2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC

Host Sponsor UNC Charlotte

ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Paper ID #50824

A review on reducing AI and machine learning induced hallucinations in LLM using agent-to-agent validation

Dr. Tahir M Khan, Western Illinois University

Dr. Tahir M. Khan is an Associate Professor of Cybersecurity in the Department of Computer Science and currently serves as the interim Director of the Cybersecurity Center at Western Illinois University. He is actively involved in teaching both undergraduate and graduate students and oversees the operations of the Cybersecurity Center, focusing on cybersecurity education, outreach, and research. Dr. Khan has extensive experience collaborating with students on research projects, redesigning curricula, and mentoring students and individuals with disabilities pursuing degrees in STEM fields. He has also successfully managed grants and organized summer camps for high school students. His research interests include computer privacy and security, digital forensics, threat detection, the Internet of Things (IoT), cloud computing, and artificial intelligence.

Osondu Chiemeka Onwuegbuchi, Western Illinois University

A Review on Reducing AI and Machine Learning-Induced Hallucinations in Large Language Models Using Agent-to Agent Validation

Abstract

The current paper addresses the Agent-to-Agent Validation (A2AV) framework as a novel approach to mitigating hallucinations in large language models (LLMs). Compared to the single-model pipelines that characterize the traditional approach, the consideration of A2AV in this work is presented as a program that can support multiple independent agents that have the capacity to generate, evaluate, and iteratively refine responses until consensus is reached. The model is shown using a Python example replicating the between-agent verification using transformer-based models, demonstrating how factual accuracy can be improved using this cycle iteratively. Healthcare, legal, and scientific applications are domains where the transformative potential of A2AV is evident, in domains where accuracy and accountability matter the most. While the framework has obvious advantages such as reduced hallucinations, improved factual grounding, and decentralized trust, challenges such as computational expense, potential for agent collusion, and latency in real-time systems remain. Proposed direction for future research are provided to include the integration of explainable AI with A2AV, symbolic reasoning, retrieval-augmented generation, and multi-agent consensus models, as well as the development of benchmark evaluation metrics for hallucination detection. The paper illustrates the promising direction that A2AV can offer for the development of trustworthy and interpretable AI systems.

Keywords: Agent-to-Agent Validation, Large Language Models, Hallucination Mitigation, Multi-Agent Systems, Trustworthy AI

1.0 Introduction

Numerous applications of large language models, such as Claude, the Generative Pretrained Transformer (GPT) series, Large Language Model Meta AI (LLAMA), Gemini, and other similar products have changed access to information, decision-making, and creation of knowledge in various fields including medicine, education, law, finance, and scientific research. Their ability to produce human-like output has been regarded as a distinct trait that emphasizes their importance in helping users undertake complex reasoning and communication processes. Nevertheless, the latest studies [5, 14] have recorded that these systems have existing weaknesses that may cause or exacerbate hallucinations. The studies by [3, 39] refer to such situations when the models come up with coherent and factually erroneous, fabricated, or logically inconsistent information as hallucinations. This means that faults may come into play to cause negative impacts of compromised trust and credibility, particularly in applications that are safety wise critical.

These issues have been studied in recent research conducted by [6] and [5]. Indicatively, according to the research of Devlin et al. [34], classical generative chat-bots could provide inaccurate information about cancer, and the rate of hallucinations reaches up to 40% according to the study. Retrieval-augmented generation (RAG) on the other hand, minimized the number of hallucinations by half to 0-6% with regard to difficult sources, which was encouraging in retrieval methods.

Equally, Darwish et al. [2] observed that in rheumatology, only the multi-step reasoning and online data integration capabilities are absent in the case of static LLMs and are limited in its clinical utility. Education is also no exception to hallucinations, and Ji et al. [13] has identified such geographic hallucinations as fictitious sources and distortions of space as obstacles to the acceptable integration of AI into the educational curricula. Joshi [3] pointed out that hallucinations undermine productivity and trust in decision-making in the business and fiscal setting and require avoidance through a systematic approach.

In order to overcome these shortcomings, an increasing amount of research has examined mitigation measures. RAG-based tools enhance grounding in facts but are susceptible to the cases of retrieved information that is not credible [1, 9]. Research measures such as prompt engineering, reinforcement learning fine-tuning, and integration with a knowledge graph have demonstrated potential, but the majority are facing serious scalability limitations [20]. Emerging paradigms are mostly directed toward multi-agent systems and competitive argumentation to verify output cross-checks. For example, Darwish et al. [19] illustrated that response consistency is enhanced by 85.5% using rule-based multi-agent paradigms, whereas Yang et al. [8] proposed adversarial argumentation and voting mechanisms that increase factual reliability and lower hallucination rates. Besides, Peasley et al. [7] showed that agent-based scientific assistants can achieve over 90% factual coherence, outperforming single-agent capability in data-intensive reasoning. Collectively, these findings show that agentic AI and multi-agent verification offer a potentially viable route to enhancing LLM reliability.

