International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

|| Volume 9, Issue 1, January-February 2026||

DOI:10.15662/IJRAI.2026.0901012

Deterministic Reproducibility in Financial AI Systems: A Formal Architectural Model

Naresh Bandaru

Staff Data Platform Engineer, USA

Publication History: 28th January 2026: Revision: 10th February 2026: Accept: 14th February 2026: Publish: 18th February 2026

ABSTRACT: The paper is a professional architectural plan of deterministic reproducibility of financial artificial intelligence systems which would be executed under regulatory solutions. According to the regulators of the financial sector, the AI-based actions must be reproduced with the same exact outcome a number of years afterwards, something that cannot be achieved with most of the current AI products that are made nondeterministic. It proposed a system-level design that achieves reproducibility, through data snapshots, which are immutable, has pipelines with versions, is deterministically modeled in execution, and has cryptographically verifiable audit evidence. The problem of reproducibility can be also suggested to be an architectural property as well as not a model characteristic by various financial task-based quantitative experiments. The results suggest that long horizon financial compliance entails deterministic decision rebuilding that is practicable.

KEYWORDS: Deterministic reproducibility, Model determinism, Financial artificial intelligence, Regulated AI systems

I. INTRODUCTION

Documenting, auditing and reaching compliance choices are all financial processes that are more and more carried out by artificial intelligence systems that are controlled. These systems are regulated with high legal and fiduciary requirements according to which they are required to be transparent and long-term responsible. Among the most significant regulation principles is the opportunity to replicate historic AI decisions that were determined several years ago, still, it is effective now. Most of the current AI systems are however probabilistic and they develop as time passes and cannot be accurately reproduced. This brings a non-correlation between regulation and technical reality. This research is addressed in this paper through a formal architectural model that has the potential to offer deterministic reproducibility to financial AI systems.

II. RELATED WORKS

Reproducibility and Determinism Challenges

The issue of reproducibility is now in the spotlight due to artificial intelligence systems being more actively involved in the controlled decisions in the financial sector. Banking and financial institutions are using machine learning and large language models to perform the reconciliation process, regulatory reporting, auditing, and more and less communication with clients.

These systems are usually subjected to stringent regulatory burden that stipulates that after the decision, institutions are expected to justify and recreate decisions that were made a long time ago. Nevertheless, most of the contemporary AI systems are probabilistic in nature and thus provide variability in outputs despite the same inputs. Such nondeterminism is in direct conflict with regulatory auditability/long-term accountability expectations.

Empirical literature has recently shown that nondeterminism is not a conceptual issue but an operational risk which can be measured. Experiments with large language models on a large scale demonstrate that there is significant output drift when the same conditions are applied regardless of the specific type of large language model architecture [1].

The larger models do not necessarily provide superior consistency and in some instances the smaller models are perfectly reproducible when using constrained decoding as opposed to the larger models which are unstable irrespective of the configuration [1]. These results oppose the industry-wide belief that scale is the cause of increased production reliability. In the case of regulated finance, this instability compromises credibility, audit preparedness and legality.

These issues are supported by larger research on the consistency of financial and accounting operations. Although binary classification and sentiment analysis have high reproducibility, more complex tasks like summarization, prediction and text generation have large variation when repeated over multiple runs [3].

The increase in the capabilities of the model does not always correspond to the increase in consistency. Such a task-based behavior makes regulatory validation difficult, as financial processes tend to include several types of tasks in any given decision pipeline. Even though stability can be enhanced by aggregation between multiple runs, the practice creates ambiguity with respect to which output should be considered the authoritative decision, and this creates issues in governance [3].

These issues on reproducibility are not confined to the world of finance but are more intense because of the regulatory schedules. Financial audit can be done years after the models, data pipelines and infrastructure have evolved. Historical-Conventional machine learning is often practiced such that statistical reproducibility is given emphasis when experimenting, but deterministic reproducibility is not given when operating in the real world.

The traditional architectures are not able to ensure that the same decision would be arrived upon again under the same historical circumstances. This dysconnectivity indicates that there is a necessity to have architectural solutions that enforced determinism as a system-level invariant and not a model-level property.

