Interactive multi-step agent for reliable artificially generated texts detection in industrial security systems*

Iryna Dumyn[1,†]

[1] Lviv Polytechnic National University, Bandery 12, 79000, Lviv, Ukraine

Abstract

This paper examines the problem of low reliability of existing automatic and human methods for detecting artificially generated texts, whose accuracy ranges from 43% to 81% and is unacceptably low in the context of critical industrial safety systems. The emergence of high-quality large language models poses a direct threat to autonomous industrial systems due to the possibility of automated creation of plausible instructions that contain incorrect and unreliable information, which can lead to system failures. The goal of this study is to develop and justify a modular architecture for an interactive dialogue agent designed to improve the reliability of detecting AI-generated texts in industrial safety systems through the use of multi-step analysis. The proposed approach is based on a combination of stylometric text analysis, automated fact-checking, and cross-examination. The agent's architecture involves sequential message processing, calculation of an integral risk score, and notification of the operator. The combination of these methods is expected to reduce the number of false positives and negatives, ensuring higher accuracy and robustness in detecting AI-generated texts compared to existing single-step AI detectors.

Keywords

interactive dialogue agent, ai-generated text detection, llm-generated content, stylometric analysis, semantic fact-checking, cross-examination dialogue strategy, industrial information security, agent-based systems

1. Introduction

Modern large language models can generate texts whose quality makes it difficult to reliably distinguish them from content created by humans. Current methods for detecting artificially generated texts are divided into five main groups: based on watermarks, analysis of linguistic and statistical features, neural approaches, hybrid models and methods involving human expert analysis. However, multilingualism, the hybrid nature of the generated content and the continuous improvement of generative algorithms limit the effectiveness of existing automatic identifiers of artificially generated text.

Empirical studies confirm the low reliability of existing tools. Systematic reviews show that the accuracy of automatic identifiers varies in the range of 43–81% with unacceptably high false positive rates, in particular for texts written by non-native speakers. Even human analysis can be ineffective, with AI text recognition accuracy showing only 19%, which is close to random guessing. This suggests that exclusive reliance on single-shot detectors is insufficient and requires the use of additional verification methods that assess style and content.

At the same time, the emergence of high-quality generative algorithms has significantly changed the list of security threats. The ability to automatically create plausible fake instructions, documents, or manipulative messages poses a direct threat to critical industrial systems[1]. In the context of autonomous dialog agents capable of multi-path thinking, the lack of reliable verification can lead to system failures or dangerous actions. The risks include data poisoning, embedding

* SmartIndustry 2026: 3rd International Conference on Smart Automation & Robotics for Future Industry, March 26-27, 2026, Lviv, Ukraine

1* Corresponding author.

† These authors contributed equally.

iryna.b.shvorob@lpnu.ua (I. Dumyn)

0000-0001-5569-2647 (I. Dumyn)

inversion, and hint manipulation attacks. To minimize these threats, it is necessary to develop more robust and flexible verification mechanisms.

Given the limited reliability of existing solutions and the growth of threats in the industrial sector, there is a need to create an interactive mechanism that combines automated analysis with deep contextual verification.

The purpose of this research is to develop and conceptually substantiate the architecture of an interactive dialog agent designed to increase the reliability of AI-generated text detection in industrial security systems by applying multi-step analysis.

Proposed approach is based on a combination of stylometric analysis, fact-checking, and cross-examination and should improve the accuracy and robustness of AI-generated text detection compared to existing one-shot AI detectors. This will ensure the prompt identification of potentially malicious or unreliable messages in critical industrial systems, contributing to improving their information security. The combination of structural analysis, contextual verification, and adaptive algorithms reduces the risks of erroneous decisions in critical environments [2, 3].

2. Review of research and systematization of methods for detecting AI texts

2.1. Research overview

Modern research in the field of interactive dialogue systems demonstrates the transition from static fine-tuning to adaptive reinforcement learning strategies. In particular, in [4], a hindsight-regeneration approach is proposed, in which dialogues are reformulated by a large language model after their completion to achieve a better target result, and the resulting optimized versions are used for offline reinforcement learning of the agent. Experimental results demonstrate an increase in the efficiency, naturalness, and controllability of the dialogue compared to traditional pre-training methods. This approach emphasizes the importance of dynamically considering the interlocutor's reactions and can be extrapolated to AI text detection systems, where clarifying questions play the role of a content verification mechanism.