Despite some progress, gaps in knowledge persist. While hallucinations are addressed to a good extent by RAG, prompt engineering, and reinforcement learning, they still don't eliminate them, particularly in real-world, dynamic, or domain-specific environments. Furthermore, most existing solutions operate at the single-agent level, which constrains their capacity for independent verification and self-correction. As an example, albeit with some improvements in the field, chain-of-thought prompting [31], constitutional AI methods, and retrieval-augmented architectures tend to be less distributed than multi-agent systems could be in terms of verification capabilities. However, more recent academic research has started to consider structured Agent-to-Agent Validation (A2AV) systems with focus on more collaborative LLMs validating, interrogating, and verifying answers. Nevertheless, these frameworks have not been studied thoroughly and are not fully applied in practice [8, 19].

This review is thus aimed at examining the phenomena of hallucinations in LLMs, and also at assessing the effectiveness of agent-to-agent validation as a novel technique in mitigation. Based on experience in healthcare, education, business, science, and computational paradigms, the paper incorporates the existing practices, introduces new agent-based solutions, and gives recommendations on how A2AV can be scaled up to real-world application.

This study is important due to the anticipated assistance in the reduction of the disparity between non-integrated mitigation approaches and growing requirements of dependable, answerable, multi-agent systems, along with precision and candidness in AI systems. A2AV is the subject of the review, and its contents can enable the intersection of the reliability of LLM and agential AI, and it is expected to be beneficial to introduce a conceptual framework and tools to apply safer generative systems in practice.

1.1 Scope and Methodology

Those mitigation strategies against hallucinations mentioned in the peer-reviewed literature from recent years (2023-2025) are within the scope of this review, with particular attention given to agent-to-agent validation frameworks, the majority of which have emerged in the past few years. Usually, the phenomenon of hallucinations has been known since the beginning of neural language models, but the explosion of atmospheric usage of LLMs in critical settings has been triggered only a few years ago, leading to a wider study of structured mitigation strategies. We use the temporal perspective to cover the post-ChatGPT time (after November 2022), where the development of hallucination mitigation turned into a systematic structure instead of unstructured methods.

We used the following keywords—LLM hallucinations, large language model hallucination mitigation, agent-to-agent validation, multi-agent AI systems, factual grounding LLMs, and AI confabulation—to conduct systematic searches in the ACM Digital Library, IEEE Xplore, arXiv, PubMed, and Google Scholar. Inclusion criteria prioritized: (1) peer-reviewed articles and preprints from reputable sources, (2) empirical studies with quantitative hallucination metrics, (3) novel mitigation frameworks with implementation details, and (4) domain-specific applications in healthcare, law, education, or science. This scoping decision reflects our objective to survey the current state rather than provide a comprehensive historical account of all mitigation attempts since neural language modeling began.

2.0 Literature Review

2.1 Nature and Implications of Hallucinations

Hallucinations in LLMs and other generative AI systems are among persistent issues for AI practice and research. Unlike human hallucinations, which involve sensory perceptions without external stimuli, Hatem et al. [12], research has shown that AI hallucinations involve the generation of text that may be linguistically coherent but factually inaccurate,

logically contradictory, or entirely fictional [14, 12]. The implication is that these outputs may be convincingly surface-level and degrade the credibility of AI systems, especially in applications that cannot compromise with accuracy, for instance in health and finance.

2.1.1 Typologies of Hallucinations

Taxonomic research by Sun et al. [16] and Ji et al. [13] has proposed several classification schemes to address the multifaceted nature of AI hallucinations. For example, Sun et al. [16] conducted large-scale empirical studies with the analysis of 243 cases of distorted information generated by ChatGPT. They observed at least eight broad error types, including overfitting, logic errors, reasoning errors, math mistakes, unfounded fabrication, factual errors, text output errors, and other errors. This taxonomy further delineated 31 second-level error subtypes within these eight broad categories, demonstrating that hallucinations arise from diverse AI system failure modes rather than constituting a monolithic phenomenon.

Other systems distinguish between intrinsic hallucinations, where outputs directly contradict source text or prior conversational context, and extrinsic hallucinations, where fabricated details are introduced without basis in the source text [13, 14, 17] also stated a corresponding three-way distinction of input-conflicting, context-conflicting, and reality-conflicting hallucinations. Together, these definitions mean that hallucinations can arise due to both internal contradiction and decoupling from facts in reality.

