The impact of language technologies in the legal domain," in Multidisciplinary Perspectives on AI and the Law, Springer, 2024.

[30] E. Bertoni et al., Handbook of Computational Social Science for Policy. Springer, 2023.

[31] P. Roy et al., (if additional low-resource or ensemble study included separately; adjust numbering if needed).