IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151059

DETECTING AND MITIGATING HALLUCINATIONS IN LARGE LANGUAGE MODELS: A REVIEW OF UNCERTAINTY, EVIDENCE GROUNDING, AND AGENTIC WORKFLOWS

Anindya Singh[1], Rishiraj Singh[2]

Department of SOCA, Babu Banarasi Das University, Lucknow, India[1,2]

Abstract: Large language models (LLMs) can give answers that sound smooth and believable but may still be wrong, unsupported, or self-contradictory. This becomes a bigger concern when LLMs are used in retrieval systems and agent-based workflows, where they may choose tools, understand user needs, search databases, and perform actions. In these situations, even one incorrect claim can affect later decisions.

This review looks at research on LLM truthfulness, hallucination detection, uncertainty, response consistency, retrieval-augmented generation (RAG), internal model signals, and agent reliability. It groups existing methods based on the evidence they use agreement between generated answers, prediction uncertainty, model-internal information, external retrieved sources, and workflow context.

One key point is that no single signal is enough to detect hallucinations. A model can be confident but wrong, retrieved information can be incomplete or misleading, and several answers may agree without being correct.

Therefore, reducing hallucinations requires combining different types of evidence.

To address this gap, the paper proposes a unified evidence-integration framework. The framework combines **token-level probabilities** and entropy, hidden-state and attention-based features where available, sensitivity to prompt changes, the quality and relevance of retrieved sources, claim-level support, and domain-specific context. In addition, it incorporates a **confidence-score check** and **intent analysis using a classifier model** to improve the reliability of the overall decision process.

Using this risk estimate, the system can answer directly, verify the response again, retrieve more information, ask the user for clarification, or avoid answering when confidence is low. The Model Context Protocol (MCP) is discussed to link agents with external tools and data, but it does not by itself ensure that the information is accurate.

Because the project materials describe a proposed method rather than completed experiments, this review does not include new benchmark results. It ends by highlighting future evaluation areas such as detection accuracy, calibration, robustness, speed, cost, and safe action execution in agentic systems.

Keywords: large language models; hallucination detection; factuality; semantic entropy; uncertainty estimation; retrieval-augmented generation; Model Context Protocol; agentic AI; selective prediction

1.INTRODUCTION

Large language models are increasingly used to answer questions, summarize information, locate relevant material, write code, and support decision-making. However, their ability to generate fluent, readable text does not guarantee factual accuracy. Researchers commonly use the term *hallucination* to refer to outputs that are incorrect, unsupported by evidence, inconsistent with a source, or fabricated yet presented convincingly. Since the term is applied inconsistently across studies, precise definitions are essential for fair and reliable evaluation.

A central issue is that these models are trained to predict the next word in a sequence. This enables fluent writing, but it does not require them to verify every claim against reliable information. Problems can arise from multiple factors: incomplete or outdated training data, limited knowledge in specialized domains, ambiguous questions, weak logical reasoning, misleading inputs, decisions made during text generation, or ineffective use of retrieved documents. As a result, a model may confidently provide an incorrect answer, while a more cautious response may actually be accurate.



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.151059

The importance of this issue depends on the context. In brainstorming or creative work, unusual or speculative ideas can be useful, provided users treat them as possibilities rather than established facts. In high-stakes domains such as medicine, law, finance, research, journalism, or business operations, however, unverified claims can mislead people and lead to poor decisions. The risk becomes greater when a language model functions as part of an automated agent. Such a system may need to interpret a request, plan a series of actions, collect information, choose tools, process data, and complete a task. An error at any stage can propagate through the workflow and result in harmful or incorrect real-world actions.

This review brings together two related strands of existing research: approaches for detecting hallucinations in individual model responses, including methods based on uncertainty, consistency across multiple outputs, retrieved evidence, and internal model signals; and the reliability of LLM-based agents, where interpreting user intent, selecting relevant context, using tools, and executing actions create additional opportunities for failure. The paper's main contribution is to integrate these ideas into a unified framework. It does not claim to have developed and experimentally evaluated a new system.

