Date of first submission	2026-06-03
Date of acceptance of article for publication after review	2026-06-25
Date of publication/publication	2026-07-25

METAMORPHIC TESTING AND AI AGENTS IN ENSURING THE RELIABILITY OF CORPORATE SOFTWARE

Mokhammed Ali Patvari
ORCID: https://orcid.org/0009-0005-7354-082X
Senior QA Automation Engineer, ECMC Group,
111 Washington Ave S, Ste 1400, Minneapolis, United States

Abstract. The article is devoted to studying the importance of AI agents in ensuring the reliability of corporate software. The purpose of the article is to substantiate the role of metamorphic testing and AI agents in ensuring the reliability, semantic consistency and security of corporate software based on artificial intelligence. The research process used methods of analysis and synthesis to generalize scientific approaches to testing AI systems; the comparative method was used to compare traditional QA, Metamorphic Testing and the agent approach. The results of the study showed that the reliability of corporate software based on AI cannot be ensured only by traditional QA approaches, since LLM, RAG systems, chat bots, recommendation services and AI agents operate in a non-deterministic mode and can change the answer depending on the formulation of the query. It has been established that Metamorphic Testing is an appropriate method for testing such systems, since it allows us to evaluate not a single result, but the constancy of the relationship between responses after a controlled transformation of the input data: paraphrasing, changing the format, rating scale, word order, or adding a small amount of noise. It has been proven that if the content of the query does not change, then the main fact, conclusion, class, recommendation, or logic of the response should remain stable, and their significant change indicates semantic inconsistency, the risk of hallucination, or weak robustness of the model. On this basis, the author's concept of AI-Agentic Testing is proposed, in which the AI-agent automates the generation of follow-up inputs, checking metamorphic relations, detecting unstable responses, hallucinations, logical contradictions, and potential security breaches in corporate AI systems. This approach moves QA from static test scenarios to dynamic control of semantic consistency, reliability, and security of AI solutions in a corporate environment. Its practical value lies in the fact that it can be used as a basis for testing LLM, RAG systems, AI agents and enterprise AI workflows before their integration into critical business processes.

Keywords: metamorphic testing, AI agents, RAG systems, LLM.

Introduction

Today, testing AI systems is becoming a separate quality assurance problem, since the traditional logic of testing through a predetermined expected result does not always work for non-deterministic models. Chatbots, recommendation systems and AI agents can give different answers to queries that are close in content, so the key is not the literal coincidence of the result, but its semantic consistency, validity and security. One of the methods that allows partially solving this problem is Metamorphic Testing, since it transfers the test from the level of a separate answer to the level of the relationship between the answers after a controlled change in the query. If the query content is preserved, then the main conclusion of the system should remain stable.

ISSN: 2787-9364

Literature Review

The issue of using metamorphic testing and AI agents in ensuring the reliability of corporate software is not widely covered in the scientific literature as a holistic research problem. Despite this, individual components of this topic are considered by scientists and expert organizations. In particular, the importance of responsible artificial intelligence systems, the problems of trust in AI and the need for testing were studied by G.M. Asaftei, R. Roberts, A. Sticha [1], who showed the connection between investments in responsible AI and the level of benefits received by organizations from the use of AI; C. Autio, M. Barrett, J. Newman, J. Nunn, A. Oprea, L. Pinelis, A. Rice, E. Tabassi and A. Vassilev [2], who revealed the risks of generative AI within the NIST AI Risk Management Framework profile; Y. Bengio [3], who systematized the global risks of the development of advanced AI systems; and the OWASP Foundation [11], which substantiated AI testing as the basis for trust, security, and controllability of AI systems.

Regarding hallucination detection, response quality assessment, and testing of RAG systems, this issue was investigated by S. Es, J. James, L. Espinosa-Anke, and S. Schockaert [5], who proposed the RAGAs approach for automated evaluation of retrieval-augmented generation based on the criteria of faithfulness, answer relevance, and context quality; A. Pesaranghader and E. Li [12], who considered the detection stage as an operational stage of hallucination management and identified approaches based on uncertainty estimation, semantic entropy, intrinsic consistency, and context-based fact-checking; W. Zhang and J. Zhang [18], who showed the role of prompt engineering in reducing hallucinations in retrieval-augmented LLMs; A.T. Kalai, O. Nachum, S.S. Vempala and E. Zhang [7], who proved the statistical nature of hallucinations and explained why they can persist even after post-training; and C. Sok, D. Luz and Y. Haddam [15], who proposed MetaRAG as a metamorphic approach to detecting hallucinations in RAG systems through the verification of atomic factoids, synonym mutations and antonym mutations.