2.1.2 Causes and Mechanisms

Özer, Lin et al, and Joshi [14, 30, 3] identify data quality deficiencies, model complexity, and inherent generative process limitations as primary causal factors. Training data may have outdated, incomplete, or inconsistent data and models may represent them without necessarily being able to distinguish fact and fiction [14]. Additionally, because LLMs are probabilistic, they may favor linguistic plausibility over factuality and generate false but coherent answers [34].

Another problem is source amnesia, a phenomenon referred to by Berberette et al. [11], in which models forget the mapping between outputs and sources within the training dataset. Once models generate a wrong output, the generative models will reinforce the error in subsequent text in a "snowball effect of hallucination" [14]. This self-consolidating mechanism stores misinformation, especially in longer conversations. Interestingly, the term hallucination itself is also a misnomer, contend some researchers, [12] and [15], in favor of instead a psychiatric-derived theory of confabulation, one that better explains the mechanism of completion of gaps in knowledge with internally consistent but actually fabricated information.

2.1.3 Implications in Real-World Applications

The effect of AI hallucinations has penetrated into various fields. The hallucinations in abstractive summarization and question-answering tests in HaluEval [28] are measured. FactCC [29] establishes factual compatibility of generated text and reference texts. Such benchmarks can be systematically compared to the mitigation strategies and may be applied as the base to create more credible models [1].

Hallucinatory behaviors in the legal profession present as a creation of fictitious precedents and fraudulent references, which leads to an ethical weakness and even discrediting the court system [3]. Education and academia are no exception: geographic hallucinations and forged references in the student outputs of AI tools have been reported by [8], and its consequences on misinformation dissemination. The fabricated references may jeopardize the academic credibility in the science field, and [10] discovered that the majority of references generated by ChatGPT are invalid.

Athaluri et al. [14] argue that AI hallucinations may decrease trust in AI systems, noting that challenges in how AI is perceived by the public—along with skepticism and opposition to its implementation—can further erode trust. Hallucinations may cause incorrect decision-making, diminished productivity, and investor confidence in enterprise and financial applications [3]. It has also been demonstrated that they can incorporate misinformation and perpetuate biases that intensify training data inequalities [14].

2.1.4 Benchmarks for Detection

In order to gauge and determine hallucinations, scientists came up with specific markers. TruthfulQA [30] determines whether the models come up with the correct answer to questions that are likely to be misconceived by humans. HaluEval [28] is a measurement of hallucinations in the abstractive summarization and question-answering problems. The factual coherence of generated text with reference texts is checked by FactCC [29]. These standards allow a comparative analysis of mitigation strategies in a systematic manner and can be analyzed in the context of establishing better credible models [22].

2.1.5 Balancing Risks and Benefits

Even though the idea of hallucinations has largely been presented as risks, Athaluri et al. [14] note a way in which some hallucinations can have creative value. As it has been demonstrated, hallucinations may lead to innovation in creative writing, design, and drug discovery, wherein plausible yet novel outputs may trigger human imagination [27, 26].

2.2 Recent Mitigation Strategies

In recent years, the body of research on mitigating hallucinations in LLMs has multiplied dramatically, with some methods being as straightforward as prompt-based interventions and others being complex multi-agent systems and architectures. The techniques have various structures, scalable features, and performance, and they are united by the aim of

enhancing the quality of facts, interpretation, and safe implementation of the LLM in a wide range of applications [36].

The best possible countermeasures to the hallucinations are Prompt Engineering and the construction of MCP tools and advanced Context Management. Yu et al. [25] have shown that the accuracy of predictions can be improved by the appropriate timely structuring, which is accomplished by self-consistency checks, chain-of-thought reasoning, or dynamic rebalancing of contextual layers. Yu et al. [25] proposed the Layer-Aware Contextual Distribution (LACD) approach, which redistributes probability weights across transformer layers to mitigate previously acquired biases that may affect newly introduced contextual information.

An additional accepted method is RAG, whereby models can ground answers in external knowledge bases. Empirical studies by Bhattacharya [18] and Gokcimen et al. [21] demonstrate that RAG substantially reduces the likelihood of hallucinations, though retrieval algorithm quality dictates its overall trustworthiness. Recent cross-LLM security architectures have enhanced RAG, in addition to the introduction of VectorDB and kernel-based designs. Gokcimen et al. [21] stated that the expected benefits are sustained stability regarding hallucinations and adversarial attacks. These developments suggest that hybrid designs can enhance factual foundation with sustainable security components capable of minimizing error spreading [38].