2.LITERATURE REVIEW

Large Language Models (LLMs) are widely used in AI-powered information-retrieval systems. By predicting the most likely next token based on extensive training data, they can produce fluent responses; however, these outputs may still include hallucinations. Early detection approaches primarily focused on analyzing the generated text, whereas more recent research explores uncertainty, consistency across outputs, retrieval-based grounding, and internal model signals.

Banerjee et al. (2024) characterized hallucinations as an inherent limitation of LLMs, arising from weaknesses in training, retrieval, reasoning, generation, and fact-checking. Methods such as Retrieval-Augmented Generation (RAG), Chain-of-Thought prompting, self-consistency, and uncertainty estimation can help reduce hallucinations, but they cannot eliminate them completely. Farquhar et al. proposed semantic entropy, which estimates hallucination risk by assessing the semantic similarity among alternative responses. Their results indicated that semantic similarity can offer a more informative uncertainty signal than surface-level textual similarity. Nevertheless, a model's confidence does not always reliably reflect factual accuracy.

Sriramanan et al. introduced LLM-Check, a method that identifies hallucinations in individual responses by analyzing hidden states, attention patterns, and token-level uncertainty. While it reduces the need for repeated generation and external retrieval, its effectiveness varies across different models, datasets, and types of hallucinations.

Farquhar et al. Farquhar et al. proposed a hallucination-detection method for large language models based on semantic entropy. Rather than focusing solely on the exact wording of different responses, the approach groups answers that express the same underlying meaning and measures the uncertainty across these semantic groups. When responses differ substantially in meaning, semantic entropy increases, indicating greater uncertainty and a higher likelihood of hallucination.

The authors evaluated this approach across multiple language models and a variety of tasks, finding that it outperformed several earlier uncertainty-estimation methods. Its main advantage is its ability to flag responses in which the model appears uncertain or inconsistent. However, the method identifies potentially unreliable outputs rather than correcting them; it does not directly verify facts or repair factual errors after they have been generated.

3.DEFINING AND CLASSIFYING HALLUCINATIONS

Hallucinations can take different forms depending on where the error arises. The following categories offer a useful framework for classifying them:

Factual or world-knowledge errors: The model makes claims that conflict with reliable external facts or fabricates names, dates, numbers, citations, or events.

Contextual or source-grounding errors: The response contradicts the provided documents or database records, misquotes them, or includes information absent from the source material.

Reasoning and consistency errors: The answer contains internal contradictions, flawed logic, or conclusions that do not logically follow from the available information.

Instruction- and intent-related errors: The model misunderstands the user's actual request, overlooks an important condition, or answers a question that was not asked.

Agentic and tool-use errors: In agent-based systems, the model may choose an inappropriate tool, fabricate a tool result, provide incorrect inputs, misinterpret returned data, or carry out an unjustified action.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151059

Temporal and domain-specific errors: The response relies on outdated information or fails to adhere to important rules, standards, or conventions specific to a particular field.

4.SOURCES AND CAUSES OF HALLUCINATION

Hallucination does not originate from a single component of a language model system. During data preparation, the training material may contain inaccurate facts, contradictory information, limited coverage of certain topics, or outdated content. At the architecture and pretraining stage, the model learns to predict the next word in a sequence; while this enables natural-sounding text, it does not ensure that the information is accurate or evidence-based.

Fine-tuning can improve a model's ability to follow instructions, but it may also make the model more inclined to provide a confident answer simply because the question appears expected. During inference, issues can further arise from ambiguous user instructions, randomness in text generation, truncated context, or incorrect interpretation of retrieved documents.

In agentic systems, the focus shifts from evaluating a single generated response to examining an entire chain of decisions. An agent may misunderstand the user's intent, assume it has sufficient information when it does not, retrieve irrelevant material, invoke a tool with incorrect inputs, or draw conclusions beyond what the tool results actually support. Across multiple steps, a minor error early in the process can compound into a more significant failure later.

