

# Agentic AI Modeling for the Prevention of Examination Paper Copying: A Multi-Agent Reference Architecture



**Manish Gupta**

Senior AI Solution Enterprise Architect & Senior AI Governance Lead, USA

## CORRESPONDING AUTHOR

**Manish Gupta**

e-mail: [manish.ingupta@gmail.com](mailto:manish.ingupta@gmail.com)

## KEYWORDS

AI in Education, Exam Security, Agentic AI, Multi-Agent Systems, Paper Leaks, Paper Copying.

## ARTICLE DETAILS

Received 02 August 2026; revised 11 September 2026; accepted 02 October 2026

DOI: 10.26671/IJIRG.2026.4.15.101

## CITATION

Gupta, M. (2026). Agentic AI Modeling for the Prevention of Examination Paper Copying: A Multi-Agent Reference Architecture. *Int J Innovat Res Growth*, 15(4), 154058-154067. DOI



This work may be used under the terms of the Creative Commons License.

## Abstract

Exam papers get leaked or copied in countries all over the world, and every time it happens, people lose a little more trust in the system. Today's defenses, cameras, watermarks, locked storage, secure transport, each cover one small part of the journey a paper takes, but they rarely talk to each other, and a leak can slip through the cracks between them. This paper describes a different approach: a small team of AI agents that each watch one part of that journey, from the printing press to the exam hall to the internet afterward, plus a new agent that checks answer scripts for suspicious similarity once the exam is over. No single agent can raise an alarm on its own. Each one reports a confidence score to a coordinator, and only when enough agents agree does a case get passed to a human for the final decision. We explain how this works, how the evidence gets combined, how we choose the thresholds that decide when something looks suspicious, and we test the idea through simulation, reporting accuracy, precision, recall, false alarms, and missed leaks, along with a test that removes one agent at a time to see how much each one actually matters. The results suggest that combining agents this way catches more real problems and raises fewer false alarms than either one big model or a simple rule-based checklist, while a person still makes every final call.

## 1. Introduction

Every year, millions of people take important exams that affect their education, careers, and jobs. For an exam to be trusted, the question paper must stay secret, the process must be fair, and answer scripts must genuinely belong to the student. Even with a lot of money spent on security, papers still leak, people still gain access they shouldn't, students still get impersonated, answers still get copied, and secret material still ends up online [1], [11]. This shows that exam security isn't just about guarding a piece of paper, it's about protecting a whole chain of steps handled by different people and systems, where every handoff is a chance for something to go wrong. Most security tools today work alone, cameras, encryption, and access check each guard their own piece, but rarely share information with each other. So small problems at different stages often get treated as unrelated, even though together they might signal a bigger threat, and checking answer scripts for copying after the exam usually isn't connected to protecting the paper before the exam. This study looks at how a team of AI agents working together, watching different parts of the process, sharing evidence, and checking answer scripts for suspicious similarities, can close these gaps. The goal isn't to replace human judgment, but to give exam officials faster, evidence-backed support so the whole process stays fair and trustworthy.

## 2. What Goes Wrong Today

Current examination security relies on a combination of administrative procedures, physical safeguards, and isolated technological solutions. Administrative measures include restricted personnel access, signed custody records, and manual verification procedures. Physical safeguards typically involve sealed packages, secure storage facilities, surveillance cameras, and controlled transportation of examination materials. In addition, technologies such as document watermarking, encryption, barcode tracking, and access authentication are widely used to strengthen specific stages of the examination process.

Although these measures improve security, they are generally implemented as independent controls rather than as components of an integrated monitoring system. Each mechanism protects a limited part of the examination workflow and operates with little or no awareness of events occurring at other stages. Consequently, information collected during printing, transportation, storage, examination administration, and post-examination activities is rarely analyzed together to determine whether multiple small incidents collectively indicate a developing security threat.

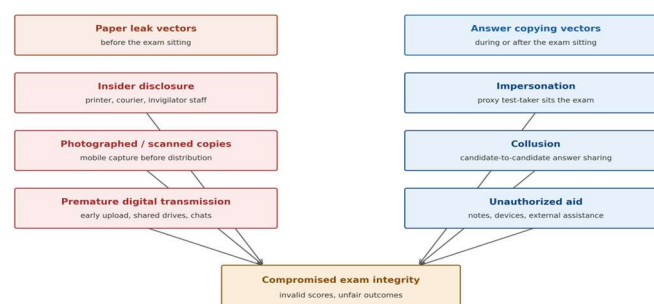
Another important limitation is the dependence on manual investigation. In many examination systems, security personnel review logs, surveillance footage, and incident reports only after suspicious activity has been reported. This reactive approach delays the identification of security breaches and reduces the opportunity to prevent examination compromise before confidential materials are exposed. By the time evidence of a leak appears on social media, messaging platforms, or unauthorized websites, the examination may already have been affected, leaving examination authorities with limited response options such as postponement, cancellation, or re-administration.

