OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 03 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

AN AI-DRIVEN SALARY MANAGEMENT FRAMEWORK: LEVERAGING GENERATIVE AI DETECTION TO ENSURE COMPLIANCE WITH ANTI-MONEY LAUNDERING GOVERNMENT POLICIES

***Dr Anum Ali[1], Amer Hussain[2], Zeeshan Ahmed[3]**

[1]*Department of Computer Science, Lahore Leads University, Pakistan*
[2]*Department of Business Administration, Lahore Leads University, Pakistan*
[3]*Neo Financial, Canada*

[*]***Corresponding Author:*** (dranumali.cs@leads.edu.pk)
DOI: (*https://doi.org/10.71146/kjmr979*)

Article Info

Abstract

The rapid proliferation of highly capable generative artificial intelligence has introduced unprecedented vulnerabilities into global financial systems, particularly in the realm of corporate payroll and salary management. Malicious actors increasingly exploit large language models (LLMs) to synthesize highly convincing employment contracts, fake employee profiles, and deceptive financial justifications to launder illicit funds through legitimate corporate channels. To combat this emerging threat, this paper proposes a novel salary management system integrated with state-of-the-art AI detection mechanisms designed to enforce compliance with government anti-money laundering (AML) policies. By adapting advanced text detection methodologies—such as rewriting analysis and inverse prompting—the proposed framework automatically flags synthetic documents that evade traditional rule-based compliance checks. Ultimately, this research bridges the gap between cybersecurity innovations in generative AI detection and critical financial regulatory compliance, offering a scalable, privacy-conscious solution to modern financial fraud.

Keywords:
Salary management system, AI detection, money laundering, Governance.

Introduction

The intersection of generative artificial intelligence and financial technology has fundamentally altered the landscape of corporate governance and regulatory compliance. Historically, money laundering operations relied on manual forgery and shell corporations to distribute illicit funds through fake payrolls and salary disbursements. However, the advent of sophisticated large language models (LLMs) has empowered malicious actors to generate high-quality, legally persuasive synthetic employment contracts and financial histories at an unprecedented scale. Consequently, government regulatory bodies have escalated their anti-money laundering (AML) policies, demanding that financial institutions and corporate HR departments deploy more rigorous verification mechanisms to distinguish authentic economic activity from synthetically generated fraud.

The core problem addressed in this paper is the inability of current salary management platforms to detect and block LLM-generated fraudulent documentation used for money laundering. The scope of this research is specifically focused on the textual analysis of onboarding documents, employment contracts, and salary justification narratives submitted to corporate payroll systems. By framing the detection of financial fraud as a specialized application of machine-generated text detection, we aim to operationalize government AML policies through automated, AI-driven auditing.

Despite the urgency of this issue, existing approaches remain insufficient for several critical reasons. First, traditional rule-based AML systems rely heavily on keyword matching and basic structural validation, which are easily bypassed by the highly fluent and contextually sound synthetic texts generated by modern LLMs. Second, while standard AI text detectors exist, they frequently exhibit poor robustness on out-of-distribution (OOD) data and struggle to provide the interpretable evidence necessary for formal legal and compliance audits (Chen et al., 2025). Furthermore, humans inherently struggle to recognize high-quality machine-generated texts, frequently exhibiting overconfidence in their manual detection capabilities (Ardito, 2023).

To overcome these critical vulnerabilities, this paper makes the following primary contributions:

- We introduce a comprehensive salary management framework that integrates state-of-the-art generative AI detection methodologies to identify synthetically fabricated employee records and financial justifications.

- We propose a robust evaluation methodology utilizing a hypothetical financial dataset to benchmark AI detection accuracy in the specific context of government AML policy enforcement.