5.EXISTING DETECTION APPROACHES

Recent surveys categorize hallucination-detection methods into several broad groups: retrieval-based, uncertainty-based, embedding-based, learning-based, and self-consistency-based approaches. However, no single method performs reliably across all situations.

5.1 Consistency-Based Detection

Consistency-based methods check whether a model gives similar answers when the same question is asked in different ways or when multiple responses are generated. If the answers disagree on an important fact, this disagreement may suggest that the model is uncertain. Self-consistency is often used as a practical way to identify unstable responses, while paraphrasing the question can show whether the answer remains the same when the wording changes.

However, agreement does not always mean correctness. A model may repeatedly give the same wrong answer, especially if that mistake is strongly linked to patterns it learned during training. Generating several responses also takes more time and computing power, which can increase cost and delay.

5.2 Predictive Uncertainty: Token Probabilities, Entropy, and Perplexity

Some detection methods rely on the model's internal prediction signals, including token probabilities, log probabilities, entropy, and perplexity. Entropy measures how dispersed the model's predictions are when selecting the next word; higher entropy may indicate lower confidence at that point in the response. By analyzing uncertainty across different parts of an answer, it may be possible to identify the sections most likely to contain errors.

Perplexity indicates how well a model can predict a given sequence of text, but it does not directly determine whether the text is true. A frequently repeated false statement may receive a high probability, whereas a less familiar but accurate statement may receive a low probability. Consequently, these signals are best interpreted as indicators of potential hallucination rather than as direct evidence of truth or falsehood.

5.3 Semantic Entropy

Farquhar et al. proposed semantic entropy to measure uncertainty by grouping answers that express the same meaning, rather than treating differences in wording as real disagreement. This approach solves a limitation of simple word-based comparison: two answers may use different words but convey the same idea, or they may use similar words while making different claims.

Semantic entropy can be a useful uncertainty signal when the model generates multiple responses. Still, it has limitations. It requires generating several answers, its accuracy depends on how well similar meanings are grouped together, and being uncertain about an answer does not necessarily mean the answer is factually wrong.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940



International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151059

5.5 Retrieval-Based Verification and RAG

Retrieval-augmented generation (RAG) connects a language model to external sources such as documents, databases, or other knowledge repositories. This approach can improve information freshness, make responses more domain-specific, and facilitate tracing claims back to their original sources.

At the same time, RAG introduces new challenges. The system may fail to locate the relevant passage, retrieve outdated or low-quality documents, rank irrelevant material too highly, or return sources that conflict with one another. Even when correct evidence is available, the model may misinterpret it or make claims that the source does not actually support. Therefore, RAG should be paired with checks on source quality, claim-level verification, and a clear process for handling missing or conflicting evidence.

6.MITIGATION STRATEGIES

6.1 Evidence-oriented prompting and abstention

Prompts can guide a model to separate facts from assumptions, refer to the exact source passages it used, admit when there is not enough evidence, and ask follow-up questions when the user's request is unclear. These instructions can improve the model's behavior, but they do not guarantee accuracy. They work best when the system also allows the model to say, "I don't know" and when evaluation gives credit for appropriately avoiding an answer when evidence is genuinely missing.

6.2 Selective Prediction and Calibrated Risk

Selective prediction means that a system responds only when its estimated risk is sufficiently low. If the risk is too high, it can search for additional evidence, request clarification from the user, escalate the case to a human, or decline to answer. For this approach to work effectively, the risk score must be well calibrated: responses assigned the same risk level should exhibit comparable error rates. Confidence should not be conflated with accuracy. Risk thresholds should also be task-dependent; a minor error may be acceptable in creative suggestions but not in financial transactions or medical summaries.

6.5 Model Context Protocol as an Integration Layer

The Model Context Protocol (MCP) is an open standard for connecting AI applications to external data sources and tools. In hallucination-reduction architecture, MCP can offer a consistent mechanism for an agent to access approved knowledge bases, databases, or verification tools. However, the ability to connect systems does not automatically ensure information reliability. MCP does not guarantee that a source is trustworthy, that a query is correctly formulated, or that a generated claim is supported by evidence. Practical implementations still require authentication, authorization, validation of tool outputs, source tracking, and policy-based checks.

