International Transactions on Electrical Engineering and Computer Science
Andra et.al., Vol: 4, No: 4, pp: 227-236, December 2025

Assessing User Confidence and Acceptance of AI Enhanced Through Multi-LLM Consensus Mechanisms

Lohit Sai Andra[1], Sandeep Kumar Srivastava[2], Shruthi Krishna[3], Phaneendra Siddana[1], Karri Sairamakrishna BuchiReddy[4]

Abstract: Large Language Models (LLMs) have transformed AI usage in many aspects, though the issue of user trust and acceptance is a crucial step to successful usage. This paper is a literature review of 16 recent publications (published in 2021-25) about the impact of LLM consensus mechanisms on user trust and the determinants that affect the acceptance of AI systems. The review outlines these structured frameworks as: LLMs-as-Judges, Mixture-of-Agents and Big Loop/Atomization as effective methods to improve the reliability, consistency, transparency, and interpretability of AI outputs. These consensus mechanisms minimize errors, reduce variability and enhance robustness, hence directly enhancing confidence in AI systems by users. Moreover, the analysis describes the most important considerations for user acceptance, such as explainability, transparency, fairness, reduction of bias, and integration into real-life practices. The interplay of these factors and their influence on the adoption and successful utilization of consensus-driven AI can be demonstrated by the applications in healthcare, education, and smart grid systems. Taken together, the results demonstrate the necessity to develop AI systems that are not only technically sound but also in line with the expectations of users. This work will be useful to researchers and practitioners who want to create reliable, user-friendly AI systems and indicate where future research can be done to maximize multi-LLM consensus mechanisms and adapt them to domain-specific situations.

Keywords: Large Language Models, LLM Consensus, AI Acceptance, Multi-LLM Frameworks, Systematic Literature Review

History
Received: 05-12-2024; Revised: 17-08-2025; Accepted: 08-12-2025

✉ Lohit Sai Andra
lohithsaiandra18@gmail.com

[1]Independent Researcher, Senior Software Engineer, SS Software Solutions LLC, Richmond - 23294, USA.
[2]Lead Software Engineer at Capital One, Richmond - 23233, USA.
[3]Independent Researcher, Senior Business Analyst, SS Software Solutions LLC, Reston - 20190, USA.
[4]Independent Researcher, Lead Software Developer, Ford Motor Company, Farmington - 48335, USA.

1. Introduction

LLM-based artificial intelligence (AI) systems have quickly become popular in the mainstream in areas like information retrieval, education, software development, healthcare, and decision support. These systems have natural language reasoning, content generation and interactive problem-solving abilities like never before in history [1]. Nevertheless, even with the increase in their use, single-model LLMs often have drawbacks, including hallucinations, erratic reasoning, and the unreliability of output changes. These difficulties have heightened doubts regarding the reliability, transparency, and trustworthiness of AI-generated answers, particularly within applications where the factual accuracy is crucial and the user's

https://doi.org/10.62760/iteecs.4.4.2025.131

trust is significant. Consequently, scientists and engineers have started to experiment with phenomena like multi-LLM consensus mechanisms systems in which multiple language models cooperate, vote, or cross-validate the output of one another, to generate a more accurate and more stable output [2].

Multi-LLM consensus can be a successful move on the way to AI reliability improvement. Consensus workflows can eliminate hallucinations, limit bias, and enhance consistency by combining responses between models with varying training pipelines or contextual strengths, or with varied ways of thinking. More recent studies on ensemble LLMs, cross-model agreement scoring, self-consistency prompting, and model voting models have demonstrated verifiably better factual accuracy and level of reasoning. The technical progress of the consensus-based LLM systems is heavily documented, whereas little is known regarding the perception of the end user of the systems. Nor should performance metrics dictate trust and acceptance they should be based on psychological, social and experiential metrics, such as perceived transparency, perceived reliability, familiarity of users with the AI systems, and clarity of communication of consensus mechanisms [3-4].