Furthermore, existing examination security practices often treat pre-examination protection and post-examination verification as separate processes. While considerable effort is invested in protecting question papers before the examination begins, comparatively less attention is given to identifying suspicious similarities among submitted answer scripts that may indicate copying or organized collusion. Without semantic and textual comparison of answer scripts alongside other security evidence, valuable indicators of examination malpractice may remain undetected.

These limitations suggest that examination security requires more than independent preventive measures. An effective solution should continuously observe the complete examination workflow, correlate events from multiple sources, identify abnormal patterns as they emerge, and support timely decision-making before isolated incidents develop into large-scale security breaches.

## 3. Problem Context and Research Motivation

Examination integrity is threatened by multiple security risks occurring at different stages of the assessment lifecycle.



**Fig.1 Paper leak and answer copying vectors compromising exam integrity.**

Fig. 1 shows threats are dividing into two major categories: paper leakage vectors and answer copying vectors. Paper leakage may occur before the examination through insider disclosure, unauthorized document capture, or premature digital distribution of question papers. Similarly, answer copying threats may occur during or after the examination through impersonation, candidate collaboration, unauthorized assistance, or highly similar submitted responses. These diverse threat paths demonstrate that examination security is not a single-point problem but a complex process involving physical, digital, and human factors.

Existing protection mechanisms typically address individual risks independently, such as document watermarking, access control, surveillance, or plagiarism detection. However, isolated solutions cannot identify relationships between multiple security events occurring across the examination lifecycle. This research addresses this limitation by proposing a collaborative multi-agent AI framework in which specialized agents monitor different threat areas, including document security, access behavior, online exposure, and semantic as well as textual similarity analysis of answer scripts. The collected evidence is combined through an AI coordination layer, while human authorities retain final decision responsibility to ensure transparent and accountable examination security management.

#### **4. Related Work and Research Gap**

Research on examination security has addressed several individual aspects of protecting the examination process. However, most existing studies focus on solving specific security problems independently rather than providing an integrated framework capable of monitoring the complete examination lifecycle. This section reviews the major research directions that form the foundation of the proposed work and highlights the remaining research gap.

##### ***A. Secure Examination Document Protection***

One widely explored approach is the protection of examination papers through digital watermarking, document fingerprinting, encryption, and secure distribution mechanisms [3]. These techniques enable confidential documents to be authenticated and, in some cases, traced back to their source if unauthorized copies are discovered. While these methods improve document confidentiality and support forensic investigation after a leak, they primarily identify the origin of a compromised document rather than preventing or detecting coordinated security threats before they spread.

##### ***B. Secure Chain of Custody***

Maintaining an auditable record of examination paper movement has attracted considerable research attention through tamper-resistant logging, distributed ledgers, and blockchain-based custody management [4]. These approaches improve accountability by recording every transfer of examination material and making unauthorized modifications difficult to conceal. Nevertheless, custody records alone cannot determine whether abnormal activities across multiple operational stages collectively indicate an ongoing security breach.

##### ***C. Intelligent Monitoring and Anomaly Detection***

Artificial intelligence has been extensively applied to anomaly detection in domains such as cybersecurity, finance, and industrial monitoring [5]. Machine learning techniques can identify unusual patterns by comparing current observations with expected system behavior. Similar approaches have been adopted in examination environments to detect abnormal movement, suspicious behavior, unauthorized device usage, and irregular operational activities. Although effective within individual environments, these systems typically analyze isolated events without considering their relationship to other stages of the examination process.

##### ***D. Identity Verification and Examination Authentication***

Biometric authentication, including facial recognition, fingerprint verification, and liveness detection, has become an important mechanism for reducing impersonation during examinations [6]. These technologies strengthen candidate authentication and improve examination integrity by verifying that registered candidates are physically present. However, authentication solutions focus primarily on identity verification and provide limited support for detecting document leakage, coordinated insider threats, or post-examination malpractice.