6.6 Two-Stage Hallucination Reduction Framework

The proposed framework rests on the hypothesis that a two-stage architecture—combining a lightweight intent classifier with a larger language model connected to a domain-specific database—can reduce both hallucination risk and computational cost by filtering irrelevant queries before they reach the main model.

In the first stage, a lightweight classifier with approximately 0.8 billion parameters is expected to analyze the user's input and determine whether the query is relevant to the information available in the connected database. If the query is unrelated or falls outside the supported domain, it would be rejected at this stage and would not be forwarded to the main language model. This filtering step is expected to reduce unnecessary inference and token consumption.

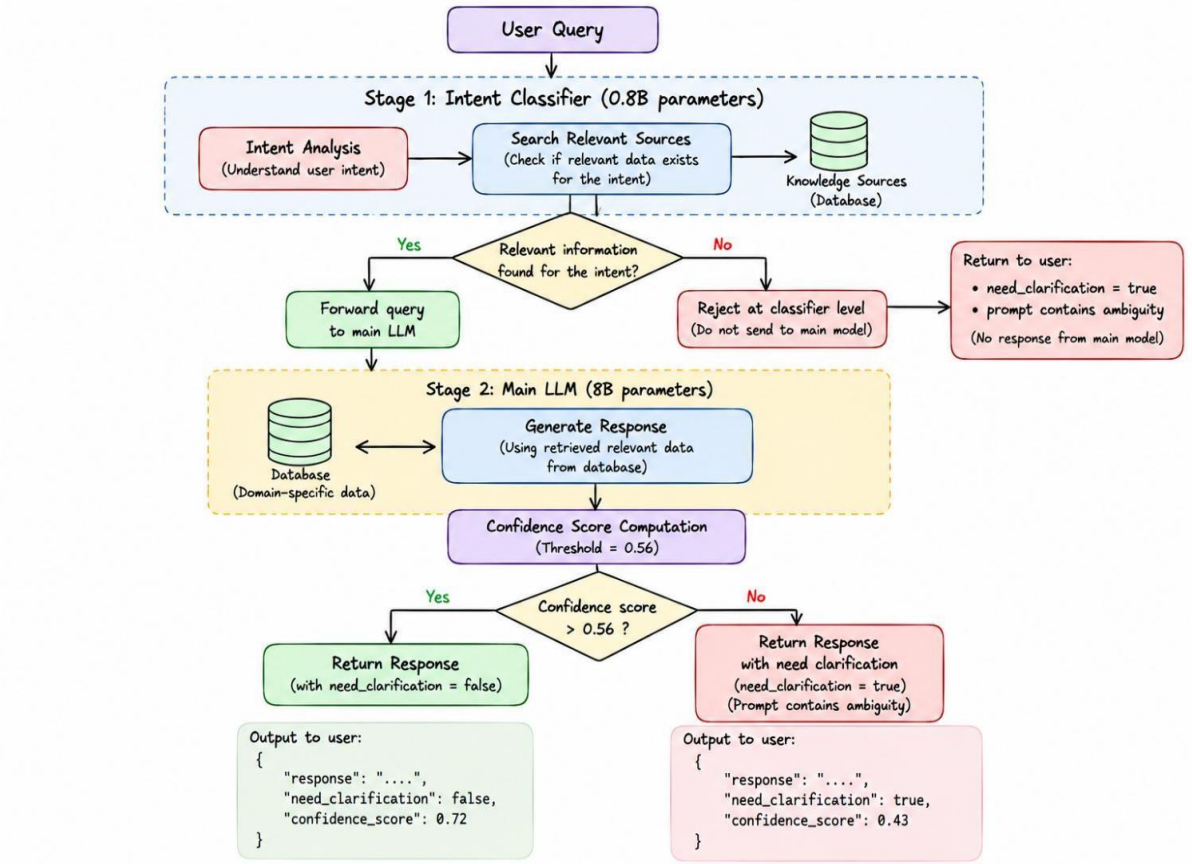
In the second stage, only relevant queries would be sent to the main language model, which has approximately 8 billion parameters. The model would generate responses using the connected database as its primary source of knowledge. A confidence-based acceptance mechanism is also proposed to assess the reliability of each response. Under this hypothesis, a response would be returned only if its confidence score exceeds a predefined threshold of 0.56; otherwise, the system would indicate whether user clarification is needed.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026