The problem of detecting AI-generated content is actively developing in several directions: watermarking, stylometric methods, neural discriminators and hybrid approaches. The review [5] systematizes watermarking methods for text, images and audio, emphasizing their proactive nature - source identification occurs at the stage of content generation. At the same time, the vulnerability of such methods to paraphrasing and the lack of unified standards are emphasized. Similarly, in [8] DeepTextMark is proposed a deep approach to text watermarking, which combines semantic coding and a transformative classifier, demonstrating resistance to obfuscation attacks. The general reviews [6] and [7] state that none of the approaches provide guaranteed accuracy, and the effectiveness of detectors significantly depends on the language, genre and evolution of generative models.

Empirical studies confirm the limitations of existing tools. In particular, in [9] it is reported that the accuracy of open AI detectors ranges from 43–81%, while human experts demonstrate results close to random guessing (≈19%). An additional problem is the phenomenon of LLM hallucinations - the generation of plausible but false statements. In [10], an unsupervised MIND system is proposed for detecting hallucinations based on the internal states of the model in real time, which justifies the need to integrate fact-checking modules into detection architectures. The study [11] demonstrates that AI texts are characterized by lower perplexity and more homogeneous structure, which confirms the feasibility of stylometric indicators. In addition, semantically enhanced frameworks, such as SEFD (He et al.), combine primary detectors with semantic similarity analysis to increase paraphrasing resistance, although they require significant resources to maintain the knowledge base [12].

In summary, current research suggests the need for combined approaches that integrate stylometric analysis, watermarking, semantic matching, and interactive cross-examination. Despite numerous proposals, the accuracy of existing systems remains unstable, and human expertise does not guarantee reliability. Promising directions include the development of interactive agents with reinforcement learning mechanisms, real-time hallucination monitoring, and semantic control that consider the context and dynamics of the dialogue.

2.2. Classification of detection approaches

An analysis of modern approaches to determining the origin of a text shows that detection methods are conditionally classified into five main groups.

Group 1. Watermarking methods are based on embedding hidden patterns in the generated text, which provides a quick verification of its source. The effectiveness of such methods is determined by a compromise between resistance to editing, minimal impact on the text and the amount of information transmitted. The main disadvantages are the need for control over the generation process and vulnerability to simple editing or formatting operations, which can lead to loss or desynchronization of the watermark.

Group 2. Feature (stylometric) approaches are based on the analysis of statistical and linguistic characteristics of the text, in particular, the average sentence length, the frequency of punctuation, lexical diversity, readability indicators and emotional polarity. Their advantage is the interpretability of the results and the possibility of applying classical machine learning methods, such as Random Forest, which in experimental studies demonstrated an accuracy of about 82.6%. At the same time, these methods are sensitive to the language and genre of the text and are ineffective when deliberately imitating human style.

Group 3. Neural approaches use large language models or discriminative transformer architectures, such as RoBERTa, to estimate token probabilities and analyze text structural properties. Systems such as GLTR, which visualizes word predictability, or GROVER, which combines generation and detection with an accuracy of about 92%, have demonstrated the ability to detect complex statistical and semantic patterns. However, such methods require significant computational resources and large training samples and remain vulnerable to deliberate obfuscation attacks.

Group 4. Hybrid approaches combine several methods, including neural analysis with stylometric features or watermarks, to increase resistance to attacks and reduce the number of false positives. The disadvantage of such solutions is the complexity of implementation and the need for careful tuning and calibration of components.

Group 5. Human-based approaches rely on expert judgment, comparative analysis of style with the author's previous work, and contextual analysis. While they allow for a broader context, these methods remain subjective, small-scale, and resource-intensive; studies have shown that human recognition of AI-generated texts is only about 19% accurate.

2.3. Reliability analysis of open detectors

Empirical research shows that none of the existing detection tools provide error-free classification of the origin of the text. A systematic review of 17 scientific studies showed that the accuracy of current AI detectors varies from 43% to 81%, with significant levels of both false positives and false negatives. Even under the most favorable conditions, such systems demonstrate misclassification in about one case out of five.