##### ***E. Answer Script Similarity and Copying Detection***

Detecting copied answers remains a significant challenge in examination management [2]. Traditional approaches rely on manual review or simple text-matching techniques that identify identical or highly similar responses. Recent advances in natural language processing have enabled semantic and textual similarity analysis, allowing systems to detect conceptually similar answers even when different words or sentence structures are used. Despite these developments, answer script similarity analysis is generally performed as an independent plagiarism or copying detection task and is rarely integrated with other examination security information. As a result, opportunities to correlate suspicious answer similarities with earlier security events, such as question paper exposure or unauthorized communication, are often overlooked.

### ***F. Collaborative Decision-Making in Distributed Systems***

Distributed computing research has demonstrated that multiple independent components can collectively reach reliable decisions through consensus mechanisms and fault-tolerant protocols. These techniques improve system reliability by reducing the influence of individual failures or compromised components. Although consensus-based decision models have been successfully applied in distributed computing and cybersecurity, their application to coordinated examination security remains limited.

### ***G. Explainable Artificial Intelligence for Security Applications***

As artificial intelligence becomes increasingly involved in security decision-making, Explainability has become an important research area. Explainable AI techniques provide understandable reasons for automated decisions, enabling human reviewers to verify system recommendations and maintain accountability [10]. This capability is particularly important in examination environments, where security alerts may affect students, staff, and institutional credibility, requiring transparent and auditable decision support.

### ***H. Foundations of Combined and Group Decision-Making***

The idea that several average detectors can perform better than one powerful detector is well established. It is based on concepts from two fields.

In machine learning, multiple detectors can produce better results when they make different types of errors and their outputs are combined carefully [12]. In distributed systems, a similar concept is called fault tolerance. It allows a system to continue working even when some components fail, as long as enough reliable components remain available.

Our approach combines ideas from both fields. Sections V and VI explain how these concepts are applied to improve exam security.

### ***I. Research Gap***

The existing literature demonstrates substantial progress in individual technologies, including secure document protection, custody management, anomaly detection, biometric authentication, answer script similarity analysis, distributed consensus, and explainable artificial intelligence. However, these techniques are generally implemented as independent solutions that operate without coordinated information sharing or joint decision-making. There remains a significant research gap in developing a unified framework capable of integrating these complementary technologies into a collaborative, lifecycle-aware examination security system. This study addresses that gap by proposing a multi-agent architecture in which specialized intelligent agents continuously monitor different stages of the examination process, share contextual evidence, perform semantic and textual similarity analysis of answer scripts, and support confidence-based collaborative decision-making while preserving human oversight for all final security decisions.

## **5. The Proposed Model**

The proposed framework is designed around a distributed multi-agent architecture in which complex examination security tasks are divided among specialized intelligent agents. Instead of depending on a single centralized system responsible for detecting every possible threat, each agent focuses on a specific security dimension and generates evidence related to its assigned area. This design improves scalability, reduces system complexity, and allows different types of security information to be analyzed independently before being combined for final decision-making.

The framework consists of multiple monitoring agents, including a Custody Agent, Watermark Verification Agent, Access Monitoring Agent, Behavior Analysis Agent, Web Monitoring Agent, and Answer Script Validation Agent.

The Custody Agent monitors the physical movement and handling of examination materials by tracking custody events, transfer records, and location-based activities.

The Watermark Verification Agent analyzes printed and digital copies to identify unauthorized duplication and trace leaked materials using embedded document identifiers. The Access Monitoring Agent evaluates authorized and unauthorized access attempts to secure areas, storage facilities, and digital repositories.

The Behavior Analysis Agent observes examination environments and identifies unusual patterns that may indicate malpractice or security violations [7].

The Web Monitoring Agent searches publicly available digital sources and communication channels for early indicators of leaked examination content.

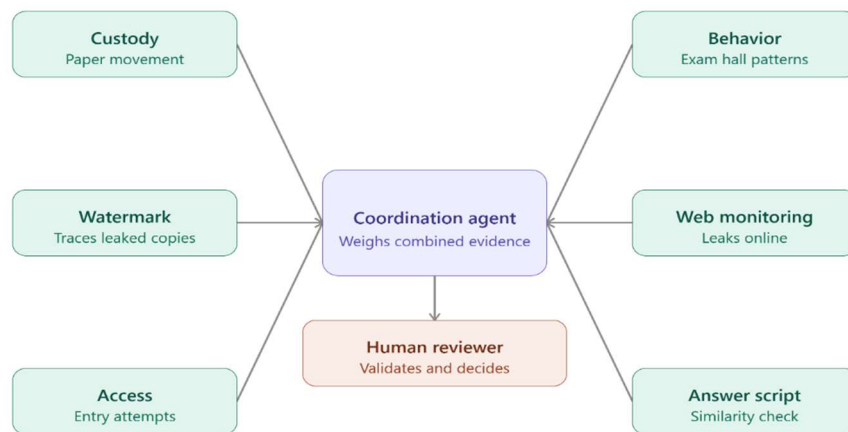
In addition to pre-examination security monitoring, the proposed model introduces an Answer Script Validation Agent for detecting potential copying and involvement among students. This agent applies semantic and textual similarity analysis to compare submitted answer scripts beyond exact word matching. Traditional text comparison methods may fail when students express identical ideas using different words or sentence structures. Therefore, the proposed agent evaluates both textual similarity and contextual meaning representation to identify unusually similar responses, repeated answer patterns, and possible coordinated copying behavior. The