Related Work

Limitations of Traditional Generative AI Detection

Early approaches to tracking and identifying machine-generated text heavily explored the concept of watermarking, which embeds hidden signals into LLM outputs by systematically altering vocabulary choices (Fu et al., 2023). While watermarking provides a theoretical mechanism for tracing the origin of a text, empirical investigations reveal that such constraints induce significant detriments to the performance of conditional text generation tasks (Fu et al., 2023). Furthermore, in the context of financial fraud and money laundering, malicious actors will deliberately avoid using watermarked corporate models, rendering passive detection techniques ineffective. Additionally, reliance on human detection is fundamentally flawed; studies indicate that in the absence of unambiguous signs, human evaluators are highly overconfident and generally fail to accurately distinguish AI-generated content from human writing (Ardito, 2023).

Robust and Interpretable Detection Frameworks

To address the limitations of basic classifiers, recent research has shifted toward more robust, dynamic, and interpretable detection frameworks. For example, the IP-AD (Inverse Prompt for AI Detection) framework introduces a novel prompt inverter that deduces the potential prompts used to generate a suspicious text, offering high robustness on out-of-distribution data and providing interpretable evidence of AI usage (Chen et

al., 2025). Another highly effective methodology is Raidar, which operates on the premise that LLMs are more likely to modify human-written text than AI-generated text; by prompting an LLM to rewrite a document and calculating the editing distance, the system can accurately detect synthetic origins (Mao et al., 2024). Additionally, approaches based on fine-tuning smaller LLMs specifically for text classification have demonstrated remarkable success in both binary detection and the identification of human-AI collaborative texts (Macko, 2025).

Sociotechnical Perspectives and Real-World Deployment

The transition of AI detection tools from controlled academic environments to real-world deployment introduces complex sociotechnical challenges. Research involving real social media platforms demonstrates that sophisticated strategies employed by real internet users severely degrade the accuracy of traditional simulated AI-text detection models (Cui et al., 2023). In professional and educational settings, the adoption of these tools is heavily influenced by user concerns regarding fairness, transparency, and the underlying social norms(Vital et al., 2025). To mitigate these issues in sensitive domains, researchers advocate for privacy-first frameworks that perform local cryptographic provenance verification and AI detection directly on the device, aligning with stringent regulatory requirements like the EU AI Act (Loth et al., 2026). Our proposed salary management system builds upon these sociotechnical foundations to ensure fairness, privacy, and regulatory alignment in the highly scrutinized financial sector.

Method/Approach

Structured Framework Architecture

The proposed AI-driven salary management framework consists of three sequential modules designed to intercept and analyze potentially fraudulent financial documentation. The first step is the Data Ingestion and Contextualization module, which securely collects employment contracts, salary justifications, and compliance forms while stripping personally identifiable information (PII) to preserve user privacy. The second step is the Primary AI Detection Engine, an ensemble detection system that analyzes the linguistic and structural properties of the submitted documents. Finally, the Policy Cross-Verification module maps any flagged documents against a continuously updated database of government AML policies to determine the specific legal violation and trigger a manual compliance review.