Particularly problematic is the high rate of false positives, where human-generated texts are mistakenly identified as AI-generated; in some cases, this rate reaches 70%, particularly for authors who are not native speakers of the language. Table 1 provides a comparative summary of the characteristics of popular open-source detectors, the evaluation of which is based on limited experimental samples.

Table 1
Open AI detectors

Detector	Features	Strengths	Weaknesses	Accuracy rating
ZeroGPT	Free web detector; estimates the probability of AI text.	Stable, predictable, efficient for blogs and mixed texts.	Prone to false positives for academic and short passages; sensitive to perfect grammar.	≈94% sensitivity / 93% specificity in a small sample.
GPTZero	Analyzes perplexity, burstiness, and text structure.	Reliable for long and structured texts; high success rate (99–100%) on experimental sample.	Overestimates AI probability for formal essays; weaker on edited content; paid for full version; length limit.	≈100% sensitivity / 99.6% specificity in a small sample.
PhraslyAI Detector	Commercial tool.	High consistency on English texts; simple interface.	Limited open data on accuracy; weak support for other languages.	n/a (0.57–0.95 ICC in the experiment).
Grammarly AI Detector	Part of the grammar service.	Integration with the editor; ease of use.	Low consistency; often incorrectly labels edited texts.	n/a (0.57–0.95 ICC).
NoteGPT	Free, focuses on the balance between sensitivity and specificity.	Fewest false positives; recognizes mixed texts and modern patterns well.	Lack of formal scientific research; figures may vary.	Blog review: high accuracy in practical tests.
Winston AI	Service for educational institutions.	Unexplained usage		

In the context of industrial systems where misclassification can lead to malfunctions or potentially dangerous actions, such instability in accuracy is unacceptable. Accordingly, modern detection tools can only be considered as auxiliary tools, the use of which is advisable only in combination with additional verification and control procedures.

2.4. Stylistic and semantic indicators of AI texts

To increase the reliability of the assessment, which complements automated detection methods, it is advisable to consider characteristic indicators that researchers and teachers identify in texts generated by artificial intelligence, the so-called "red flags." These include factual errors and fictitious sources, when the model makes plausible but incorrect statements or refers to non-existent publications, which is a manifestation of hallucinations.

Another indicator is an impersonal and predictable style of presentation: such texts are usually grammatically correct, but devoid of an individual authorial "voice", are characterized by a template structure, uniform paragraph length and limited depth of analysis. In addition, there may be a discrepancy with the requirements of the task, in particular a tendency to overly generalized answers or incorrect execution of instructions that require precise data or specific sources. A separate indicator is excessive linguistic correctness, when a sharp change in the author's style — particularly the appearance of impeccable grammar, uncharacteristic of his previous works — may indicate automated generation.

Practical recommendations boil down to a comprehensive approach that includes a comparative analysis of style with previously completed works, verification of the facts and sources cited, and the use of cross-questions that allow you to reveal a superficial or formal understanding of the subject area.

3. Threat analysis in agent systems and hallucination monitoring

3.1. Risks of using AI texts in industrial systems

Hallucination in the context of large language models is defined as the generation of plausible, grammatically correct, and coherent text that contains factually false, unverified, or logically contradictory information. Such responses may include fictional events, incorrect references, distorted data, or confabulations presented as reliable facts. Unlike deliberate misinformation, hallucinations are a consequence of the statistical nature of the model and the limitations of its training, rather than deliberate behavior. Conceptually, this phenomenon is associated with human confabulation, the construction of a convincing but untrue narrative based on incomplete or incorrect knowledge.