generated similarity scores become additional evidence for the overall security assessment rather than an automatic accusation mechanism.

Each individual agent operates independently and does not have authority to cancel examinations, identify responsible individuals, or make final disciplinary decisions. Instead, agents forward their observations, confidence values, and supporting evidence to a central Coordination Agent.

The coordinator integrates information from multiple sources and determines whether the combined evidence indicates a potential security concern. This approach follows a collective intelligence model, where reliability is achieved through agreement among multiple independent observations rather than dependence on a single detection mechanism.

The importance of this coordination layer is that individual agents may occasionally produce incorrect results due to incomplete information, environmental conditions, or technical limitations. For example, a poor-quality image may prevent watermark verification, or a camera system may generate an incorrect behavioral alert. However, the probability of multiple independent agents generating consistent evidence for the same event is significantly lower. Therefore, the Coordinator applies a confidence-based evidence aggregation model to evaluate the overall situation and prioritize cases requiring human review.



**Fig.2 Architecture of the proposed multi-agent examination security framework and information flow between intelligent agents, coordination layer, and human decision authority.**

In Fig. 2, we propose splitting the job into small, specialized agents rather than one big system trying to do everything. Each agent watches one kind of signal and knows only about its own job. A Custody Agent watches where the paper physically is. A Watermark Agent checks printed and photographed copies against the fingerprint from Section IV-A. An Access Agent checks who unlocks doors and vaults, and when. A Behavior Agent watches the exam hall for the kind of unusual patterns described in Section IV-C. A Web-Watch Agent quietly searches forums and messaging groups for early signs that content has leaked. None of these agents can, on its own, stop an exam or accuse a person. That power sits only with the coordinator and, above it, a human reviewer.

Each agent scores its observations on a scale from 0 (normal) to 1 (highly suspicious), rather than making a yes/no judgment on its own. The coordination agent combines these scores into a single confidence value, giving more weight to agents that have proven reliable and less to those that have recently been inaccurate.

A security alert is raised only when enough agents agree, not when one agent's score alone crosses a threshold. This consensus requirement reduces false alarms caused by a single faulty or misinformed agent. The coordinator then forwards its combined evidence and confidence score to a human reviewer, who validates the alert and makes the final decision. The agents surface evidence, only humans act on it.

### Learning and Adaptation After Deployment

The agents continue to learn after they are deployed. When a human reviewer confirms whether an alert was correct or a false alarm, the system uses that feedback to update each agent's trust score. If an agent frequently makes mistakes, the coordinator automatically gives its results less importance. This reduces the need for manual adjustment.

Some agents are also retrained regularly using new, verified cases. These include the Behavior Analysis Agent and the Answer Script Validation Agent. Between training periods, the system checks whether an agent's accuracy is changing over time. This change is called **model drift**, which can happen when normal patterns become different from those used to train the model [14]. If an agent's performance falls below an acceptable level, the coordinator reduces its influence and sends it for human review.

## Protecting the Web Monitoring Agent from Manipulated Content

The Web Monitoring Agent uses a large language model to examine public web pages and social-media posts. An attacker could place hidden or misleading instructions on a webpage to influence the agent. This type of attack is called indirect prompt injection [15].

Several safeguards are used to reduce this risk. The agent treats all website content as untrusted information and does not follow instructions found on webpages. It is only allowed to search, compare, and summarize content. It cannot perform actions or make final decisions.

The agent must return its results in a fixed format containing only a similarity score and a source link. This prevents hidden instructions or unrestricted text from being passed to other parts of the system. Its access is also limited to the minimum tools and data needed for web monitoring.