Some evidence has been documented; although fine-tuning, reinforcement learning with human feedback (RLHF), and reinforcement learning with AI feedback (RLAIF) methods have potential, they require extensive resources and often depend on well-annotated datasets. Darwish et al. [9] discovered that RLHF can align model outputs to be more factual but is fragile when presented with adversarial or out-of-domain examples. Likewise, fine-tuning alone cannot repair intrinsic knowledge gaps even if it can adjust tone and style.

Another fundamental direction is inherent in Verifier and Multi-Agent Models because instead of employing a single generative model, researchers are progressively investigating frameworks where independent agents verify, interrogate, or adjudicate outputs. Darwish et al. [19] illustrated an 85.5% increase in response consistency with rule-based multi-agent schemes. Yang et al. [8] presented adversarial debate and voting systems in which agents dispute one another until agreement is attained, dramatically decreasing hallucination rates in factual question answering.

Farquhar et al. [20] documented entropy-based uncertainty estimators, which detect confabulations by perceiving semantic-level uncertainty rather than token-level probabilities. Peasley et al. [7] illustrated that agent-based scientific assistants can surpass 90% factual consistency in empirical data summaries, significantly outperforming their single-agent competitors.

Joint mitigation of bias and hallucination is a topic of interest to researchers. Lin et al. [22] performed an overview of methods that act on both biases and hallucinations in synchrony, acknowledging that misleading outputs are frequently amplified by biased training data. Eliminating sociodemographic biases from datasets and outputs can decrease propagation of fake but culturally aligned narratives [37].

Some studies have also led to the conclusion that Security-Augmented Architectures are the fundamentals of current innovation. For example, Gokcimen et al. [21] observed a cross-LLM security architecture that simultaneously detected injection attacks in addition to hallucinations. The aggregation of multiple model outputs led to a significant improvement of adversarial robustness with a reported 98% accuracy, which matched an eligibility scoring on benchmark tasks. This observation indicated that the assurance on greater system reliability and robustness can be violated by divorcement of mitigation of hallucination.

Table 1 is a synthesis of the available mitigation techniques to LLM hallucinations, based on type of strategy, functional processes, exemplary implementations, and reported effectiveness. This taxonomy explains why a variety of interventions are available and can be more or less effective in a variety of application scenarios.

Table 1. Taxonomy of LLM Hallucination Mitigation Strategies

Strategy	**Key Mechanism**	**Example Implementations**	**Efficacy**
Prompt Engineering & Context Management	Self-consistency checks, chain-of-thought, contextual rebalancing	LACD [25]	Moderate; domain-dependent
Retrieval-Augmented Generation (RAG)	External knowledge grounding	VectorDB, cross-LLM RAG [21]	High; depends on retrieval quality
Fine-Tuning & RLHF/RLAIF	Reinforcement from feedback	GPT-4, Claude (Anthropic)	Moderate to high; resource-intensive
Multi-Agent Verification	Collaborative cross-checking, debate, voting	Rule-based systems [19], adversarial frameworks [8]	High (85-90% improvement)

Uncertainty & Entropy Estimation	Semantic uncertainty detection	Entropy estimators [20]	High for detection; lower for correction
Bias and Hallucination Joint Mitigation	Simultaneous debiasing and fact-checking	Integrated frameworks [22]	Emerging; shows promise
Security-Augmented Architectures	Concurrent attack and hallucination detection	Cross-LLM security [21]	Very high (98% accuracy on benchmarks)

Sources: [24, 19, 23, 25, 21, 22]

Table 1 illustrates that the avoidance of hallucination has a strategy that switches in the scope of operation and effectiveness such as lightweight changes to complex multi-agent and security centric infrastructure. Although no method offers a complete solution, the new development trends evenly support hybrid and multi-agent methods which integrate retrieval, verification, and continuous feedback to achieve the best degree of reliability.

3.0 Agent-to-Agent Validation (A2AV) Framework

A2AV is a novel framework designed to address the problem of hallucination in LLMs. While single-model systems have relied in the past on individual output alone, this framework is based on multiple independent agents that interact with one another. Each of these agents creates, validates, and refines responses, guaranteeing the distribution of trust between models and significantly minimizing the chances of false or fictitious information being taken as final output.

3.1 Concept

A2AV has its basis on the practice of cross-checking models. The system, rather than having one model generate an answer and accepting it as true to face value, is a well-structured dialogue where one agent produces an answer and another agent checks it to find out the accuracy of facts as well as logical consistency. The agent that is being reviewed authenticates the original output, or detects discrepancies and in this case, revisions are created and the process recurs until the two agents agree with a validated response. It is similar to the peer-review system of research, in which unbiased reviewers verify claims to minimize bias and error.