Hallucinations can be classified into several categories. Factual inaccuracies include errors in dates, names, scientific statements, or biographical information. Nonsense manifests itself in the form of responses that lose connection with the context of the query or contain semantically incorrect statements. Contradictions can occur both within a single text and between different model responses when key facts change or internal logical consistency is violated. One of the key reasons is the limitations of training data: incompleteness, noise, bias, or insufficient representation of certain topics lead to inaccurate generalizations. Additional factors include architectural and algorithmic limitations of models, such as overfitting, insufficient modeling of cause-and-effect relationships, and the lack of a mechanism for internal verification of the truthfulness of statements. A limited context window also contributes to information loss in long dialogues, which can lead to inconsistent responses. Linguistic subtleties such as irony, sarcasm, cultural allusions, or ambiguity pose an additional challenge for models, as their interpretation requires extra-model knowledge and deep pragmatic analysis. As a result, the model may generate responses that formally comply with syntax but do not reflect the actual meaning or relevant

context. This highlights the need to integrate fact-checking mechanisms, external knowledge bases, and authenticity monitoring into systems that use LLM in critical applications.

To formalize the process of hallucination generation and detection, it is useful to view it as a sequence of interrelated stages of information processing. In the first stage, the input query is passed to a large language model, which generates a response based on statistically trained parameters and internal knowledge representations. Due to limitations in training data, architecture, or context window, factual errors, logical contradictions, or semantic distortions may occur at this stage. In the next stage, the generated text is sent to a verification subsystem, which performs a multi-level analysis: stylometric characteristics are evaluated, including sentence length, lexical diversity, and structural homogeneity; automated fact-checking is performed using external knowledge bases; and internal contradictions or nonsense are identified. If factual inaccuracies or logical inconsistencies are detected, the response is marked as potentially hallucinatory, after which a correction procedure is initiated, for example, by generating clarifying questions, re-asking with additional context, or requesting references to sources. This algorithmic approach not only describes the mechanism of hallucination generation, but also forms the basis for their systematic detection and minimization, combining methods of stylometric analysis, semantic verification, and external fact validation.

In the context of the spread of the Internet of Things and autonomous agent systems, the lack of proper verification of information flows poses a direct threat to critical industrial infrastructures. Automated generations of falsified instructions or manipulative entries in event logs can lead to disruption of controllability of technological processes, system failures and potentially dangerous actions. Research on agent-based security identifies several key attack vectors. These include cue injection, which involves embedding hidden instructions into input data to force a model to behave in an undesirable way; data and model poisoning, which involves introducing corrupted data into training or refresh samples; model inversion, which allows private information to be recovered from training data by analyzing the model's responses; and risks associated with weak identity and supply chain controls, such as unauthorized access through improper token management or reliance on third-party APIs. To mitigate these threats, it is recommended to implement multi-layered protection mechanisms. These include continuous anomaly monitoring and behavioral analysis of AI systems for early detection of deviations, verification and sanitization of input data to minimize prompt injection risks, the use of a zero-trust architecture with strict access control and token protection, and systematic adversarial testing to identify and eliminate vulnerabilities.

3.2. Monitoring hallucinations

A particular threat to the reliability of autonomous agent systems is the hallucinations of large language models, i.e. the generation of convincing, but factually incorrect or fabricated statements, which can lead to disinformation and undermining trust in the system. To counteract this phenomenon, automated monitoring mechanisms are proposed that analyze each agent's message in real time, compare it with authoritative knowledge bases, and record semantic or factual deviations. At the same time, effective control involves not only detecting hallucinations, but also analyzing their causes, which necessitates the need for transparency in the logic of the agents' functioning and the interpretability of their decisions.

Considering the limited reliability of isolated detectors and the growing range of threats, a modular conceptual architecture of an interactive dialog agent for detecting texts generated by artificial intelligence in industrial systems is proposed. The architecture is based on sequential multi-stage message processing and involves the integration of human control mechanisms, which allows to increase the reliability of decision-making and reduce the risks of critical errors in real operating conditions.

The system architecture consists of several interconnected functional modules, each of which is responsible for a separate stage of message analysis (Fig. 1). The preprocessing block ensures the reception of the incoming message, performs its normalization, tokenization and automatic

language detection. The style and structure evaluation module analyzes the statistical and linguistic characteristics of the text, in particular sentence length, lexical diversity, readability indices and emotional polarity, and compares the indicators of perplexity and burstiness with reference style profiles.