Despite the significant body of literature on trust in AI in the context of human-computer interaction (HCI), explainable AI (XAI), and technology adoption literature, there is little literature on the topic of user trust in multi-LLM consensus systems. The literature emphasizes the trust in single-model AI, accuracy increases through ensembling, or technical reliability. Yet, such studies do not often explore the question of whether users do have trust in responses produced by several cooperating models or whether they are aware and supportive of consensus as a valuable reliability characteristic [5]. As an example, users can interpret consensus as a demonstration of strength, or vice versa, to think of it as an overly complex or opaque process that lowers confidence. Besides, other variables that affect user acceptance (perceived usefulness, ease-of-interaction, transparency of a consensus process, and perceived fairness) have not been empirically proven in the environment of multi-LLM. The need to identify the gaps between AI reliability mechanisms and user-centered evaluation frameworks is illustrated by this gap [6-7].

In order to counter these limitations, the study will seek to determine the effects of multi-LLM consensus on the level of user trust and the determinants that affect user acceptance of this model. In particular, the study examines the trust of users in the responses generated by AI with the assistance of the model agreement, the significance of perceived reliability, and the psychological and experiential factors according to which confidence is formed [8-10]. A guide to the study is presented in the following research questions:

***RQ1*:** How does LLM consensus influence user trust?

***RQ2*:** What factors shape user acceptance of AI systems using LLM consensus?

The study makes three contributions to the developing field of literature on human-AI interaction. To begin with, it provides a contribution to the existing theories and research on trust and acceptance by adding consensus-based reasoning to the list of factors that determine user perception. Second, it provides empirical data about the effect of consensus on user confidence and decision-making behaviour, which is a research gap in the area of AI. Third, it suggests a conceptual model that combines AI reliability in consensus and the concept of user trust and acceptance, providing a basis for future HCI and AI design research. Altogether, this piece of work can inform designers, AI developers, and policymakers to design more reliable, transparent, and user-consistent AI systems.

The rest of this paper is structured in the following manner: Section 2 covers the methodology, such as research design, data collection and analysis process. In section 3, the results are discussed and the implications of the results are interpreted. Section 4 summarizes the paper by providing important findings, limitations and future research directions.

2. Materials and Methods

The proposed work is based on the Systematic Literature Review (SLR) methodology to synthesize the research results on user trust, confidence, and acceptance of AI systems that have been improved with the help of multi-LLM consensus mechanisms. The review was conducted in accordance with

PRISMA to ensure transparency, reproducibility, and methodological rigor.

2.1 Review Protocol and Objectives

An organized review process was designed, which included research questions, search strategy, databases, screening process, and inclusion and exclusion criteria and data extraction process. The SLR had a goal of finding and assessing research that covers:

- The influence of LLM consensus on user trust
- The factors shaping user acceptance of AI systems using consensus mechanisms

The protocol guided all stages of search, screening, and analysis as mentioned in Fig. 1.

Fig. 1: Protocol Guided All Stages

2.2 Databases and Search Strategy

Two major databases were selected for their relevance to AI, computer science, and user-centric technology research:

- Google Scholar
- Science Direct

Searches were conducted between 2021–2025 to capture recent advancements in LLM technology and consensus-based AI reliability research.

The Boolean search string used was:

("multi-LLM" OR "multiple language models" OR "LLM consensus" OR "AI consensus mechanism" OR "AI reliability mechanisms") AND ("user trust" OR "user confidence" OR "trust in AI" OR "AI acceptance)

2.3 Inclusion and Exclusion Criteria

Inclusion Criteria

- Publications from 2021–2025
- Journal articles or conference papers that are peer reviewed.
- Research on the topic of LLCMs, AI consensus, ensemble reasoning, or AI trustworthiness.
- Research on user trust, user confidence, acceptance or adoption factors.

Exclusion Criteria

- Other (non-academic) materials (blogs, editorials, commentaries)
- Non peer reviewed material
- Research that is not related to LLMs or human perception.
- Articles that are not adequately detailed in their methods.

2.4 Study Selection Process

Screening was conducted in three stages:

- ***Initial screening:*** based on publication year (2021–2025)
- ***Title and abstract screening:*** to remove irrelevant works
- ***Full-text review:*** to determine final eligibility

A total of 240 articles were initially retrieved. After applying all filters, 16 studies were included in the final analysis. Fig. 2, presents a detailed breakdown of the screening stages across each database.