The issue of Metamorphic Testing as a method of testing systems in the absence of a classical test oracle was investigated by S. Segura, G. Fraser, A.B. Sanchez and A. Ruiz-Cortés [13], who systematized the theoretical foundations of metamorphic testing and showed its importance for complex software systems; T. Kanstren [8], who revealed the practical logic of transforming seed input into follow-up input and verifying metamorphic relations in machine-learning based systems; X. Xie, J.W.K. Ho, C. Murphy, G. Kaiser, B. Xu and T.Y. Chen [16], who applied Metamorphic Testing to test and validate supervised machine learning classifiers; S. Cho, V. Terragni, S. Kang, T. Ahmed and S. Yoo [4], who investigated metamorphic testing of large language models in NLP tasks; J. Yang, D. Chen, Y. Sun, R. Li, Z. Feng and W. Peng [17], who considered semantic consistency as a separate problem of large language models; M. Khirbat, Y. Ren, P. Castells and M. Sanderson [9], who tested ChatGPT as a recommender system and showed instability of recommendations under semantically equivalent prompt changes; E. Manino, J. Rozanova and R. Morante [10], who applied metamorphic relations to analyze systematicity, compositionality and transitivity in deep NLP models.

Problem Statement

The purpose of the article is to substantiate the role of metamorphic testing and AI agents in ensuring the reliability, semantic consistency and security of corporate software based on artificial intelligence. To achieve the goal, the following tasks will be performed during the study: to determine the significance of responsible artificial intelligence systems and testing for ensuring trust in corporate AI solutions; to reveal the essence, features and applied value of Metamorphic Testing in checking the stability of AI system responses when changing the form of the request without changing its content; to form the author's concept of AI-Agentic Testing as an adaptive model of testing corporate software using AI agents.

ISSN: 2787-9364

Methodology and Materials

The methodological basis of the study is a systematic approach to the analysis of AI system testing in corporate software, which made it possible to consider responsible artificial intelligence, hallucination detection, RAG evaluation, Metamorphic Testing and AI-Agentic Testing as interconnected elements of ensuring reliability. The research materials were scientific works devoted to metamorphic testing, semantic consistency of LLM, hallucination detection, RAG system evaluation and ML model testing, as well as expert reports, regulatory and recommendation documents and analytical sources on responsible AI, enterprise AI and generative AI risks. The research used methods of analysis and synthesis to generalize scientific approaches to AI system testing; comparative method – to compare traditional QA, Metamorphic Testing and agent approach; logical generalization method – to determine the advantages and limitations of metamorphic testing; tabular method – to systematize the expected behavior of AI when changing the query form and the advantages and disadvantages of Metamorphic Testing; conceptual modeling method — to form the author's concept of AI-Agentic Testing as an adaptive model for checking the stability, security, and semantic consistency of corporate AI systems.

Results and Discussion

In 2026, the development of artificial intelligence is characterized not only by the rapid growth of the capabilities of models, but also by the accumulation of risks of their use. According to the 2026 AI Index report, generative AI has reached almost 53% of the level of distribution among the population within three years, and the organizational implementation of AI has increased to 88% [6].

This means that AI has already moved from the phase of experimental use to the phase of mass application. At the same time, evaluation, testing, and control mechanisms are not keeping up with the pace of such distribution. For the corporate sector, this problem has a practical dimension. According to Reuters, the forecast for the global AI market has been increased to over $ 4 trillion. USA, which is associated with the growth of enterprise AI, agent-based workflows and task automation [14]. By 2030, this could shift competition from simply comparing models to verifying their stability, security and suitability for integration into business processes. McKinsey also shows that the level of benefit from AI increases along with investment in responsible AI systems. In the group of organizations with investments of less than 1 million US dollars, the high level of use of AI with significant benefit was 22%, while in the group with investments of 50 million US dollars and more this figure reached 72% (see Figure 1) [1].