DOI: 10.17148/IJARCCE.2026.151059



The central assumption underlying this architecture is that separating query classification from response generation can improve efficiency. By using a smaller model for initial filtering, the framework could avoid invoking the larger model for irrelevant queries, thereby reducing computational overhead and unnecessary token usage. It is further hypothesized that combining database-grounded generation, relevance filtering, and confidence-based response acceptance would decrease the likelihood of unsupported or hallucinated outputs.

7.A UNIFIED EVIDENCE-INTEGRATION FRAMEWORK

Building on the limitations identified in the reviewed literature, this paper proposes an integrated framework that combines predictive signals, internal model features, contextual information, and external evidence to estimate hallucination risk and support evidence-based decision-making.

Intent parsing: Convert the user's request into a structured intent, including the required action, entities, constraints, output format, and any ambiguities. If essential information is missing, ask a clarifying question.

Context requirement analysis: Identify only the evidence necessary for the task, such as relevant conversation history, business rules, database schema, or domain knowledge.

Evidence retrieval: Retrieve information from approved sources and record source details, timestamps, permissions, and retrieval scores. Preserve missing or conflicting evidence rather than silently filling gaps.

Controlled generation: Generate a response using evidence-aware instructions, validated tool inputs, and authorized actions. Verify database queries and returned records before using them.

Multi-signal risk estimation: Collect signals such as token probabilities, entropy, perplexity, internal model features, prompt sensitivity, retrieval quality, and the degree of support between claims and evidence.

A simple conceptual formulation is $R = f(U, I, C, G, V, D)$, where Rrepresents the estimated hallucination risk; Udenotes predictive uncertainty; I refers to internal-state signals; C captures consistency and prompt sensitivity; G reflects

DOI: 10.17148/IJARCCE.2026.151059

grounding and retrieval quality; V represents claim-level verification outcomes; and D denotes domain- or task-specific risk. The function f should be learned and calibrated using labeled validation data; no specific weights or performance values are assumed here. The decision policy then maps R and the level of task risk to an appropriate action. This formulation makes the proposed framework modular: when internal model signals are unavailable, the system can still rely on external evidence, consistency checks, and deterministic validation.

8.RESEARCH GAPS AND FUTURE DIRECTIONS

Current approaches to hallucination detection generally rely on consistency checks, uncertainty estimation, or external retrieval. Although these methods are useful, each has notable limitations. Consistency-based approaches may fail when a model repeatedly produces the same incorrect answer. Uncertainty-based methods can incorrectly classify unfamiliar but accurate information as unreliable. Retrieval-based methods depend on the quality and completeness of external sources, which may be outdated, incomplete, or contradictory.

The main research gaps identified in this work are as follows:

Absence of a unified detection framework: Existing techniques typically focus on a single type of signal, such as consistency, probability, internal model states, or retrieved evidence. An integrated framework is needed to combine these signals and provide a more reliable estimate of hallucination risk.

Confidently incorrect outputs: Large language models can generate incorrect answers with high confidence, making confidence alone an unreliable indicator of correctness.

High computational cost: Some detection methods require generating multiple responses, performing additional retrieval steps, or running separate verification models. These requirements can increase latency, token usage, and overall computational cost.

Incomplete or unreliable retrieved evidence: Retrieval-augmented systems may fail to identify the most relevant information, return low-quality sources, or retrieve documents that conflict with one another. Consequently, even grounded responses may contain unsupported claims.

Limited refusal mechanisms: Models often lack a reliable way to decline answering when available evidence is insufficient. A stronger mechanism is needed to enable systems to request clarification, seek human review, or abstain from answering rather than producing a potentially incorrect response.