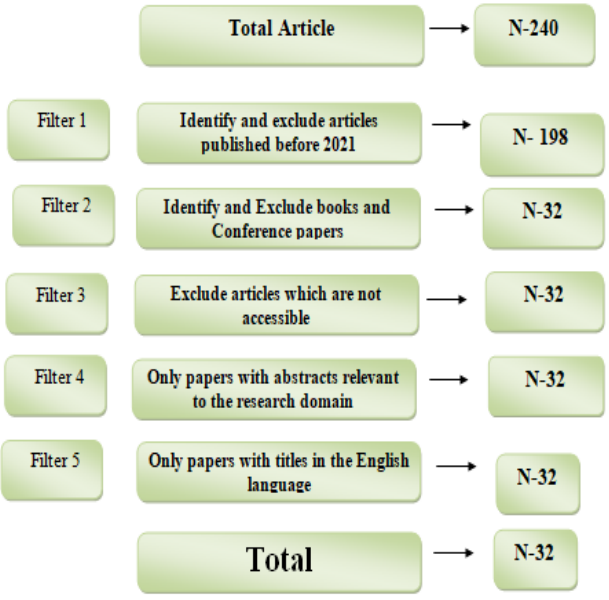

Fig. 2: PRISMA Framework

2.5 Quality Assessment

Quality assessment was done strictly to determine the credibility and reliability of studies used in this review. The predefined criteria were used to evaluate each article in regard to its methodological clarity, the strength of the experimental design, the clarity of the data and model use, the validity of the evaluation measures, and applicability in answering the research questions. Research was also reviewed based on the completeness of reporting, reproducibility of findings and the degree to which limitations and biases were recognized. Peer-reviewed journal publications and fully documented conference papers were assigned higher weights, whereas preprints were subject to critical evaluation of methodological soundness. On the whole, the chosen papers were of good technical consistency and great conceptual relevance, which could be considered a good base on the analysis of the impact of LLM consensus mechanism on user trust and its influences on user acceptance in AI systems.

Table. 1: Quality Assessment

Ref No.	Relevancy	Findings	Contributions	Overall Quality
[1]	Relevant	Defines the frustration in sentiment analysis with LLMs, the Model Variability Problem (MVP), demonstrating how the stochastic inference, prompt sensitivity and training biases lead to inconsistency. Underlines the effect of the instability on user trust and reliability.	Gives a systematic study of variability reasons, justifies the position of temperature in randomness, gives mitigation methods, and underscores the clarity and reliability of LLM results.	High
[2]	Highly Relevant	Introduces the LLMs-as-Judges paradigm, which describes how it works, how it is done, how it can be meta-evaluated, and what weaknesses it has as a way of reliably judging outputs. Focuses on the generalizability, interpretability and strength of multi-LLM evaluation systems.	Gives a thorough overview of multi-LLM evaluation and presents information on how to create reliable, interpretable and consensus-driven systems of LLM. It is a basis on which user trust and acceptance can be improved.	High
[3]	Relevant	Discusses the reviews of AI agents using the LLM with the focus on efficiency, consistency, transparency, adaptive reasoning, and collaborative workflows. Talks about issues like quality assurance, mitigation of bias and scalability.	Gives a systematic view of the annotation systems based on the LLM, including the reliability mechanisms, transparency, as well as the guidelines in designing trustworthy AI workflows.	High
[4]	Relevant	Discusses multimodal and longitudinal data. Reviews XAI systems with an emphasis on clinical usability, interpretability, uncertainty management, and transparency.	Introduces the XAI orchestrator framework, which focuses on the adaptive, interactive, and uncertainty-aware design to increase user trust and confidence.	High
[5]	Relevant	Techniques of survey alignment, value alignment, commonsense reasoning, factuality, and robustness in neural language models focus on the work aimed at improving the reliability, fairness and trustworthiness.	Gives a thorough account of the human-centered approaches to the LLM, and resources, giving the background on how to create trustworthy and user-friendly AI systems.	High
[6]	Highly Relevant	Survey LLM-based recommendation systems, such as hybrid models, RAG, assessment techniques, fairness and LLM-as-a-Judge models. Lays stress on reliability, trustworthiness, and reduction of bias and hallucination.	Gives recommendations on how to design trustworthy, interpretable and user-confidence-based systems in which an LLM is used to make recommendations, points out evaluation strategies and gaps in research.	High
[7]	Relevant	Review of the use of LLM in healthcare, emphasizing the quality of outcomes in diagnosis, practicality, reliability, transparency, ethical issues, and human supervision.	Generalizes the results of several systematic reviews and offers comments on the issues and possible utility of LLMs in practice, especially user trust and acceptance.	Medium
[8]	Relevant	Assesses the combination of LLMs with visual analytics with a focus on interpretability, interactive exploration, visualization, ethical implications, and aspects of trust.	Brings to light the understanding of how to design user-facing LLM-assisted analytics systems that can increase reliability, transparency, and user trust.	High