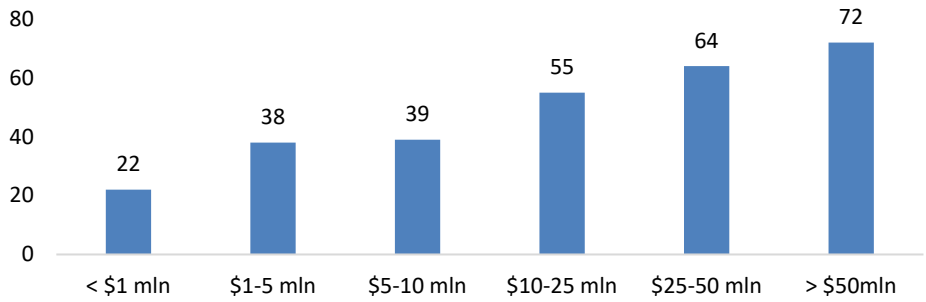

Fig. 1. Benefits of AI depending on the level of investment in responsible AI systems, %
Note: compiled by the author based on the source [1].

Testing of AI systems differs from traditional software testing in that the result of the model is not always deterministic. In a classic software product, the test is often built on the logic of checking the compliance of the actual result with the expected one. In AI systems, such logic works to a limited extent, since the model can generate different, but partially correct

ISSN: 2787-9364

answers to similar queries. That is why the OWASP Foundation considers AI testing as the basis for trust in AI systems, and not only as a security check or search for technical vulnerabilities [11].

In practical terms, AI testing should cover not only functionality, but also a wider set of characteristics. This includes resistance to manipulative requests, no leakage of sensitive information, reduction of hallucinations, control of excessive autonomy, transparency of answers, and compliance with user intent. Today, AI systems can make mistakes due to adversarial manipulation, bias and fairness failures, hallucinations, data or model poisoning, excessive agency, model drift and other factors that are not covered by conventional security tests [11].

One of the central areas of such testing is the detection of hallucinations. A. Pesaranghader and E. Li consider the detection stage as an operational input to the management of hallucinations. Its goal is to confirm the fact of a hallucination and at the same time preserve information about possible root causes [12]. This is important because the mere fixation of an incorrect answer is not enough. It is necessary to understand why it arose: due to outdated knowledge, weak retrieval, poor confidence calibration, ambiguity of the query or internal instability of the model.

Several approaches to the detection of hallucinations are distinguished in the scientific literature:

- uncertainty assessment. The model or external module analyzes how confidently the system generates the answer. Pesaranghader and Li describe token-level entropy and semantic entropy as two different ways to assess uncertainty. The first works at the token probability level, the second assesses the discrepancy between several answers at the meaning level [12];
- checking contextual faithfulness. In RAG systems, the answer is compared with the retrieved context. Es et al. define faithfulness as the ability of the answer to be deduced from the given context. If the statements in the answer are not confirmed by the context, this is a signal of a possible hallucination [5];
- checking the relevance of the answer. The answer may be grammatically correct and even factually correct, but not answer the question posed. That is why Es et al. separately highlight answer relevance as a criterion that shows how well the answer is consistent with the initial query [5];
- checking internal consistency. If the system gives contradictory answers to semantically identical queries, this indicates instability. Pesaranghader and Li describe intrinsic consistency as checking whether multiple generations of a model are consistent with each other [12];
- contextual fact-checking. The answer is compared with an external source of truth, such as a database, document, ontology, or current retrieval result. This approach is especially important in law, finance, medicine, and other fields with a high cost of error [12].

Testing of RAG systems is of particular importance, as they are often used to reduce hallucinations by connecting external sources [5]. Such an algorithm allows you to move from a general assessment of “good or bad answers” to a more formalized verification. This is important for corporate systems, where it is necessary to explain why a particular answer is considered unreliable.

Another way to reduce errors is to properly design prompts. W. Zhang and J. Zhang note that prompt engineering is an important tool for reducing hallucinations in retrieval-augmented LLMs, as it allows you to increase the semantic relevance of the query response without changing the model architecture or retraining it [18]. The simplest level is clear and precise prompts, which reduce ambiguity and help the model understand the task more

ISSN: 2787-9364

accurately. For more complex tasks, few-shot prompts, Chain-of-Thought prompts, and Program-of-Thought prompts can be used [18].