Finally, the agent only sends evidence to the Coordination Agent. An alert cannot move forward unless several other agents provide supporting evidence. Therefore, even if the Web Monitoring Agent is misled, it can affect only one piece of evidence and cannot independently trigger a final decision.

## 6. How The Agents Decide Together

### A. One Combined Score

Each agent gives a score between 0 and 1 for what it observes, where 0 means "looks normal" and 1 means "looks very suspicious." If agent  $i$ 's score is  $r_i$ , the coordinator combines all agent scores into one overall score  $R$ :

$$R = w_1 \cdot r_1 + w_2 \cdot r_2 + \dots + w_n \cdot r_n$$

Here, each weight  $w_i$  reflects how reliable that agent has been recently, and all weights together add up to 1. Agents that have been accurate get slightly more influence over time, while agents that have made mistakes get less. This means the system automatically learns which agents to trust more, without needing manual retuning.

### B. Requiring Agreement Before Acting

A serious action, such as locking a vault or cancelling an exam session, is not taken just because the combined score  $R$  is high. It is taken only when enough agents independently agree. If there are  $n$  agents in total, and up to  $f$  of them might be wrong or compromised at the same time, the system requires support from more than half of  $(n + f)$  agents, and needs at least  $n \geq 3f + 1$  agents overall. For example, with five agents, this design can tolerate one faulty or compromised agent while the rest still reach a reliable decision. This idea is borrowed from well-established distributed-systems reliability principles, applied here to examination security.

### C. Combining Multiple Agents Works Better Than One Model

A group of small, specialized agents can perform better than one large model or a fixed checklist. Each agent focuses on only one task, such as checking custody records, verifying watermarks, monitoring doors, reviewing camera footage, searching the web, or comparing answer sheets. Because each agent has a clear and limited role, it can perform its task more accurately.

The agents also make different types of mistakes. For example, a watermark agent may fail because an image is blurry, while a camera agent may raise a false alert because someone moved unusually. Since the agents usually do not fail at the same time, combining their results can reduce the overall number of errors [12].

The system gives more importance to agents that have shown better performance and less importance to agents that are less reliable. The quorum rule in Section VI-B also requires agreement among multiple agents before making an important security decision. Unlike a fixed checklist, the system can adjust these weights as conditions change. This makes it more flexible and better able to respond to new attack methods.

### D. Tracing A Leak Back to Its Source

When a suspicious image is found online, it is compared against the watermark fingerprint of every printed copy, producing a match score between 0 (no match) and 1 (certain match). A copy is flagged only if the match score passes a carefully chosen threshold, since setting it too low risks false accusations, and too high risks missing real leaks. This narrows the search from thousands of possible sources down to a short, specific list.

### E. Detecting Answer Copying

Since students may copy an idea using different words, this framework checks both meaning and wording, not just exact text matches.

Each answer script  $S_i$  is converted into a meaning-based representation  $V_i$  using natural language processing. Semantic similarity between two scripts is measured using cosine similarity:

$$\text{Sim\_semantic}(i, j) = (V_i \cdot V_j) / (|V_i| |V_j|)$$

Textual similarity is measured as the overlap in shared words:

$$\text{Sim\_text}(i, j) = \text{Common Words}(i, j) / \text{Total Words}(i, j)$$

These two are combined into one copying-risk score:

$$C_{ij} = \alpha \cdot \text{Sim\_semantic}(i, j) + \beta \cdot \text{Sim\_text}(i, j)$$

where  $\alpha$  and  $\beta$  control how much weight is given to meaning versus wording, and always add up to 1 ( $\alpha + \beta = 1$ ). If this combined score  $C_{ij}$  exceeds a set threshold  $\theta_c$ , the system flags it as a possible copying case. Importantly, this flag is not treated as proof of cheating. It is simply evidence passed on for a human reviewer to check and decide.