[9]	Highly Relevant	Presents the framework of Big Loop and Atomization to combine the tool learning, multi-agent systems and the collaboration with LLM, with focus on the coordination, optimization and reliability.	Provides a systematic viewpoint of multi-LLM collaboration, which offers practical information about how to build upon trust, resilience, and user trust in AI-based consensus systems.	High
[10]	Highly Relevant	Surveys LLM-based semantic integration (semantic join) with heterogeneous sources based on in-context learning, RAG, reliability and automation.	Offers a holistic system of enhancing multi-LLM thinking and dependability, informing reliable and user-confidence-building AI system design.	High
[11]	Relevant	Surveys establish narrative visualization on the basis of foundation models, emphasizing on insight extraction, multimodal integration and interactive interaction.	Huge offers insights into the design of interpretable and user-trust-friendly AI systems that increase reliability and involvement in the visual storytelling application.	High
[12]	Highly Relevant	Surveys of AI-based student-centred learning frameworks presented on the platform of multi-LLM consensus, showing an increased level of reliability, decision-making, and end-user trust.	Shows how LLM consensus mechanisms can be applied in practice in the field of education, which gives information about the factors that affect user confidence in consensus-driven AI systems and their acceptance.	High
[13]	Highly Relevant	Reviews NLP-based mental health The idea of reviews NLP-based mental health research, in which custom LLM modules were used to automate the review procedures with much emphasis on dataset handling, reproducibility, and efficiency of the workflow.	Visions is an example of a practical application of the LLM consensus mechanisms to improve reliability, transparency and user faith in the research pipeline.	High
[14]	Relevant	Discusses LLM-based Generative Agent-Based Models (GABMs) that emphasize predictive power, simulation of human behavior, and such issues as prompt sensitivity and hallucinations.	Presents information on the implications of the actions of the LLM agent regarding user trust and reliability, which would be applicable in understanding the aspects of influencing user confidence and acceptance.	High
[15]	Highly Relevant	Introduces the Mixture-of-Agents (MoA) system, a system that incorporates multiple specialized LLMs to improve reasoning, flexibility and reliability.	Shows that organized multi-LLM cooperation makes AI systems used in decision-making more accurate, trustworthy and more confident by users.	High
[16]	Relevant	Surveys AI/ML systems in smart grids, such as LLM systems and multi-agent systems, with a focus on context-based reasoning, reliability, and coordination.	Shines light on elements that have a secondary impact on the user trust and acceptance of consensus-driven AI, especially in complex, distributed systems.	Medium

2.6 Data Extraction and Synthesis

Data from 16 selected studies were systematically extracted, taking into account relevance, key findings, contributions, and overall quality, and synthesized to determine patterns in multi-LLM consensus mechanisms, reliability, user trust, and acceptance of AI systems.

- Article [1]: Recognizes the Model Variability Problem (MVP) of the sentiment analysis of LLMs and emphasizes the factors that contribute to inconsistency, their effects on user trust, and how the latter can be improved to increase the explainability and reliability of the results.
- Article [2]: Introduces the concept of the LLMs-as-Judges paradigm, the main points in which are the multi-LLM evaluation paradigms to enhance the interpretability, consistency, and user confidence in AI results.
- Article [3]: presents reviews of AI agents powered by LLM and shows that these agents lead to increased efficiency, transparency, and reliability of collaborative AI workflows.
- Article [4]: In takes a look at XAI systems in multimodal and longitudinal data, and suggests adaptive and uncertainty-aware frameworks to increase clinical usability and user trust.
- Article [5]: Surveys alignment and human-centered LLM methodology, providing advice of making AI systems more fair, robust, and trustworthy.