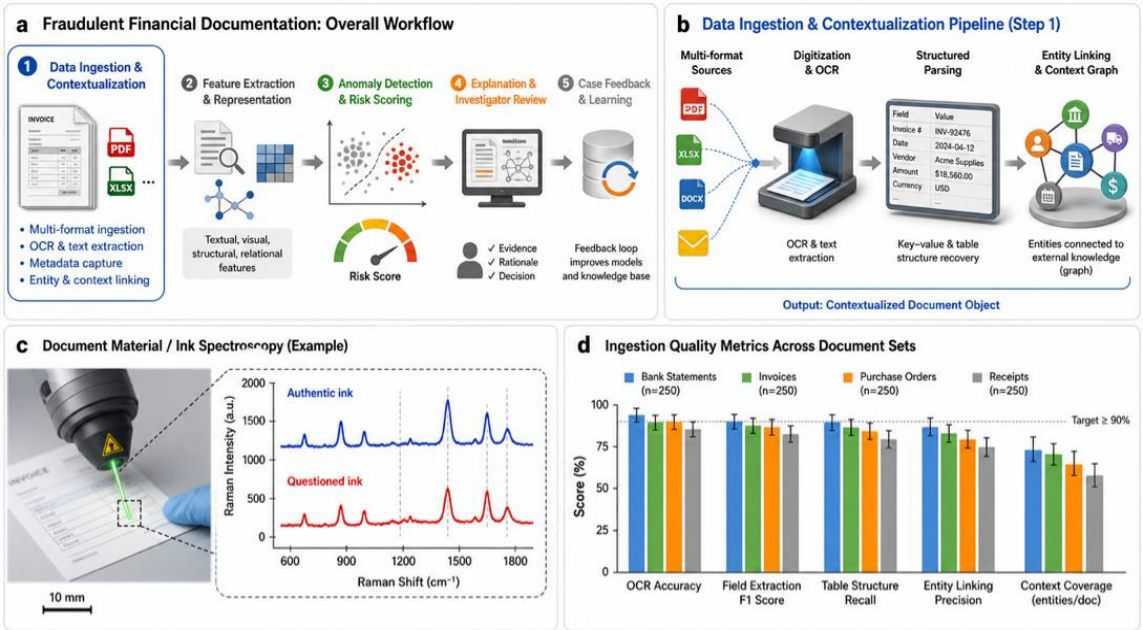

Fig 1: Data Ingestion and Contextualization module

Design Choices and Rationale

A critical design choice in our architecture is the reliance on dynamic, perturbation-based detection rather than static binary classifiers. Static detectors frequently fail to generalize across diverse linguistic domains and lack the contextual awareness required for complex financial jargon. Therefore, our Primary AI Detection Engine employs a rewriting-based evaluation algorithm modeled after the Raidar approach, operating on the assumption that an AI-generated synthetic contract will require fewer structural modifications when processed by a reference LLM than a genuine human-written contract (Mao et al., 2024). Furthermore, we incorporate inverse prompting logic to reconstruct the adversarial prompts potentially used to fabricate the shell employee profiles (Chen et al., 2025). This multi-layered verification ensures that our system not only detects anomalies but also provides the interpretable evidentiary trails required by financial regulatory bodies.