At the same time, prompting cannot be considered a full-fledged replacement for testing. It reduces the risk of error, but does not eliminate its cause. A.Kalai et al. show that hallucinations are of statistical nature and can occur even at the pretraining stage, even with high-quality training data [7]. The authors also explain why hallucinations persist after post-training: models can generate overly confident answers instead of reporting uncertainty or answering "I don't know" [7]. This leads to an important conclusion for QA: it is necessary to check not only the correctness of the answer, but also the behavior of the model in situations of lack of knowledge. The Metamorphic Testing method deserves special attention, which should be considered as a testing method used when it is impossible or too expensive to determine the only correct result of the system in advance. Segura et al. explain that a conventional test oracle determines the presence of an error by comparing the actual result with the expected one, but in complex systems such a check is not always practical [13]. Metamorphic Testing offers a different logic: it is not a single result that is tested, but the relationship between the results for the original and modified input. Kanstren describes it as transforming the seed input into a follow-up input and then checking whether the expected metamorphic relation is preserved [8]. For AI systems, this is especially important because their responses are probabilistic, depend on the formulation of the query, and often do not have a single reference version. In this case, testing moves from the question of "did the response match the expected one" to the question of "did the model behave consistently after a controlled change in the input data."

Examples of the application of Metamorphic Testing in AI systems can be given through several typical scenarios:

- search and recommendation systems. Kanstren gives an example with a search query, where the general word "car" is supplemented with the qualifying component "autonomous". In this case, what is expected is not a specific number of results, but a relation: a narrower query should return a smaller or more focused set of results [8];
- autonomous transportation systems. In examples for autonomous cars, the initial image of the road can be transformed by changing the viewing angle, adding snow, fog, rain, blurring or scaling. The metamorphic relation is that the system's decisions regarding the trajectory of movement, steering or acceleration should not change significantly if changes to the image do not change the real road situation [8];
- machine learning classifiers. Xie et al. apply Metamorphic Testing to test supervised classification algorithms. The authors note that even if the correct output cannot be known in advance, the expected change in the result after a change in the input may indicate an error or deviation in the behavior of the classifier [16];
- LLM in NLP tasks. Cho et al. show that for large language models, a metamorphic relation can be based on paraphrase. If two inputs have the same meaning but different wording, the result should remain the same for the corresponding NLP task. In the natural language inference example, the GPT-4 model gave different results for paraphrased inputs, which was recorded as a violation of the metamorphic relation [4];
- RAG systems and hallucination detection. Sok et al. in MetaRAG propose to decompose the response into atomic factoids, generate synonym mutations and antonym mutations, and then check them against the retrieved context. If synonymous variants are not confirmed by the context or antonym variants are not denied, this increases the hallucination score [15];
- compositionality analysis in NLP models. Manino et al. use metamorphic relations to test whether hidden representations of an NLP model exhibit compositional behavior. In this case, not only the final responses but also the hidden representations for pairs of related inputs are compared [9].

ISSN: 2787-9364

In practice, the main problem is as follows: if only the form of the query changes, but its content remains the same, the response of the AI system should not change significantly. Stylistic differences, a different order of sentences, a different level of detail or minor variability of formulations are permissible. Changing the main fact, conclusion, class, recommendation or logic of the response is unacceptable. This is what semantic constancy of the model means. The main expected reactions of the AI to a change in the query are given in Table 1.

Table 1 – Expected behavior of the AI system when changing the form of the query without changing its content

Type of query change	Expected response	Interpretation of the result
Query paraphrasing [17]	The main conclusion should remain the same	A change in the conclusion indicates a violation of semantic consistency
Changing the rating scale without changing relative preferences [9]	The recommendations should be similar in composition and ranking	A radical change in the list indicates instability in the recommender system
Adding spaces or minor noise to the prompt [9]	The model should maintain its understanding of the query	A loss of relevance indicates low robustness to formatting
Adding random words without changing the key meaning [9]	The response should not change its core logic	A significant deviation shows that the model depends on the surface form of the query
A semantically equivalent query in an NLU/NLG task [9]	The class, fact, or semantic result should remain stable	Instability requires additional testing or model correction

Note: compiled by the author based on sources [17; 9].