The fact-checking block implements semantic matching of statements with reliable sources, such as knowledge bases, regulatory documents or RAG systems. The cross-examination module, based on the results of the previous stages, automatically generates several clarifying questions regarding content, terminology or procedures and analyzes the received answers; the presence of contradictions or the inability to provide reasonable explanations is considered as an indicator of automated generation. The decision-making engine integrates the results of all modules into a single integrated risk indicator in the form of a coefficient or probability, based on which a conclusion is formed about the origin of the text ("human" or "AI") and further actions are initiated notifying the operator or isolating the message. User interaction and logging are implemented through a specialized interface that provides risk level display, feedback and dialogue history storage, and supports security mechanisms, including authentication, communication channel encryption and protection against prompt injection attacks. The system's operating algorithm is consistent and includes message reception, its pre-processing, style and structural analysis, fact checking, formation and analysis of clarifying questions, calculation of an integral risk score, decision making and recording of results with appropriate informing of the operator.

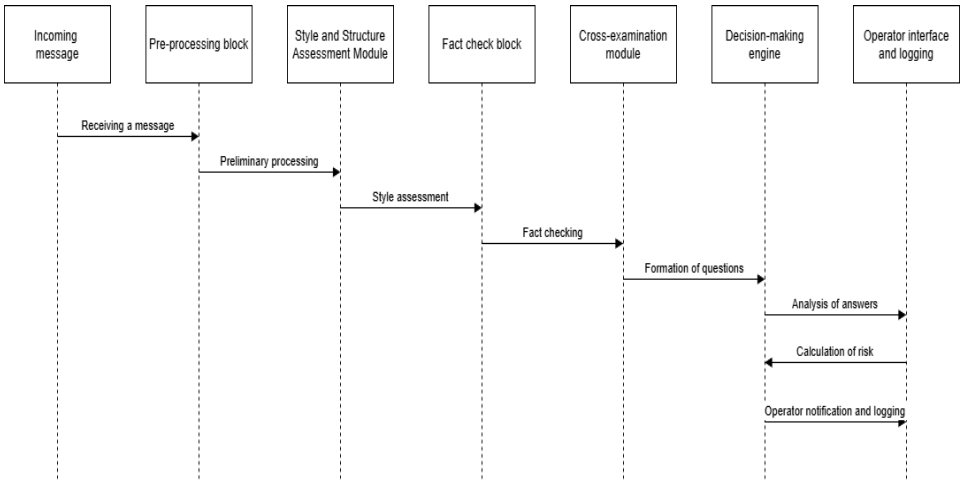

Figure 1: Conceptual diagram of the proposed approach.

3.3. Expert test methodology

For the experimental evaluation of the interactive dialog agent, a pilot sample was generated from several text fragments of 5–10 sentences each, dedicated to production procedures and equipment validation. Four fragments were prepared manually by engineers on pressure sensor calibration, welding robot inspection, conveyor cleaning after batch completion, and testing of the control and measurement system, while another four were generated by a large language model in the format of generalized instructions. The generated materials are not included in the training samples of the studied models, which minimizes the risk of their prior memorization and ensures the correctness of the experiment.

During the creation of the generated fragments, characteristic stylistic features of AI-generated texts were purposefully reproduced predictable compositional structure, repetitive wording, lack of specific technical details, and an increased proportion of passive constructions. These features are considered as typical indicators of automated generation and are used to assess the effectiveness of the proposed approach.

A fragment of the formed sample is presented in a tabular format (Table 2). For each text fragment, a unique identifier is given, the source of origin is indicated, and a brief description of

the content is provided. This structure ensures transparency of the experimental design and convenience of further analysis of the results.

Table 2
Fragment of the formed sample

ID	Source	Fragment description
1	Human	Calibration of the pressure sensor in the pipeline; checking the shut-off of the supply, connecting the reference pressure gauge, recording deviations and tightness.
2	Human	Welding robot inspection: lubrication of moving parts, test run and positioning accuracy control.
3	Human	Cleaning the conveyor after a batch: turning off the line, using brushes and cleaning solution, visual inspection for damage
4	Human	Testing the control and measurement system: checking sensor signals, comparing with reference values, and documenting deviations
5	AI	General instructions for checking the performance of equipment without specific devices: check stability, monitor parameters, adjust processes
6	AI	Guidelines for process compliance with standards: monitoring parameters, adherence to general control principles, documenting stages.
7	AI	General recommendations for system testing: performing tasks according to criteria, eliminating deviations using standard procedures, updating documentation.
8	AI	Temperature parameter control: checking device settings, stability of modes, documentation and correction of discrepancies