3.2 Workflow

The functionality of the A2AV system may be explained in the form of a series of actions. The conversation starts with a question by a user and continues. The generator is known as agent 1 and generates an initial response depending on the available context and its training.

The validator (agent 2) then analyses this response in terms of factual coherence, logical consistency or compatibility with reliable sources. In case of errors or ambiguity shown by Agent 2, it gives particular feedback and the response is refined by the Agent 1. This validation is done by organized questioning of the factual assertions of the first response. This process is repeated over and over again-revision, re-evaluation, and further refinement until the agents are satisfied that the output has reached pre-defined quality standards. This is when the final consensus answer is presented to the user.

3.3 Benefit

There are various benefits associated with this model compared to single-model strategies. When the system is heavily monitored by independent agents to cross-check each other, it gains increased strength against the hallucinations that may go unnoticed during a one-pass production. Decentralized trust implies that there is no one model that is the single-hand decision maker, which lowers the number of systemic biases and enhances trustworthiness. The refinement process is iterative and enables reduction of errors as the cycle progresses with each cycle reducing the error. Moreover, recording all the steps of the validation process, A2AV systems can provide the features of transparency and explainability that are essential in applications where accountability is an issue: in law, health, and scientific research.

3.4 Code Demonstration: Agent-to-Agent Validation in Python

We created the Python code below illustrating the concept of the A2AV with the help of the available open-source models. This code (proof-of-concept) demonstrates a simple example of the basic workflow, but it could be implemented with advanced validation code, larger models, and domain-specific knowledge bases.

The most typically used models include GPT-2 [32], DistilGPT-2, or other open-source transformers, which can be accessed in Hugging Face [35].

```
# Agent-to-Agent Validation Example

# Install transformers if not already: pip install transformers

from transformers import pipeline

# Define two agents (can be the same or different models)

# Using GPT-2 as example; can substitute with gpt2-medium, gpt2-large, etc.

agent1 = pipeline("text-generation", model="gpt2")

agent2 = pipeline("text-generation", model="gpt2")

def agent_to_agent_validation(query, max_length=80):
```

```
    """
    Agent 1 generates a response, Agent 2 validates it,
    and returns both the original and validation response.
    """
    # Step 1: Agent 1 generates an answer
    response1 = agent1(query, max_length=max_length,
num_return_sequences=1)[0]['generated_text']
    # Step 2: Agent 2 validates Agent 1's answer
validation_query = f"Verify if the following statement is factually correct: {response1}"
    response2 = agent2(validation_query, max_length=max_length,
num_return_sequences=1)[0]['generated_text']
    return {
        "Agent1_Response": response1.strip(),
        "Agent2_Validation": response2.strip()
    }
# Example usage
if __name__ == "__main__":
    query = "Who discovered Penicillin?"
    result = agent_to_agent_validation(query)
print("User Query:", query)
print("\nAgent 1 Response:", result["Agent1_Response"])
print("\nAgent 2 Validation:", result["Agent2_Validation"])
```

This developed implementation serves as a pedagogical illustration rather than a production-ready system. Researchers deploying A2AV in critical applications should incorporate domain-specific validators, retrieval-augmented knowledge bases, and robust consensus mechanisms as discussed in Section 5.0.

4.0 Agent-to-Agent Validation Applications

A2AV is more than a theory. It has actual practical application in domains where factual dependability is the primary requirement. The solution solves the historical hallucination issue of large language models by virtue of the ability to allow multiple agents to cross examine, authenticate outputs and guarantee information shown to users is reliable. The three broad areas that this method is most appropriate include healthcare, law and science research.

Indeed, no single area is as susceptible to A2AV as healthcare. LLM's are increasingly being applied to clinical decision support systems to either provide diagnosis or treatment recommendations or produce summary medical records. Uncontrolled errors may have catastrophic consequences for patient safety. With one agent conducting a diagnosis and another cross-checking the same with proven medical databases or guidelines, the chances of the hallucinated illness or not-otherwise-medically-recommended interaction are significantly lower. Such a cross-verify guarantee is a clinically accurate production prior to release to the healthcare professionals.

Precision is also important in legal systems. Contracts, case summaries, and precedent identification are all being done using LLLMs. One attorney may misinterpret precedents and provide poor legal advice. A2AV is used to reduce such a risk by ensuring that a single agent produces arguments of law and a second agent confirms citations to authoritative legal databases, such as Westlaw or LexisNexis. Such a two-step verification goes a long way in eliminating the chances of offering cases that are not actual or creating false interpretations of the ruling under legal briefs.