This problem has also been widely studied in academic circles. Yang et al. show that semantic consistency is a separate problem of LLM, since the model can respond differently to semantically close or equivalent queries [17]. In this context, Metamorphic Testing provides a practical test criterion: if the input relation preserves the content of the query, then the output relation should preserve the content of the response. In other words, the paraphrase should not change the model's decision. In recommender systems, this issue has practical significance. Khirbat et al. tested ChatGPT as a recommender system and changed the prompts without changing the essence of user preferences. For example, the rating, which was presented in a numerical definition such as 40%, 60%, 80%, was converted to 4/10, 6/10 and 8/10 [18], and the system's responses became different [9]. Accordingly, it is concluded that if a semantically equivalent query leads to the opposite conclusion or a radically different recommendation, this is a sign of instability of the AI system. That is why this method is promising for chatbots, LLM, RAG systems, recommendation services and autonomous agents. However, it also has a number of weaknesses, which are presented in Table 2.

Table 2 – Advantages and disadvantages of Metamorphic Testing in testing AI systems

Aspect	Advantages	Disadvantages
Test oracle problem [13]	Enables the system to be tested when there is no single reference output	Does not eliminate the need to formulate metamorphic relations correctly
Test scaling [8]	Makes it possible to automatically generate follow-up inputs from seed inputs, including through augmentation or reformulation	Test quality depends on the extent to which the transformations genuinely preserve or change the required property

ISSN: 2787-9364

Testing ML classifiers [16]	Detects implementation errors or unexpected deviations in classifier behavior	Not every deviation from the expected relation is direct evidence of a defect; sometimes it is a signal for additional validation
Testing LLMs and NLP tasks [4]	Makes it possible to verify response consistency under paraphrasing, reordering, or controlled text modification	NLP is prone to false positives, since queries that are similar in form are not always fully equivalent in meaning
Detecting hallucinations in RAG [15]	Makes it possible to localize unreliable statements at the factoid level and calculate a hallucination score	Requires high-quality retrieved context, since a retrieval error may be mistakenly interpreted as a generator hallucination
Verifying complex model properties [9]	Can be applied not only to final responses but also to hidden representations and compositionality	This approach requires access to the model's internal states, which is difficult for closed API-based models

Note: compiled by the author based on sources [8; 13; 16; 4; 15; 9].

Due to the identified shortcomings and problems, today there is a demand for new testing models that combine semantic consistency checking with intellectual analysis of the behavior of the AI system. One of such models is the author's concept of AI-Agentic Testing. Its peculiarity is that it combines two directions: the metamorphic testing methodology, which allows you to check the consistency of results in the absence of a classic test oracle, and the agent approach, within which the AI agent independently perceives the environment, analyzes the situation, performs actions and evaluates the consequences. In such logic, testing ceases to be a set of static scripts and turns into an adaptive quality control system capable of detecting hallucinations, unstable responses, logic errors, security violations and weaknesses in corporate software.

Conclusions

The spread of AI in the corporate sector transforms testing from an auxiliary technical procedure into a key condition for the trust, security and manageability of digital systems. The mass implementation of generative AI and enterprise AI is accompanied by an increase in risks: hallucinations, bias, adversarial manipulation, model drift, data/model poisoning, excessive agency and leaks of sensitive information. Therefore, responsible AI systems should be evaluated not only for performance, but also for stability, transparency, compliance with user intent, the ability to recognize uncertainty and the safety of behavior.

Metamorphic Testing is justified as a method that allows testing AI systems in the absence of a classic test oracle: not a separate answer is checked, but the relationship between the answers to the initial and controlled modified request. The main discovered property is semantic consistency: if only the form of the query changes – paraphrasing, formatting, rating scale, word order, or adding minor noise – but the content remains unchanged, then the main fact, conclusion, class, recommendation, or logic of the answer should also remain stable.

Based on the identified limitations of traditional testing and Metamorphic Testing, the author's concept of AI-Agentic Testing is proposed as an adaptive model for ensuring the quality of corporate software. Its essence lies in the combination of metamorphic testing with an agent approach: the AI-agent not only performs predefined tests, but also independently generates query variations, checks metamorphic relations, analyzes responses, detects hallucinations, logical contradictions, security violations, and instability of system behavior.