The interactive agent is implemented as two sequentially connected modules. The first module performs stylometric analysis, within which basic quantitative characteristics are calculated for each text fragment, in particular, the average sentence length, the lexical diversity index (the ratio of the number of unique words to the total number of word forms), and the readability index, implemented in the prototype in a simplified form based on the length of words and sentences. It is known that texts generated by AI are characterized by reduced perplexity, a more uniform sentence structure, and repetition of formulations, while texts created by humans demonstrate greater structural variability. In the case when at least two of the three indicators indicate a generalized style, the module records an increased risk of automated generation. The second module implements a cross-examination mechanism, forming two clarifying questions regarding the content of the text. The meaningfulness of the answers is assessed by length and specificity; excessively short or generalized answers increase the likelihood of classifying the text as generated by AI. The final decision is made based on the principle of logical combination of results: if there is a risk signal from at least one module, the text is classified as "AI", otherwise - as "Human".

The experimental study was conducted in three stages:

Stage 1. Launch of the prototype with calculation of stylometric indicators and cross-examination.

Stage 2. Recording the results of each module, the final classification, and the actual category of the text.

Stage 3. Comparison with the performance of open AI detectors.

An example of the work of the developed agent is shown in Figure 2.

```
======= Аналіз фрагменту 1 =======
Текст: Інженер виконує калібрування датчика тиску трубопроводу. Спочатку закриває вхідні та вихідні клапани...
Статистика: avg_len=9.5, diversity=0.97, long_words=2.63% -> AI Marker: False

Фрагмент №1: відповідайте на запитання. (Для AI можна писати 'не знаю' або коротку відповідь).
Яка основна мета цієї процедури?
> калібрування датчика трубопроводу
Назвіть один конкретний інструмент або прилад, згаданий у тексті.
> датчик

ФІНАЛЬНЕ РІШЕННЯ: AI (stylometric=False, cross_exam=True)

======= Аналіз фрагменту 2 =======
Текст: Перед запуском зварювального робота працівник оглядає шви, видаляє окалину та тестує механізм на хол...
Статистика: avg_len=13.3, diversity=0.97, long_words=5.00% -> AI Marker: True

Фрагмент №2: відповідайте на запитання. (Для AI можна писати 'не знаю' або коротку відповідь).
Яка основна мета цієї процедури?
> зварювання
Назвіть один конкретний інструмент або прилад, згаданий у тексті.
> зварювальний робот

ФІНАЛЬНЕ РІШЕННЯ: AI (stylometric=True, cross_exam=True)
```

Figure 2: Example of agent operation.

In a small pilot sample, the agent correctly identified all texts, achieving 100% accuracy, which exceeds the average performance of open tools (43–81%), which are characterized by significant levels of false positives and false negatives. The automatic processing time was insignificant (about one second for stylometric analysis), and the total number of clarification questions was 16. The results obtained indicate the potential effectiveness of the interactive approach, while requiring confirmation on larger and more representative samples.

The results obtained require careful interpretation. First, the sample size is insufficient to form statistically sound generalizations, which creates a risk of overfitting the methodology to specific texts. Second, the simulated responses for AI-generated fragments do not fully reflect the diversity of real-world interaction scenarios with language models, which may affect the assessment of the effectiveness of cross-examination.

At the same time, the results of the pilot study indicate that the integration of stylometric analysis with the mechanism of clarifying questions can improve the quality of detection of

automatically generated texts. This confirms the feasibility of further research in the direction of developing interactive approaches using wider and more representative experimental samples.