A2AV is also of great help to scientific research. The applications of LLMs by scientists have become common in literature reviews, hypothesis generation, or summarization of findings. Nonetheless, fabricated references and imaginary research are some of the most widespread issues that typically undermine the quality of AI-assisted studies. A reference can be created by one agent through validation loops and then another agent is used to confirm the existence of the reference by searching through large science databases like PubMed, Scopus, or Web of Science. This helps minimize the spread of false or fabricated references, ensuring that academic communications remain as credible as possible.

To conclude, the above applications reveal that A2AV goes beyond conceptual modeling to offer a viable tool of improving safety, credibility, and accountability in AI-assisted decision-making in critical areas.

Figure 1 shows the entire A2AV workflow cycle both the refinement loop (for the cases when consensus is not attained in the first place) and the straight path towards the validated output (when the early validation by Agent 2 proves the correctness). The continuous improvement is provided by the bidirectional feedback mechanism and the ability to have

sustained multi-turn interactions with the system is exhibited by the user query input return loop.

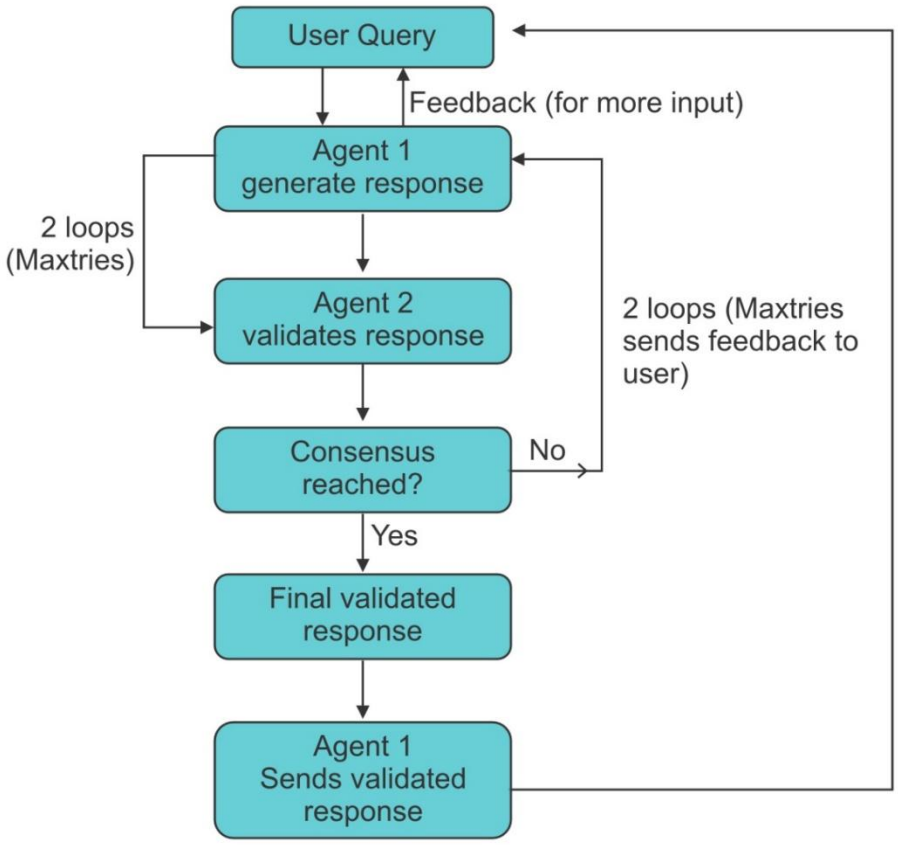

Figure 1: Agent-to-Agent Validation (A2AV) Workflow Diagram

Figure 1 depicts the iterative validation process in which Agent 1 produces an initial response to a user query, Agent 2 checks the response both in terms of factual accuracy and logical consistency and successive revisions are produced until a unanimous decision is reached. The bidirectional arrows represent feedback that allows continuous refinement. Once consensus (the final validated response) is reached, the system returns the output to the user or restarts to process new queries.

5.0 Challenges and Future Directions

Despite the A2AV framework offering good directions into enhancing the credibility of large language models, there exist challenges that have to be overcome in order to ensure that it becomes widely adopted.

5.1 Challenges

The framework maintains the existence of a variety of independent agents and is more correct in terms of facts due to the decentralized trust system and reduction of systematic bias. Its transformative potential has been demonstrated through applications in medicine, law, and science, and the Python implementation confirms its practical feasibility. The key issues have been identified in the discussion e.g., the fact that it is computationally expensive, prone to the agent collusion and experiencing latency in real-time systems implying areas that should be developed and optimized further. These results indicate that A2AV is one of the directions that can be chosen to create more trustful, transparent, and reliable AI systems. It provides decentralization of validation controls among various independent agents and thereby the dependency on individual model outputs is minimized and instead a kind of peer review environment is created where the credibility is enhanced.