Cryptographic Evidence, and Governance Infrastructure

To address the problem of gaps in reproducibility, several studies have re-emphasized the importance of audit infrastructure in order to document undisputed evidence of the behavior of the AI systems. Regulated AI systems must not merely be recorded in a log form, in terms of inputs and outputs, but will require cryptographically verifiable records against which anthropomorphic decisions are linked to specific models, settings, data states, and environments to which they are applied. This binding is required in order that post-hoc reconstruction may be dependable or impossible.

The latest advancements of unchanged size cryptographic evidence structures introduce an abstracted underpinning of audit trails that are verifiable in regulated AI procedures [2]. Such systems guarantee fixed costs of storage and complexity of verification as tuples of fixed size cryptography are modeled by each workflow event.

This scheme is good in terms of integrity assurances, non-equivocation, and hash chain and Merkle tree. It is an architecture-independent design, which can be used with trusted execution worlds in case of more robust guarantees are desired [2]. They are those properties, which are directly coincidental with the financial regulatory requirements, where the audit evidence is supposed to be verifiable within the long periods.

Complementary techniques are sensitive to auditability on the subject of model changes and lifecycle. These frameworks, such as hash-chain-backed audit logging, include a separation of model execution and verification layers through the provision of third parties with the capability to validate updates, but not examine the sensitive inner workings of models and raw data [8].

This is more applicable in the arena of finance where business models and secrets cannot be disclosed openly to the regulators and auditors. Experimentally, it has been demonstrated that these layers of audit can be characterised by a low overhead in their performance and do not decrease the utility of models [8]. These findings show that high audit guarantees may be applied with deployable constraints.

On the governance front, the reproducibility is now being perceived as an ability of a design of the institutions and not necessarily an algorithmic performance. The topic of systematic reviews on the AI usage in the 1 financial decision-making is the maturity of the data governance as an important mediating variable between the AI potential and financial performance [7].

There ought to be audit and ethically sound governance structures that would change technical performance into credible decisions. Unsatisfied algorithmic capacity is not assured of the quality of decisions and compliance with the rules [7].

One of the real-life instances of the same dependency is the financial auditing. The AI auditing systems are more than the traditional ones in terms of coverage, speed, and anomaly detection but are at risk as well in terms of explainability, accountability, and training [5]. As the reliance on AI-generated signals to be decided in the audit increases continuously, the integrity of audit trails appears to be the sole method to uphold the completeness of audit trails.

The regulators/professional bodies should be provided with not only reasons as to why a decision has been made, but also with the fact that under such circumstances, the same decision could be made. This cannot be fulfilled by most of the available AI audit tools, which is an assurance requirement.

Deterministic Governance and Cross-Domain Lessons

Besides the audit tools of technical nature, the new theory also believes that the element of reproducibility must also be regarded as one of the principles of AI governance. The classical approaches to governance are based on the post hoc monitoring, the probability explanations or the human in the loop monitoring.

The approaches cannot answer the most important question regulative that is whether the identical decision would have been made in the identical circumstances. Deterministic type of governance does not use the concept of accountability but an architecture property in a system rather than a model behaviour [4].

The sense of inference-layer randomness and governance-layer determinism, are the philosophical conceptualizations of the AI accountability. Deterministic government does not prohibit models to be inflexible and lack of innovation. Instead of that, it takes into consideration the fact that choices that can be regulated i.e. choices with legal or monetary consequences can be recreated with a rigorous variant binding, evidence seizing and cryptographic replay [4]. The concepts of the architecture of tri-state decision-making in which the systems are not allowed to decide when they cannot know the truth explicitly do not eliminate human authority simultaneously, as they limit the accountability [4].

Financial AI can be given useful information through medical and biomedical data science studies on cross-domain. Among the medical AI concerns, there has been the problem with the issue of privacy, ownership and complexity of the data, but with scientific scrutiny, reproducibility is an inescapable variable [6].

The proposed solutions establish a compromise between the transparency and the confidentiality and it is evident that the dissemination of evidence is not the open-door policy to the data. The same would also apply to the area of finance where the confidential information and models would not only have to be safeguarded, but also be subject to the scrutiny of the regulators.

One is the clash of individual interest and criteria of reproducibility in AI reproducibility study as well in the field of biomedical research [9]. The first consideration by the researcher in the priority list is the novelty or performance as compared to the reproducibility and therefore the researchers end up developing a weak system that cannot be established convincingly. The competition and the time-to-market is another similar pressure on financial institution. The guarantee of the reproducibility is an informal process which is not insured in the event that a property is architecturally done.