Lack of standardized evaluation: Hallucination is measured differently across domains, tasks, and research studies. There is no widely accepted evaluation standard that enables fair comparison of methods across different LLM applications and agentic systems.

8.Limitations

This review is limited to the source materials and selected literature used during manuscript preparation. It is a narrative technical review rather than a registered systematic review or meta-analysis, and it does not claim to cover every hallucination-detection method or benchmark available in the field.

No new experiments or numerical results are presented in this paper. The evaluation protocol described earlier should therefore be understood as a proposed plan for future validation rather than a completed empirical study. A future version of this work reporting experimental findings should also include clearly defined research questions, datasets, models, baseline methods, evaluation metrics, implementation details, obtained results, and a discussion of limitations once those experiments have been conducted.

9.CONCLUSION

Hallucination remains one of the most significant reliability challenges for large language models and the agent-based systems built around them. Researchers have explored various detection methods, including consistency checks, semantic entropy, predictive uncertainty, internal model signals, retrieval-based grounding, and model-assisted verification. However, no single method fully resolves the problem on its own.

The reviewed literature also highlights the need for a better understanding of user intent, more selective retrieval of relevant context, and stronger safety controls in agentic systems. To address these gaps, this review proposes an integrated framework that combines multiple signals to estimate hallucination risk in a calibrated manner. Depending on the

 This work is licensed under a Creative Commons Attribution 4.0 International License



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 10, October 2026
DOI: 10.17148/IJARCCE.2026.151059

available evidence and the risk level of the task, the system could respond directly, perform additional verification, request clarification from the user, or decline to answer.

Although the Model Context Protocol (MCP) can enable agents to connect with external tools and data sources, connectivity alone does not ensure reliable outcomes. Trust still depends on the quality of the sources, proper validation of outputs, appropriate access controls, and accurate interpretation of retrieved evidence.

Future work should focus on reproducible evaluation across both individual-response tasks and database-driven agent tasks. Such evaluations should measure detection accuracy, calibration, robustness, computational cost, and the frequency of unsafe or unauthorized actions.

REFERENCES

[1] Y. Wang, M. Wang, M. A. Manzoor, F. Liu, G. N. Georgiev, R. J. Das, and P. Nakov, "Factuality of Large Language Models: A Survey," *Proceedings of EMNLP 2024*, pp. 19519– 19529, 2024, doi: 10.18653/v1/2024.emnlp-main.1088.

[2] P. N. Venkit, T. Chakravorti, V. Gupta, H. Biggs, M. Srinath, K. Goswami, S. Rajtmajer, and S. Wilson, "An Audit on the Perspectives and Challenges of Hallucinations in NLP," *Proceedings of EMNLP 2024*, pp. 6528–6548, 2024, doi: 10.18653/v1/2024.emnlp-main.375.

[3] 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," *Proceedings of ACL 2024*, pp. 10879–10899, 2024, doi: 10.18653/v1/2024.acl-long.586.

[4] X. Lin et al., "LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions," *arXiv preprint*, arXiv:2509.18970, 2025. This is a preprint; any later peer-reviewed version should be checked and cited before final submission.

[5] 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/s41586024-07421-0.

[6] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, "LLM-Check: Investigating Hallucination Detection in Large Language Models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 34188–34216, 2024, doi: 10.52202/0790171077.

[7] Anthropic, "Model Context Protocol," official technical documentation. Available: https://docs.anthropic.com/en/docs/mcp (accessed October 14, 2026). This is technical documentation rather than a peer-reviewed research paper.

[8] S. Banerjee, A. Agarwal, and S. Singla, "LLMs Will Always Hallucinate, and We Need to Live with This," *arXiv preprint*, arXiv:2409.05746, 2024. This is a preprint; its conclusion should be presented as the authors' argument rather than an established fact.

[9] Alansari and H. Luqman, "Large Language Models Hallucination: A Comprehensive Survey," *Computer Science Review*, vol. 61, Art. no. 100970, August 2026, doi: 10.1016/j.cosrev.2026.100970.