Table 3
Frequency of Special Characters

ID	Fact source	Stylometry result (risk of AI)	Interrogation result (AI risk)	Final agent classification
1	Human	No	No	Human
2	Human	No	No	Human
3	Human	No	No	Human
4	Human	No	No	Human
5	AI	No	Yes	AI
6	AI	No	Yes	AI
7	AI	No	Yes	AI
8	AI	No	Yes	AI

4. Conclusions and discussion

After the experimental study is completed, it is advisable to generalize the obtained results, which involves a comparative analysis of the accuracy of the interactive agent with the indicators of open detectors, a detailed analysis of false positives, as well as an assessment of the time costs for message processing. It is expected that the combination of stylometric analysis, automated fact-checking, and cross-examination mechanisms will contribute to a reduction in the number of both false positive and false negative decisions, ensuring an increased level of reliability and trust in information flows.

At the same time, the proposed approach is characterized by several limitations, in particular, the limited scope of pilot testing, the subjectivity of evaluating answers to clarifying questions, the dependence on the participation of specialists in fact-checking, as well as the potential influence of the respondent's stylistic features on the classification results.

When integrating an interactive agent into industrial systems, it is necessary to take into account the requirements for user authentication and encryption of communication channels, ensuring compatibility with existing enterprise monitoring and information security systems, supporting the scalability of the architecture for multilingual dialogues, as well as regular updating of rules and knowledge bases in order to adapt to the evolution of large language models.

Prospects for further research include expanding the experimental base by using more representative datasets, automated integration with corporate knowledge bases and regulatory documents, combining an interactive approach with next-generation neural detectors, and applying reinforcement learning methods to optimize the cross-examination formation strategy.

Declaration on Generative AI

During the preparation of this work, the author used AI tools in order to: Grammar and spelling check.

References

[1] N. Shakhovska and I. Shvorob, "The method for detecting plagiarism in a collection of documents," 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, Ukraine, 2015, pp. 142-145, doi: 10.1109/STC-CSIT.2015.7325453.

[2] I. Dumyn, O. Basystiuk, and A. Dumyn, “Graph-based approaches for multimodal medical data processing,” in Proc. 7th Int. Conf. on Informatics & Data-Driven Medicine, Birmingham, UK, Nov. 2024, pp. 349–362.

[3] I. Shvorob, “New approach for saving semistructured medical data,” in Advances in Intelligent Systems and Computing, vol. 512, N. Shakhovska, Ed. Cham, Switzerland: Springer, 2017, pp. 29–39. doi: 10.1007/978-3-319-45991-2_3.

[4] J. Hong, J. Lin, A. Dragan, and S. Levine, “Interactive dialogue agents via reinforcement learning on hindsight regenerations,” arXiv preprint arXiv:2411.05194, 2024.

[5] L. Cao, “Watermarking for AI content detection: A review on text, visual, and audio modalities,” arXiv preprint arXiv:2504.03765, 2025.

[6] S. Fariello, G. Fenza, F. Forte, M. Gallo, and M. Marotta, “Distinguishing human from machine: A review of advances and challenges in AI-generated text detection,” International Journal of Interactive Multimedia & Artificial Intelligence, vol. 9, no. 3, 2025.

[7] J. Wu, S. Yang, R. Zhan, Y. Yuan, LS Chao, and DF Wong, “A survey on LLM-generated text detection: Necessity, methods, and future directions,” Computational Linguistics, vol. 51, no. 1, pp. 275–338, 2025.

[8] T. Munyer, AA Tanvir, A. Das, and X. Zhong, “DeepTextMark: A deep learning-driven text watermarking approach for identifying large language model generated text,” IEEE Access, vol. 12, pp. 40508–40520, 2024.

[9] AM Elkhatat, K. Elsaid, and S. Almeer, “Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text,” International Journal for Educational Integrity, vol. 19, no. 1, pp. 1–16, 2023.

[10] W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu, Y. Zhou, and Y. Liu, “Unsupervised real-time hallucination detection based on the internal states of large language models,” arXiv preprint arXiv:2403.06448, 2024.

[11] HM Jaashan and WRA Bin-Hady, “Stylometric analysis of AI-generated texts: A comparative study of ChatGPT and DeepSeek,” Cogent Arts & Humanities, vol. 12, no. 1, Art. no. 2553162, 2025.

[12] W. He, B. Hou, T. Shang, DA Tarzanagh, Q. Long, and L. Shen, “SEFD: Semantic-enhanced framework for detecting LLM-generated text,” in Proc. IEEE Int. Conf. on Big Data (BigData), 2024, pp. 1309–1314.