- Article [6]: surveys Artificial intelligence DLM-based recommendation systems, focusing on hybrid systems, evaluation, and fairness as an attempt to enhance their reliability and user acceptance.
- Article [7]: Provides a meta-review of the applications of LLM in the healthcare sector, summarizing the findings on the accuracy of the results, usability, trust, and ethical implications to the real-world application of the findings.
- Article [8]: Discusses the idea of the combination of LLMs with visual analytics and the need to focus on the interpretability, interactive exploration, and transparency to establish user trust.
- Article [9]: It presents the framework of the Big Loop and Atomization, where the multi-LLM work is coordinated to improve the reliability, trust, and confidence of users in AI systems.
- Article [10]: Surveys the semantic integration (semantic join) based on LLM, introductions of in-context learning and RAG techniques to enhance the reasoning and reliability of multi-LLM.
- Article [11]: Surveys foundation models of narrative visualization, including the multimodal insight extraction and interaction to boost trust and interpretability of AI results.
- Article [12]: Relates to AI-based student-centered learning systems based on multi-LLM consensus, which involve enhanced decision-making, reliability, and trust in the system by the users of educational applications.
- Article [13]: Having reviewed NLP-based mental health studies with custom LLM modules, the article [13] focuses on the reproducibility, efficiency, and trust-enhancing mechanisms of AI-assisted research.
- Article [14]: Generative Agent-Based Models Driven by LLM that discusses the ability to predict, recreate the behavior of humans, and the consequences of these findings on user trust and AI acceptance.
- Article [15]: introduces the Mixture-of-Agents (MoA) concept, showing the organized multi-LLM cooperation to improve the reasoning, adaptiveness and reliability of AI systems.
- Article [16]: Surveys AI/ML use in intelligent grids with LLM architecture and multi-agent systems and points out factors that determine reliability and indirectly the trust and acceptance of users.

3 Results and Discussion

The chapter reports the results of the systematic literature review and extracts the knowledge of the chosen studies. The discussion is aimed at the interpretation of the multi-LLM consensus mechanism's effect on the system reliability, user trust, and acceptance of AI in various fields of application. They have grouped the findings in a thematic manner to bring out the methodological trends, emerging frameworks and the common challenges that have been prevailing in the literature. This section is an overview of the existing research environment by using quantitative tendencies and qualitative insights and forms the basis of further discussion in the other subsections.

3.1 Overview of Selected Studies

This systematic review included 16 studies published in the last five years (2021-2025) and was found through two major academic databases: Google Scholar (11 studies) and ScienceDirect (5 studies). These works cover a very broad range of fields, such as sentiment analysis, AI evaluation and alignment frameworks, healthcare diagnostics, educational avenue, smart grid systems, mental health research, visual analytics, and generative agent-based models, which are all indicative of the fast-growing field of applications of LLM. Table. 1, will give a summary concerning the sources of study, their relevance scores and areas of thematic focus.

Out of the reviewed literature, 7 studies were deemed as highly relevant and provided direct information on the subject of LLM consensus, user trust formation, and multi-agent collaboration frameworks. A few other 9 studies were marked as relevant that added support evidence to explainability, fairness, reliability, human-centered design, and a

real-world integration of AI systems. Taken together, the studies signify various yet intertwined areas of focus, such as the use of LLM consensus mechanisms, the calibration of trust, AI reliability and robustness, multi-LLM coordination architecture, transparency XAI techniques, and the use of LLM in recommendation and decision-support systems. The extensive disciplinary dissemination and high relevance scores indicate the increased significance of learning the impact of multi-LLM consensus on user trust and acceptance in diverse sectors and application environments.

3.2 Answer to Research Questions

The systematic review of 16 studies provides clear insights into the research questions as mentioned in Table. 2.

RQ1: How does LLM consensus influence user trust?

The evidence provided in [1], [2], [6], [9], [10], [11], [12], and [15] all point to the fact that the consensus mechanisms of LLM have an impressive effect on the enhancement of reliability, consistency, and transparency as the key characteristics that directly impact the user loyalty towards the AI systems and have a beneficial effect on them.