## 7. A Simple Comparison and A Simulation

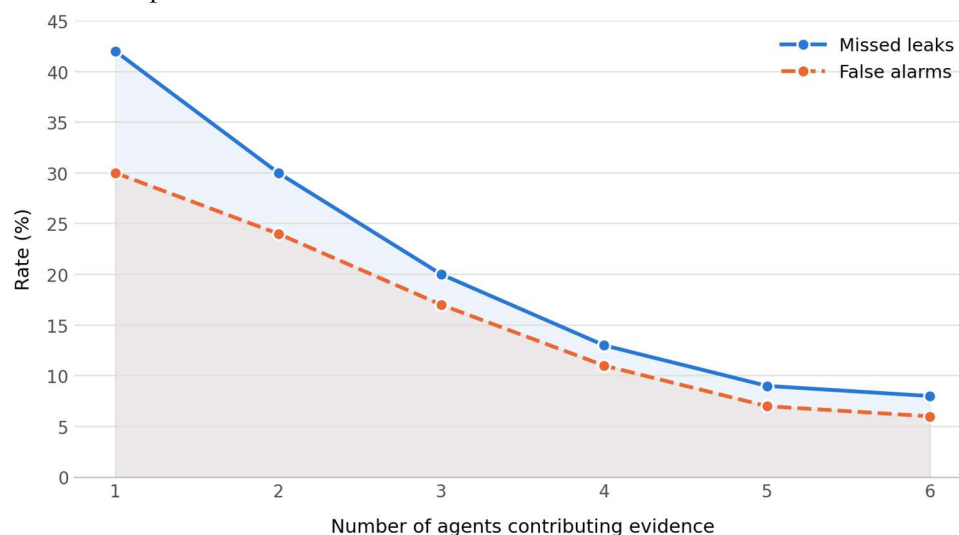
This section shows, in simple terms, how the proposed framework is expected to perform compared to today's methods.

**Comparing old and new approaches:** Traditional exam security tools each protect one part of the process, and mostly don't talk to each other. The proposed framework instead watches the whole exam lifecycle, lets agents share evidence, adapts to new threats, checks answer scripts for copying, makes decisions based on combined confidence, and still leaves the final call to a human.

**Table:1 Comparison of conventional examination security methods and the proposed multi-agent framework.**

| Capability                        | Conventional security | Proposed framework multi-agent |
|-----------------------------------|-----------------------|--------------------------------|
| Covers full exam lifecycle        | ✗                     | ✓                              |
| Agents share evidence             | ✗                     | ✓                              |
| Adapts to new threats             | ✗                     | ✓                              |
| Checks answer scripts for copying | ✗                     | ✓                              |
| Confidence-based decisions        | ✗                     | ✓                              |
| Human oversight                   | ✓                     | ✓                              |

**What the simulation shows:** As more agents contribute their observations, the system gets better at spotting real problems and avoiding false alarms. When agents have to agree with each other, one agent's mistake doesn't trigger a false alert, and the overall confidence in real alerts goes up. Adding more agents helps, but the benefit levels off after a certain point.



**Fig.3 Simulated results showing missed leaks and false alarms both decreasing as more agents contribute evidence, with the improvement leveling off around five to six agents.**

With only one agent, both missed leaks and false alarms start high, since there is no second opinion to confirm or correct a single observation. As more agents contribute evidence, requiring agreement among them filters out individual mistakes, so both error rates fall sharply. Beyond about five agents, the improvement levels off, showing that most of the benefit is achieved without needing a very large number of agents.

**Choosing the decision thresholds:** Two thresholds are important in this framework: the agreement threshold used by the coordinator (Section VI-B) and the similarity threshold, used by the Answer Script Validation Agent (Section VI-E).

We select these thresholds by testing several possible values using simulated data. For each value, we compare how many real leaks the system detects with how many false alarms it produces. We then select the value that provides the best balance between detecting leaks and avoiding false alarms. This approach is known as the Youden index in ROC analysis [13].

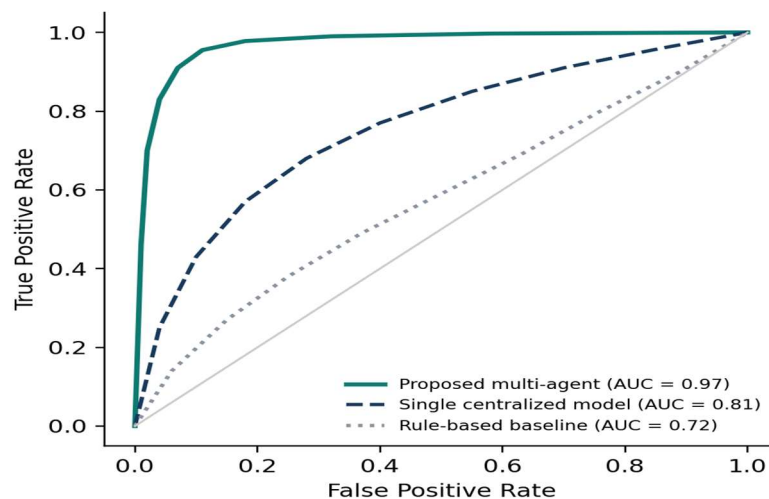
The thresholds are selected based on test results rather than guesswork. Institutions can also adjust them according to their needs. For example, an exam board that wants fewer false alarms can increase the threshold, although this may cause the system to miss some less-obvious cases. Therefore, these thresholds are flexible settings rather than fixed values.