ISSN: 2787-9364

References

1. Asaftei, G.M., Roberts, R., Sticha, A. (2026, March 25). State of AI trust in 2026: Shifting to the agentic era. *McKinsey & Company*. URL: https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era
2. Autio, C., Barrett, M., Newman, J., Nunn, J., Oprea, A., Pinelis, L., Rice, A., Tabassi, E., & Vassilev, A. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). *National Institute of Standards and Technology.* https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence
3. Bengio Y. International AI safety report 2026. *International AI Safety Report*. (2026). URL: https://internationalaisafetyreport.org/sites/default/files/2026-02/international-ai-safety-report-2026_1.pdf
4. Cho, S., Terragni, V., Kang, S., Ahmed, T., & Yoo, S. (2025). Metamorphic testing of large language models for natural language processing. *arXiv*. https://arxiv.org/pdf/2511.02108
5. Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations,* 150–158. https://aclanthology.org/2024.eacl-demo.16.pdf
6. Gil, Y., Perrault R. (2026). The 2026 AI Index report. *Stanford Institute for Human-Centered Artificial Intelligence.* URL: https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf
7. Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. *arXiv.* https://arxiv.org/pdf/2509.04664
8. Kanstren T. (2020, June 23). Metamorphic testing of machine-learning based systems. *Medium.* https://medium.com/data-science/metamorphic-testing-of-machine-learning-based-systems-e1fe13baf048
9. Khirbat M., Ren Y., Castells P., and Sanderson M. (2024). Metamorphic evaluation of ChatGPT as a recommender system. *Conference acronym 'XX, June 03–05, 2024*. URL: https://arxiv.org/pdf/2411.12121
10. Manino, E., Rozanova, J., & Morante, R. (2022). Systematicity, Compositionality and Transitivity of Deep NLP Models: a Metamorphic Testing Perspective. *Findings of the Association for Computational Linguistics:* ACL 2022, pages 2355 – 2366. URL: https://aclanthology.org/2022.findings-acl.185.pdf
11. OWASP Foundation. (2025). OWASP AI Testing Guide. https://owasp.org/www-project-ai-testing-guide/
12. Pesaranghader, A., & Li, E. (2026). Hallucination detection and mitigation in large language models. *arXiv*. https://arxiv.org/pdf/2601.09929
13. Segura, S., Fraser, G., Sanchez, A. B., & Ruiz-Cortés, A. (2016). A survey on metamorphic testing. *IEEE Transactions on Software Engineering, 42(9),* 805–824. https://doi.org/10.1109/TSE.2016.2532875
14. Singh R. (2026, April 28). Citigroup lifts AI market view to over $4 trillion on enterprise adoption. *Reuters.* URL: https://www.reuters.com/business/finance/citigroup-lifts-ai-market-view-over-4-trillion-enterprise-adoption-2026-04-28/
15. Sok, C., Luz, D., & Haddam, Y. (2025). MetaRAG: Metamorphic testing for hallucination detection in RAG systems. *CEUR Workshop Proceedings*. https://ceur-ws.org/Vol-4136/iaai6.pdf

ISSN: 2787-9364

16. Xie, X., Ho, J. W. K., Murphy, C., Kaiser, G., Xu, B., & Chen, T. Y. (2011). Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software, 84(*4), 544–558. https://doi.org/10.1016/j.jss.2010.11.920

17. Yang, J., Chen, D., Sun, Y., Li, R., Feng, Z., & Peng, W. (2024). Enhancing semantic consistency of large language models through model editing: An interpretability-oriented approach. *Findings of the Association for Computational Linguistics:* ACL 2024, 3343–3353. https://aclanthology.org/2024.findings-acl.199.pdf

18. Zhang, W., & Zhang, J. (2025). Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. *Mathematics, 13(5),* 856. https://doi.org/10.3390/math13050856

19. Lavrik, N. (2026). Transparent financial architecture as a tool for feducing graud risks. *Global Prosperity*, *6*(1). https://gprosperity.org/index.php/journal/article/view/256

20. Lavrik, N. (2026). Transformation of accounting into a strategic asset based on the practice of clear finance architecture. SWorldJournal, 4(36-04), 16–30. https://doi.org/10.30888/2663-5712.2026-36-04-013

ISSN: 2787-9364