Table. 2: Research questions and supporting studies

Research Question	Articles
RQ1: How does LLM consensus influence user trust?	[1], [2], [6], [9], [12], [15], [10],[11]
RQ2: What factors shape user acceptance of AI systems using LLM consensus?	[3], [4], [5], [7], [8], [12], [13], [16], [14]

One such case is the LLMs-as-Judges model [2], where it is shown that a combination of many models can be trusted to evaluate or predict far better (regarded as less random, less model-specific, and more consistent and reliable) and more reliably. Likewise, other models such as Mixture-of-Agents (MoA) [15] and Big Loop/Atomization [9] also focus on how the richness of an agent interaction can be refined through agent coordination and reasoning, error reduction and predictable behavior, and reducing ambiguity in model behavior. Additional evidences of the results of the searching in the sphere of the LLM-based recommendation systems [6], semantic integration and multi-LLM reasoning [10], and narrative visualization models [11] suggest that the results of the consensus-based decisions are considered to be more credible, specifically, those that contribute to raising the interpretability and reducing the number of hallucinations. In addition, thematic use in learning scenarios [12] proves consensus mechanisms not only to make accuracy better but also user confidence, because transparent, verifiable decisions are provided. Altogether, the findings can be summarized to suggest that multi-LLM consensus techniques have significant positive impact on the perceived credibility of AI systems, as they are more reliable, interpretable, and user-congruent.

RQ2: What factors shape user acceptance of AI systems using LLM consensus?

Articles [3], [4], [5], [7], [8], [12], [13], [14], and [16] indicate that user acceptance of AI systems utilizing AI consensus based on LLM is influenced by the combination of cognitive, technical and contextual variables. One core factor is explainability and transparency, which are highly prioritized when designing XAI frameworks [4] and allow users to have an idea of how multi-LLM decisions are created and why certain outputs are generated. Even fairness, the mitigation of bias, and the general performance of a system, which have been emphasized in the human-centric alignment research [5], are crucial factors in whether users can trust an AI system and perceive it as morally grounded to human values. Data annotation systems [3] and healthcare meta-reviews [7] evidence indicate that users tend to be more accepting of AI systems when they are demonstrated to be consistent, clear, and have a low cognitive burden. Further validation of the fact that interpretability and interactive transparency enhance user trust, particularly in data-intensive setting comes in visual analytics studies [8]. Hands-on workflow integration and empirical utility are also essential: in educational contexts [12], in research automation in mental health research [13], in smart grid decision-support systems [16], seamless workflow integration is important to improve acceptance significantly. Furthermore, the experience of generative agent-based modeling studies [14] shows, hallucination reduction and stable behavior of agents are the factors that define user buy-in.

3.3 Implications for User Trust and Acceptance

The synthesis of the chosen studies proves that multi-LLM consensus mechanisms are an important and consistent feature of enhancing user trust and the general acceptance of AI systems. In many different areas, these mechanisms enhance reliability, transparency, and predictability, which are core concepts of trustworthy AI.

First, reliability is increased by reducing errors and improving consistency because a consensus between many models reduces stochastic variability and alleviates the weaknesses of a given LLM. This enhances the predictability and repeatability of outputs, which is vital in high-stakes situations in which users are unwilling to take risks.

Second, interpretable consensus mechanisms and the structured outputs of reasoning are used to enhance transparency and allow users to understand the way decisions are constructed. Research indicates that users tend to trust and adopt an AI system when they can follow the path and processes used to form the conclusions of that specific system, especially in consensus-based systems.

Third, user confidence rises as the reproducibility, robustness and less sensitivity to timing or model-related biases of decision-making have risen. The consensus methods reveal the possibility of obtaining balanced, diversified, and harmonized outcomes, which leads to enhanced perceived system reliability.

Collectively, integrated frameworks incorporating evaluation mechanisms, explainable AI (XAI) concept, and multi-agent orchestration become effective solutions to decrease the overall variability, uncertainty management, and the trust ecosystem around AI systems. Such results highlight the need to create AI technologies that are user-centered in terms of trust building.

3.4 Recommendations for Future Research

Based on the results of this review, certain significant recommendations are advanced as to future research to enhance the andragogy, dependability and feasibility of multi-LLM consensus systems:

- ***Establish Generalized Standards and Assessment procedures:*** Development of domain-independent benchmarks that can be used to measure the performance of multi-LLM consensus models should be prioritized in future product development activities. This involves coming up with measures that measure accuracy, consistency, robustness, and computing efficiency. The standardized evaluation systems would allow researchers to evaluate various consensus methods and detect performance flaws, as well as facilitate repeatable experimentation.
- ***Explore Contextual User Perception, Trust and Acceptance:*** More research needs to be conducted to understand how various groups of users, including students, professionals, clinicians or the general consumer, view consensus-based AI systems. Studies must also be conducted to understand the effects that issues like the design of UI, clarity in explaining, familiarity with the domain and perceived autonomy of the system have on user trust, satisfaction and adoption of such a tool in making their day-to-day decisions.
- ***Inculcate Ethical, Fairness and Transparency:*** With the increase in the use of multi-LLM consensus systems, ethical issues gain greater significance. Fairness, bias detection constructs, and clear decision-rationalization processes should be included in future studies.