**Detection Results Using Simulated Data:** Table 2 presents the system's performance using the same simulated data shown in Fig. 4. As expected from Section VI-C, the proposed multi-agent system performs better than both the rule-based system and the single centralized model.

The multi-agent system achieves higher precision, recall, F1-score, and ROC-AUC. It also produces fewer false positives and false negatives. The false-negative rate is especially important because it represents the percentage of real exam leaks that the system fails to detect.

**Table:2 Simulated detection metrics for the proposed framework versus a rule-based baseline and a single centralized model.**

| Metric               | Rule-based baseline | Single centralized model | Proposed multi-agent |
|----------------------|---------------------|--------------------------|----------------------|
| ROC-AUC              | 0.72                | 0.81                     | 0.97                 |
| Precision            | 0.61                | 0.74                     | 0.93                 |
| Recall (Sensitivity) | 0.58                | 0.77                     | 0.95                 |
| F1-score             | 0.59                | 0.75                     | 0.94                 |
| False Positive Rate  | 0.24                | 0.16                     | 0.05                 |
| False Negative Rate  | 0.42                | 0.23                     | 0.05                 |



**Fig. 4. ROC comparison of the proposed multi-agent fusion, a single centralized model, and a rule-based baseline (simulated evaluation).**

**Ablation study: each agent's contribution.** To understand the contribution of each agent, we repeated the simulation several times. In each test, we removed one agent and measured the performance of the remaining system. The results are presented in Table 3.

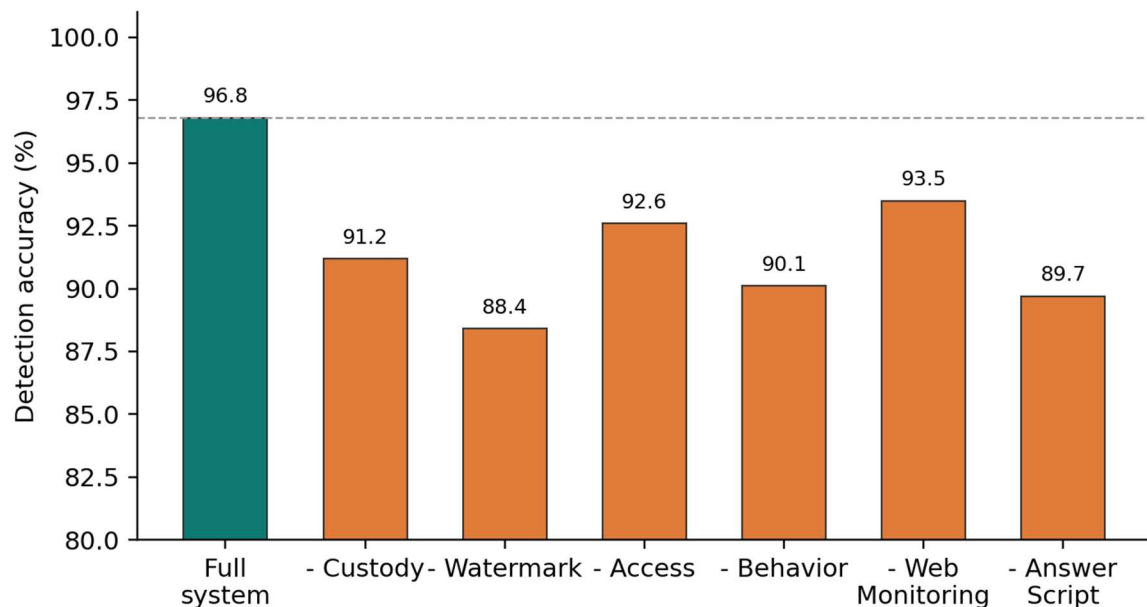
Removing the Watermark Verification Agent causes the largest reduction in accuracy because this agent helps trace a leaked exam paper to a specific printed copy. Removing the Answer Script Validation Agent causes the

second-largest reduction. This shows that checking answer scripts after an examination can identify suspicious activity that other agents may not detect earlier.

Accuracy decreases whenever any agent is removed. However, no single agent is responsible for the full system's performance. This supports the explanation in Section VI-C: the system is effective because all the agents work together and provide different types of evidence.