The methodology is well aligned with the increasing demand of explainable and trusted AI systems even though the computational and methodological limitations have been documented. The future needs to work on the computation efficiency and maybe by the lightweight validator modelling or asynchronous verification protocol which minimizes latency.

An extension to three or more consensus multi-agents would provide an increased robustness as well, in the form of majority checks, and standard benchmarking protocols must be developed to determine the effectiveness of various A2AV designs. The framework could be further reinforced with the help of integration of the symbolic reasoning engines and the database with retrieval-augmented databases. Also, considering hybrid models with some elements of A2AV and explainable AI would provide the user not only with precise outputs but also with clear explanations of the way that consensus was reached to form trust in AI-based decision-making systems.

6.0 Conclusion

This review has critically analyzed the phenomenology of hallucinations of large language models and discussed the promising applicability of the Agent-to-Agent Validation as a new framework of mitigation. Through our analysis, we establish that albeit hallucinations are caused by various factors such as: training data shortages, probabilistic generation strategies, and through source amnesia, the hallucinations may be significantly reduced by using structured multi-agent verification systems.

The A2AV framework has been designed to counter the inherent drawbacks of single-model architectures through the spreading of the trust among independent agents who are able to jointly produce, assess and refine outputs until a consensus has been reached. The practical viability of this method is demonstrated by our Python implementation, and it promises to be useful in healthcare, law enforcement, and scientific fields where its utility and reality cannot be compromised.

However, significant issues remain. At this point, computational overhead, agent collusion possibilities, real-time latency concerns, and the standardized evaluation measures require further research. Research in the future ought to focus on the incorporation of explainable AI elements, symbolic reasoning systems, and sophisticated consensus systems to improve the efficiency and interpretability of A2AV systems.

With the growth in the deep implantation of LLMs in critical decision-support infrastructure, the pressure of hallucination mitigation increases. The present review aims to support that endeavor by summarizing the existing knowledge, suggesting the organizational validation models, and mapping the ways of achieving improved reliability and responsible AI systems. The switching of single-model output to multi-agent verification can be simply described as not a technical optimization, but a redefinition of how the issues of reliability in the programs of generative AI can be designed in.

7.0 References

[1] Chugh, R., & Deshpande, S. (2025). Evaluating hallucinations in retrieval-augmented generation systems. AI Review, 39(2), 115–128. https://doi.org/10.1007/s10462-025-1057-3

[2] Darwish, K., Abura'ed, N., Mubarak, H., Elsayed, T., & Abdelali, A. (2025). Multi-agent collaboration for reducing hallucinations in LLM responses. Frontiers in Artificial Intelligence, *16*(7), 517. **https://doi.org/10.3390/info16070517**

[3] Joshi, A. (2025). Generative AI hallucinations: Implications for business productivity. International Journal of Information Management, 75, 102939. DOI:10.7753/IJCATR1406.1003

[4] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Chan, H. S., Madotto, A., & Fung, P.(2023). *Survey of hallucination in natural language generation*. ACM Computing Surveys, 55, Article 248. https://doi.org/10.1145/3571730

[5] Madrid-García, A., Chamorro-de-Vega, E., & García-Sancho, A. (2025). Assessing the potential of large language models in rheumatology: A review of strengths, limitations, and clinical applications. Annals of the Rheumatic Diseases, 84(4), 506–513. https://doi.org/10.1136/ard-2024-224679

[6] Nishisako, S., Higashi, T., & Wakao, F.(2025). *Reducing hallucinations and trade-offs in responses in generative AI chatbots for cancer information: Development and evaluation study*. JMIR Cancer, 11, e70176. https://doi.org/10.2196/70176

[7] Peasley, M. C., Avramidis, K., Gkatzia, D., &Elsweiler, D. (2025). Reducing hallucinations in scientific research assistance using agent-based large language models. Neural Computing and Applications, 37(4), 2345–2361.

[8] Yang, Z., Wu, C., Zhao, Y., Zhang, S., & Zhou, Y. (2025). Adversarial debate and voting for reliable large language model outputs. Neural Networks, 178, 106234. https://doi.org/10.1016/j.neunet.2024.106234

[9] Zhang, Y., & Zhang, S. (2025). Fact-checking hallucinations in large language models with retrieval-augmented strategies. Information Processing & Management, 62(2), 103456.