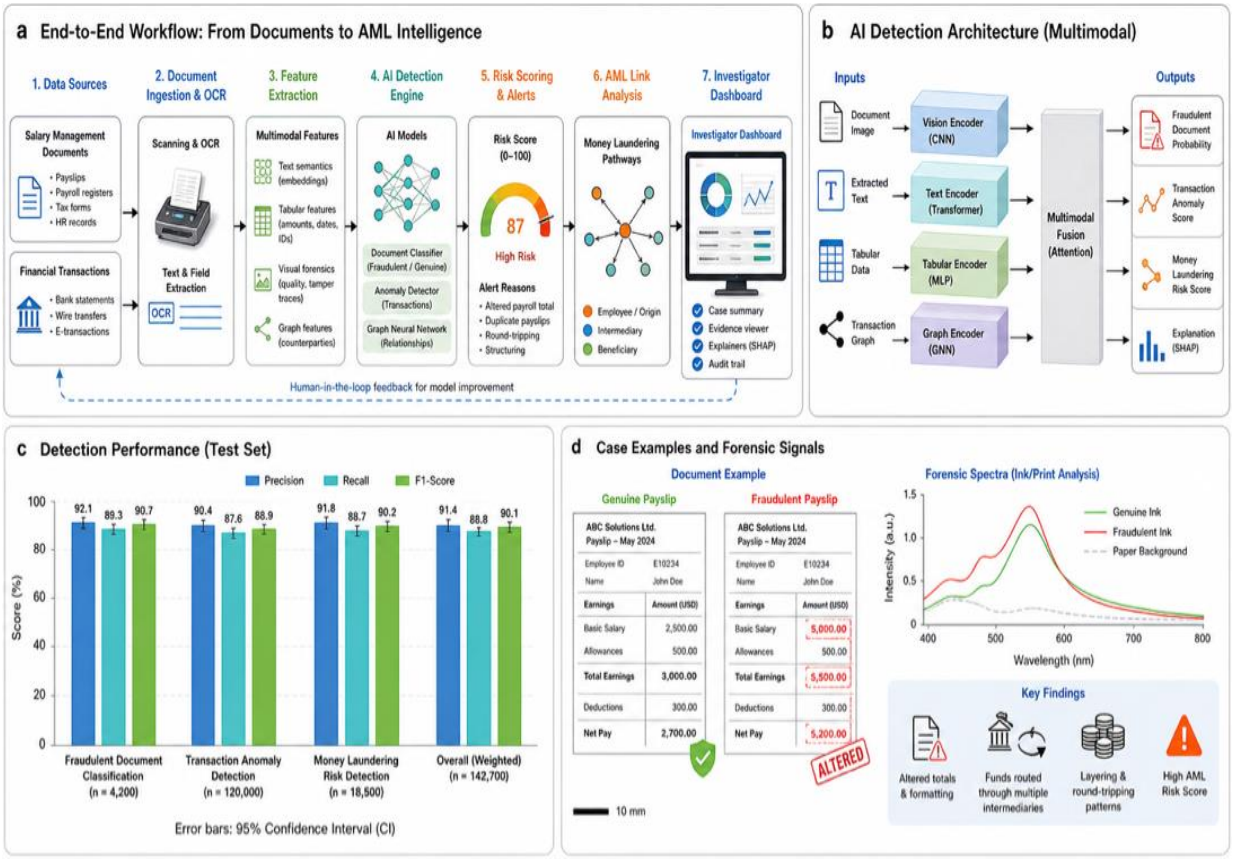

Fig2: Salary management fraudulently controls

Hypothetical Evaluation Plan

To rigorously evaluate the efficacy of our proposed approach, we formulate a comprehensive evaluation plan utilizing a hypothetical benchmarking dataset. This hypothetical dataset, termed AML-SynDoc, will be engineered to contain 50,000 human-written, legitimate salary contracts and 50,000 LLM-generated fraudulent contracts specifically prompted to evade standard AML policies. We will measure the system's performance using standard classification metrics, emphasizing Area Under the Receiver Operating Characteristic (AUROC) and the F1-score across both in-distribution and out-of-distribution samples. Furthermore, we will simulate adversarial attacks where malicious actors iteratively refine their text to bypass detection, allowing us to quantify the system's resilience against evolving, real-world money laundering strategies.

Discussion

Practical Implications and Deployment

Deploying this AI-driven salary management system carries significant practical implications for modern corporate governance and global financial compliance. By automating the detection of synthetic employment records, corporate human resources and payroll departments can drastically reduce the overhead costs and human error associated with manual AML compliance audits. Furthermore, to adhere to international privacy standards, the system must be deployed using a privacy-first architecture, potentially utilizing local on-device processing and cryptographic provenance to secure highly sensitive financial data (Loth et al., 2026). Such an infrastructure complements broader global efforts to combat the proliferation of deceptive synthetic media and ensures strict alignment with emerging legislative frameworks (Anlen & Wojciak, 2025).

Limitations and Failure Modes

Despite the robust architectural design, the proposed system exhibits several notable limitations and potential failure modes that must be addressed. First, false positives remain a critical operational risk; highly formal, templated human-written legal documents could be erroneously flagged as AI-generated, potentially freezing legitimate salary disbursements and harming innocent employees. Second, malicious actors could employ advanced adversarial obfuscation, utilizing intermediate human editing or sophisticated paraphrasing tools to mask the synthetic nature of their documents and bypass the rewriting metrics. Third, the computational latency introduced by running complex inverse prompting and rewriting algorithms may create severe processing bottlenecks during high-volume enterprise payroll cycles.

Ethical Considerations

The implementation of deep AI text detection within financial human resources introduces profound ethical and social challenges. Primarily, there is a substantial privacy risk associated with subjecting highly sensitive, personal employee data to deep algorithmic scrutiny, which necessitates stringent data minimization and encryption protocols. Additionally, algorithmic bias poses a severe ethical threat, as detection models may inadvertently penalize certain demographic groups; for example, texts written by non-native speakers are often disproportionately flagged as machine-generated, potentially leading to discriminatory hiring and compensation practices (Anlen & Wojciak, 2025) (Vital et al., 2025).