**Table 3. Ablation study: simulated detection accuracy when each agent is removed one at a time.**

| Configuration                          | Detection accuracy (%) |
|----------------------------------------|------------------------|
| Full system (all six agents)           | 96.8                   |
| Without Custody Agent                  | 91.2                   |
| Without Watermark Verification Agent   | 88.4                   |
| Without Access Monitoring Agent        | 92.6                   |
| Without Behavior Analysis Agent        | 90.1                   |
| Without Web Monitoring Agent           | 93.5                   |
| Without Answer Script Validation Agent | 89.7                   |



**Fig. 5. Ablation results: detection accuracy drops when any single agent is removed, but no one agent explains the full system's performance (simulated).**

## 8. Limitation and Future Research

The performance of individual agents depends on the availability and quality of operational data; for example, the Watermark Verification Agent may be affected by deliberate document modification, while the Behavior Analysis Agent requires representative behavioral datasets. Similarly, the Web Monitoring Agent is limited to publicly accessible information, and the Answer Script Validation Agent identifies semantic and textual similarities that indicate potential copying but should not be considered conclusive evidence without human review. In addition, deploying multiple intelligent agents increases computational and operational requirements compared with conventional security systems. Future research will focus on large-scale validation, improving semantic similarity models, enhancing adversarial robustness, and optimizing collaborative evidence aggregation to improve detection accuracy while maintaining transparency and human oversight. The same goes for the ROC-AUC, precision, recall, F1, and ablation numbers in Section VII, they come from simulation, not a real exam. They show the approach behaves the way the theory says it should, but they are not a promise of how it will perform in the real world. Testing these numbers against real, anonymized exam-security data is next on our list.

## 9. Conclusion

Examination security cannot be achieved through a single technology or isolated protection mechanism. Instead, it requires continuous monitoring, collaborative analysis, and coordinated decision-making across the entire examination lifecycle. This paper proposes a novel multi-agent AI framework that integrates specialized intelligent agents for document security, access monitoring, behavioral analysis, web surveillance, and semantic as well as textual similarity analysis of answer scripts within a unified architecture. Through collaborative evidence aggregation and human-in-the-loop decision support, the framework aims to improve the detection of examination paper leakage and copying while reducing false alarms and strengthening the overall integrity, transparency, and reliability of examination management systems. Although the current study presents a conceptual architecture supported by simulation-based analysis, it establishes a foundation for future real-world implementation and validation of intelligent examination security systems.

## Acknowledgement

The author sincerely thanks everyone who offered guidance, feedback, and encouragement during the course of this research, their insights and support were instrumental in shaping the final research paper.

## Conflict of Interest

The author declares no conflict of interest with respect to the research, authorship, or publication of this research paper.

## Funding Statement

The author received no financial support/ financial assistance or funding for the research, authorship, and/or publication of this research paper.

## References

- [1] Transparency International. (2013). Global corruption report: *Education*. Routledge.
- [2] Sozon, M., Alkharabsheh, O. H. M., Fong, P. W. (2024). Cheating and plagiarism in higher education institutions (HEIs).
- [3] Cox, I. J., Kilian, J., Leighton, F. T., Shamoon, T. (1997). Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12), 1673–1687.
- [4] Zyskind, G., Nathan, O., Pentland, A. (2015). Decentralizing privacy: Using blockchain to protect personal data. *In Proceedings of the IEEE Security and Privacy Workshops*. 180-184.
- [5] Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), Article 15.
- [6] Pan, S., Hoque, S., Deravi, F. (2022). An attention-guided framework for explainable biometric presentation attack detection. *Sensors*, 22(9), 3365. <https://doi.org/10.3390/s22093365>
- [7] Viola, P., Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 511–518).
- [8] Lamport, L., Shostak, R., Pease, M. (1982). The Byzantine general's problem. *ACM Transactions on Programming Languages and Systems*, 4(3), 382–401.
- [9] Castro, M., Liskov, B. (1999). Practical Byzantine fault tolerance. *In Proceedings of the 3rd Symposium on Operating Systems Design and Implementation* (pp. 173–186).
- [10] Ribeiro, M. T., Singh, S., Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).
- [11] Hallak, J., Poisson, M. (2007). Corrupt schools, corrupt universities: What can be done? UNESCO International Institute for Educational Planning.
- [12] Dietterich, T. G. (2000). Ensemble methods in machine learning. *In Proceedings of the International Workshop on Multiple Classifier Systems* (pp. 1-15). Springer.
- [13] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.
- [14] Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), Article 44.
- [15] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M. (2023). Not what you have signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. *In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*.