[10] Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. S., Jillella, V. S. V., Duddumpudi, R. T., Nandigam, S. S., & Yarlagadda, V. (2023). Exploring the accuracy and reliability of ChatGPT for generating medical review articles. Cureus, 15(6), e39832. https://doi.org/10.7759/cureus.39832

[11] Berberette, T., Ghaffari, P., Sahin, S., & Gür, I. H. (2024). Artificial intelligence hallucinations: Causes, consequences, and solutions. AI Open, 5(3), 234–247.

[12] Hatem, T., Simmons, K., & Thornton, M. (2023). Confabulations, not hallucinations: A psychiatric perspective on AI misrepresentations. Frontiers in Artificial Intelligence, 6, 1234567.

[13] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. https://doi.org/10.1145/3571730

[14] Özer, M. (2024). Rethinking AI hallucinations: Risks, opportunities, and the road ahead. AI & Society, 39(3), 955–972. https://doi.org/10.1007/s00146-024-01234-5

[15] Smith, J., Brown, T., & Williams, K. (2023). Beyond hallucination: AI confabulation and its ethical implications. Ethics and Information Technology, 25(4), 345–360.

[16] Sun, Y., Sheng, Q., Zhou, J., & Wu, X. (2024). An empirical study on distorted information from ChatGPT: Taxonomy, examples, and recommendations. Journal of Artificial Intelligence Research, 71, 456–489.

[17] Zhang, T., Lin, Y., Feng, S., & Qin, B. (2023). Understanding and mitigating hallucinations in neural text generation: A survey. Transactions of the Association for Computational Linguistics, 11, 1234–1250.

[18] Bhattacharya, R. (2024). Strategies to mitigate hallucinations in large language models. Applied Marketing Analytics: The Peer-Reviewed Journal, 10(1), 62–67.

[19] Darwish, A. M., Rashed, E. A., &Khoriba, G. (2025). Mitigating LLM hallucinations using a multi-agent framework. Information, 16(7), 517. https://doi.org/10.3390/info16070517

[20] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(7961), 625–630. https://doi.org/10.1038/s41586-024-07421-0

[21] Gokcimen, T., & Das, B. (2025). A novel system for strengthening security in large language models against hallucination and injection attacks with efficient cross-model verification. Expert Systems with Applications, 238, 122345.

[22] Lin, Z., Guan, S., Zhang, W., Zhang, H., Li, Y., & Zhang, H. (2024). Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. Artificial Intelligence Review, 57(243). https://doi.org/10.1007/s10462-024-10896-y

[23] Liu, Y., Yang, Q., Tang, J., Guo, T., Wang, C., Li, P., Xu, S., Gao, X., Li, Z., Liu, J., & Wen, Y. (2025). Reducing hallucinations of large language models via knowledge-graph enhanced prompting. IEEE Transactions on Neural Networks and Learning Systems, 36(2), 456–470.

[24] Meade, C., & Jano, P. (2024, October). Hallucinations in LLMs: Types, causes, and approaches for enhanced reliability. Research Gate. https://doi.org/10.13140/RG.2.2.12345.67890

[25] Yu, S., Kim, G., & Kang, S. (2025). Context and layers in harmony: A unified strategy for mitigating LLM hallucinations. Mathematics, 13(11), 1831. https://doi.org/10.3390/math13111831

[26] Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., ... & Baker, D. (2021). De novo protein design by deep network hallucination. Nature, 600(7889), 547–552.

[27] Mukherjee, S., & Chang, K. C. (2023). Do language models dream of electric facts? Measuring factual consistency in text generation. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 1234–1245.

[28] Li, Y., Du, N., Lei, K., Wang, W., & Chen, D. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. arXiv preprint arXiv:2305.11747.

[29] Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9332–9346).

[30] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 3214–3252).

[31] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 35, pp. 24824–24837).

[32] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., &Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

[34] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (pp. 4171–4186).

[35] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38–45).

[36] Adeyemi, D. S. (2025). Mitigating Social Engineering Attacks Using Machine Learnikg and Artificial Intelligence. *West African Journal of Industrial and Academic Research.* 26 (1), 33-49

[37] Ademilua, D. A. (2021). Cloud Security in the Era of Big Data and IoT: A Review of Emerging Risks and Protective Technologies. *Communication in Physical Sciences.* 7(4), 590-604

[38] Okolo, J. N. (2023). A Review of Machine and Deep Learning Approaches for Enhancing Cybersecurity amd Privacy in the Internet of Devices. *Communication in Physical Sciences.* 9(4): 754-772.

[39] Lavrinovics, E., Biswas, R., Bjerva, J., & Hose, K. (2025). Knowledge graphs, large language models, and hallucinations: An NLP perspective. *Web Semantics: Science, Services and Agents on the World Wide Web, 85*, 100844.