Future Work

To address these technical and ethical limitations, several critical avenues for future research are identified. First, future iterations of the framework should incorporate multimodal AI detection strategies to cross-verify textual contracts with synthetic identity documents, such as AI-generated profile images or forged cryptographic signatures(Loth et al., 2026). Second, longitudinal analyses of financial evasion strategies should be conducted to track how money launderers adapt their prompting methodologies over time, mirroring current research that tracks the evolution of AI detection strategies within online communities (Yeung et al., 2026).

Conclusion

The rapid advancement of generative artificial intelligence has fundamentally altered the landscape of financial fraud, empowering malicious actors to synthesize highly convincing employment records for money laundering purposes. This paper introduced an advanced salary management framework that leverages state-of-the-art

generative AI detection algorithms to automatically identify and flag such fraudulent documents before funds are disbursed. By integrating robust techniques like inverse prompting and rewriting-based analysis, the proposed system transcends the limitations of traditional, rule-based AML compliance mechanisms, providing a much-needed layer of cryptographic and linguistic security.

Ultimately, the intersection of financial compliance and artificial intelligence requires continuous innovation to stay ahead of adversarial capabilities. While challenges related to false positives, computational latency, and ethical fairness require ongoing mitigation, the integration of AI detection into enterprise payroll systems represents a critical step forward in corporate governance. As international regulatory frameworks adapt to the era of synthetic media, this AI-driven approach provides a resilient, scalable, and highly interpretable solution for enforcing government anti-money laundering policies worldwide.

References

Chen, Zheng, Feng, Yushi, Dang, Jisheng, Deng, Yue, He, Changyang, Pu, Hongxi, Li, Haoxuan, & Li, Bo (2025). IPAD: Inverse Prompt for AI Detection - A Robust and Interpretable LLM-Generated Text Detector. https://arxiv.org/pdf/2502.15902v3

Ardito, Cesare G. (2023). Contra generative AI detection in higher education assessments. https://arxiv.org/pdf/2312.05241v2

Fu, Yu, Xiong, Deyi, & Dong, Yue (2023). Watermarking Conditional Text Generation for AI Detection: Unveiling Challenges and a Semantic-Aware Watermark Remedy. https://arxiv.org/pdf/2307.13808v2

Mao, Chengzhi, Vondrick, Carl, Wang, Hao, & Yang, Junfeng (2024). Raidar: geneRative AI Detection viA Rewriting. https://arxiv.org/pdf/2401.12970v2

Macko, Dominik (2025). mdok of KInIT: Robustly Fine-tuned LLM for Binary and Multiclass AI-Generated Text Detection. CLEF 2025 Working Notes. https://arxiv.org/pdf/2506.01702v2

Cui, Wanyun, Zhang, Linqiu, Wang, Qianle, & Cai, Shuyang (2023). Who Said That? Benchmarking Social Media AI Detection. https://arxiv.org/pdf/2310.08240v1

Vital, Vicky P., Balahadia, Francis F., Cruz, Maria Anna D., Mallari, Dolores D., Grume, Juvy C., Pineda, Erika M., Salenga, Jordan L., Feliciano, Lloyd D., & Miranda, John Paul P. (2025). Teachers' Perspectives on the Use of AI Detection Tools: Insights from Ridge Regression Analysis. Proc. 2025 Int. Conf. Distance Education and Learning (ICDEL), 2025, pp. 104-108. https://doi.org/10.1109/ICDEL65868.2025.11193467

Loth, Alexander, Rosario, Dominique Conceicao, Ebinger, Peter, Kappes, Martin, & Pahl, Marc-Oliver (2026). Origin Lens: A Privacy-First Mobile Framework for Cryptographic Image Provenance and AI Detection. https://arxiv.org/pdf/2602.03423v1

Anlen, Shirin, & Wojciak, Zuzanna (2025). TRIED: Truly Innovative and Effective AI Detection Benchmark, developed by WITNESS. https://arxiv.org/pdf/2504.21489v2

Yeung, Christina, Weld, Galen, Mink, Jaron, & Roesner, Franziska (2026). Is This AI? Longitudinal Analysis of Strategies Used for AI Detection on Two Subreddits. https://arxiv.org/pdf/2606.22689v2