Additional Directions

- Multi-LLM systems should be expanded in real-world applications in healthcare, education, finance, industrial operation, and public decision-support.
- More cost-effective orchestration strategies could be considered to decrease the computational burden without compromising the precision or strength.
- Exploring human-in-the-loop feedback with hybrid methods to enhance the reliability and accountability of the LLM consensus.

4 Conclusion

This systemic literature review highlights the importance of the use of the LLM consensus

mechanisms in enhancing user trust and acceptance in the current AI systems. The results indicate that multi-LLM strategies, including LLMs-as-Judges, Mixture-of-Agents and Big Loop/Atomization are far more effective in ensuring the reliability, stability and interpretability of model outputs, including through a reduction in stochastic variability, the alleviation of hallucinations, and the promotion of consistent reasoning across models. These advancements are directly related to the rise in user confidence since users are likely to trust systems that give consistent, reproducible and transparent results. In addition to technical reliability, the review reports a range of human-oriented factors, most notably explainability, decision processes transparency, assurance of fairness, mitigation of bias, and effortless integration with existing workflows as the key factors that impact the user acceptance. In a wide variety of application fields, such as healthcare diagnostics, education, visual analytics, recommendation systems, and smart grid management, consensus-based mechanisms have been demonstrated to increase the robustness, usability and ethical appropriateness of the system. Taken together, all these results indicate that the use of multi-LLM consensus is an encouraging direction toward the development of trustworthy, user-friendly, and socially responsible AI ecosystems.

4.1 Practical Implication

The review results have a number of implications that should be considered when designing and developing a language model (LLM)-based application, or using LLMs in the implementation of a product. In particular, implementing a consensus mechanism with multiple LLMs is an effective way to increase the reliability of an LLM-based application. This is particularly beneficial in applications that operate in areas where high levels of reliability are required, such as healthcare, finance, the legal system, and critical infrastructure management. Additionally, an LLM-based application will require greater levels of transparency, explainability, and reduction of hallucinations, as these elements will lead to greater levels of user trust in an LLM-based application and ultimately greater success in the marketplace, whether that be for consumers or enterprises. Finally, the evidence presented demonstrates that LLMs with a consensus approach can serve as a means for multi-tiered verification of a decision to move forward with the application, as well as ensuring that the decision is made within the parameters of ethical compliance. Additionally, the use of these types of architectures will enable enterprises to ensure that any AI tool they deploy meets current and emerging regulatory requirements regarding fairness, accountability, and responsible use of AI.

4.2 Limitation

In light of the rapidly evolving research area of multi-LLM consensus, there are important limitations to this review. First, the vast majority of studies referenced herein are early-stage or prototype studies; hence, none of the results can be generalized for large-scale applications. Second, due to the diverse nature of the types of experiments conducted, reporting standards used, and metrics employed to evaluate the results in the various studies, it is difficult to directly compare them to one another; therefore, most of the findings contained within will not be able to provide clear insights into how LLMs might function within a multi-LLM consensus framework. Finally, because of the newness of the area, long-term user perceived and real-world performances have not yet been evaluated comprehensively enough to develop a sufficient body of evidence to support multi-LLM consensus capability. Thus, additional longitudinal and cross-domain, standardized studies are needed to create a solid evidence base for multi-LLM consensus.

Conflict of Interest

The authors declared “No conflict of Interest”

References

[1] D. H. Poyatos, C. P. González, C. Zuheros, A. H. Poyatos, V. Tejedor, F. Herrera, R. Montes “An overview of model uncertainty and variability in LLM-based sentiment analysis. Challenges, mitigation strategies and the role of explainability”, *arXiv.org*, 2025. https://arxiv.org/abs/2504.04462

[2] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu “LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods”, *arXiv.org*, 2024. https://arxiv.org/abs/2412.05579

[3] M. M. Karim, S. Khan, D. H. Van, X. Liu, C. Wang and Q. Qu "Transforming Data Annotation with AI Agents: A Review of Architectures, Reasoning, Applications, and Impact," *Future Internet*, Vol. 17, No. 8, art. no. 353, 2025. https://doi.org/10.3390/fi17080353

[4] A. P. D. Mortanges, H. Luo, S. Z. Shu, A. Kamath, Y. Suter, M. Shelan, A. Pöllinger & M. Reyes "Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging", *npj Digital Medicine*, Vol. 7, art. no. 195, 2024. https://doi.org/10.1038/s41746-024-01190-w

[5] S. Sicari, J. F. Cevallos, A. Rizzardi, and A. Coen-Porisini, "Open-Ethical AI: Advancements in Open Source Human Centric Neural Language Models", *ACM Computing Surveys*, Vol. 57, No. 4, pp. 1-47, 2024. https://doi.org/10.1145/3703454

[6] D. Nawara and R. Kashef "A Comprehensive Survey on LLM-Powered Recommender Systems: From Discriminative, Generative to Multi-Modal Paradigms", *IEEE Access*, Vol. 13, pp. 145772–145798, 2025. https://doi.org/10.1109/access.2025.3599832

[7] A. Triantafyllidis, S. Segkouli, S. Kokkas, A. Alexiadis, E. Lithoxoidou, G. Manias, A. Antoniades, K. Votis, D. Tzovaras "Large Language Models for Cardiovascular Disease, Cancer, and Mental Disorders: A Review of Systematic Reviews", *Preprints.org*, 2025. https://doi.org/10.20944/preprints202510.2480.v1

[8] N. S. Agarwal and S. S. Kumar "A Review on Large Language Models for Visual Analytics", *arXiv.org*, 2025. https://arxiv.org/abs/2503.15176

[9] Z. Hu, Y. Huang, J. Feng, and C. Deng, "Big Loop and Atomization: A Holistic Review on the Expansion Capabilities of Large Language Models", *Applied Sciences*, Vol. 15, No. 17, art. no. 9466, 2025. https://doi.org/10.3390/app15179466

[10] K. Hong and Y. Park "Large Language Models for Semantic Join: A Comprehensive Survey", *IEEE Access*, Vol. 13, pp. 184478-184493, 2025. https://doi.org/10.1109/access.2025.3625753

[11] Y. He, K. Xu, S. Cao, Y. Shi, Q. Chen, and N. Cao, "Leveraging Foundation Models for Crafting Narrative Visualization: A Survey," *IEEE Transactions on Visualization and Computer Graphics*, Vol. 31, No. 10, pp. 9303–9323, 2025. https://doi.org/10.1109/tvcg.2025.3542504

[12] H. Albasry, E. Carmona-Cejudo, A. Rauf, and D. Chen "A systematically derived AI-based framework for student-centered learning in higher education", *Social Sciences & Humanities Open*, Vol. 12, art. no. 102085, 2025. https://doi.org/10.1016/j.ssaho.2025.102085

[13] D. A. Scherbakov, N. C. Hubig, L. A. Lenert, A. V. Alekseyenko, and J. S. Obeid "Natural Language Processing and Social Determinants of Health in Mental Health Research: An Artificial Intelligence Assisted Scoping Review", *JMIR Mental Health*, Vol. 12, 2024. https://doi.org/10.2196/67192

[14] Y. Lu, A. Aleta, C. Du, L. Shi, and Y. Moreno "LLMs and Generative Agent-Based Models for Complex Systems Research", *Physics of Life Reviews*, Vol. 51, pp. 283-293, 2024. https://doi.org/10.1016/j.plrev.2024.10.013

[15] S. S. Bavirthi, D. P. Sreya, and T. Poojitha "Comparative analysis of Mixture-of-Agents models for natural language inference with ANLI data", *Natural Language Processing Journal*, Vol. 11, art. no. 100140, 2025. https://doi.org/10.1016/j.nlp.2025.100140

[16] Y. M. Banad, S. S. Sharif, and Z. Rezaei, "Artificial intelligence and machine learning for smart grids: from foundational paradigms to emerging technologies with digital twin and large language model driven intelligence", *Energy Conversion and Management: X*, Vol. 28, art. no. 101329, 2025. https://doi.org/10.1016/j.ecmx.2025.101329

***