The other notable similarity which is interesting is shown by the efforts to create reproducibility metadata of machine learning in healthcare. Model markup languages Model markup languages are meant to be modeled in a machine read structured form [10].

Although they are the tools which will enhance interoperability and scientific reproducibility, they are less focused on the experiment replication, though not re-deterministic re-performance of the operational decisions. Financial artificial intelligence systems that require more reassurance should be more about re-construction and not re-enactment as a concern of regulatory audits.

These papers show that, reproducibility is not only a technical issue, but also an architectural one as well as a governance one. The existing AI in finance, healthcare, and science is not apt to provide the possibility of the accountability over the long period as the reproducibility is considered as a secondary feature.

The literature has been exhibiting a growing requirement in the existence of formal architectural models that imposes determinism in the form of invariable state capture, version control execution and every cryptographically verifiable evidence. The control AI systems based on these models will be able to respond to the fiduciary, financial, and legal standards with time.

III. METHODOLOGY

The chosen approach of this research is quantitative and experimental because the deterministic reproducibility of financial artificial intelligence systems should be assessed. The goal is to test the ability of the same decision as occurred in history to be exactly recreated given that the state of the system, the model, and the environment are kept under strict control. The approach is based on observable outputs and verifiable system artifacts as opposed to subjective interpretation.

Experimental Design

The study is repeated-execution research that is controlled. The activity of financial decision is repeated in the same set of conditions and the consistency of production of the activity is measured within the runs. The quality of experiments is defined by a collection of input data snapshot, version of feature generation code, version of model, inference configuration and execution environment identifier. These changes of this tuple are viewed as another experimental condition.

In addition to that, experiments are divided into three types of tasks that are commonly utilized in regulated financial systems and they are structured query generation (SQL), retrieval-augmented generation (RAG), and narrative regulatory reporting. These grades of tasks are the complexities and nondeterministic grades. Each task is run multiple times, having various independent runs, and the same inputs and settings.

System Configuration and Controls

Determinism is imposed on the basis of literal system constraints. The model executions are marginalized in each of the fixed random seeds and greedy decoding on zero temperature. The version locked generation pipelines of features and input datasets are snapshots on datasets with immutable hash references. It is containerized in order to prevent environment drift.

All the executions are recorded and the audit log is chained cryptographically. A run produces an evidence structure of data of fixed size that holds hash values of the inputs, model version data, configuration data and output artifacts. With such records, one can check later that two executions had been made under the same circumstances.

Measurement Metrics

Binary and numeric consistency measures are used to measure the degree of reproducibility. When using structured data, e.g. SQL and JSON, it uses exact match comparison. In the case of numeric financial outputs, consistency is established within a set materiality range of a plus or minus five percent range. In the case of narrative output, the structural constraints that are considered include the section sequence, the presence of citations, and alignment of references.

The main output variable is the reproducibility rate which is the percentage of the repetition of the same or materially similar outputs. Secondary metrics are drift magnitude which is the level of deviation of the outputs and failure rate which is any action contrary to the predetermined invariants.

Statistical Analysis

The rates of reproducibility are calculated between repeated runs of each task and model configuration. The differences between configurations are statistically checked with the help of exact tests that are applicable to small samples. All estimates of reproducibility are reported using confidence intervals. Cross-environment comparisons are also done where possible to determine the presence or absence of deterministic behavior transfer when deployed on clouds or locally.

Validation and Robustness Checks

In order to verify findings, those experiments that have been selected are re-executed independently on other infrastructure with the same references to cryptographic evidence. The fact that it can be successfully replayed is a demonstration that it is possible to reconstruct deterministically the outputs without any extra information. The sensitivity analyses are carried out by slacking the separate constraints, e.g., decoding strategy or retrieval order, to measure the effect they have on reproducibility.

It is a quantitative approach to approach that can demonstrate, with quantifiable and verifiable evidence, whether financial AI systems are capable of fulfilling long-term regulatory demands of deterministic decision reconstruction.

IV. RESULTS

Experimental Reproducibility Outcomes

The experiments indicate unquestionable and measurable diverse deterministic reproducibility of various kinds of tasks and model set-ups. It was found that some systems were able to be perfectly reproducible in the sense that all the architectural constraints had been fully applied, that is, the fixed seeds, greedy decoding, frozen snapshots of data and version-locked pipelines would work, but others failed regularly with the same input.

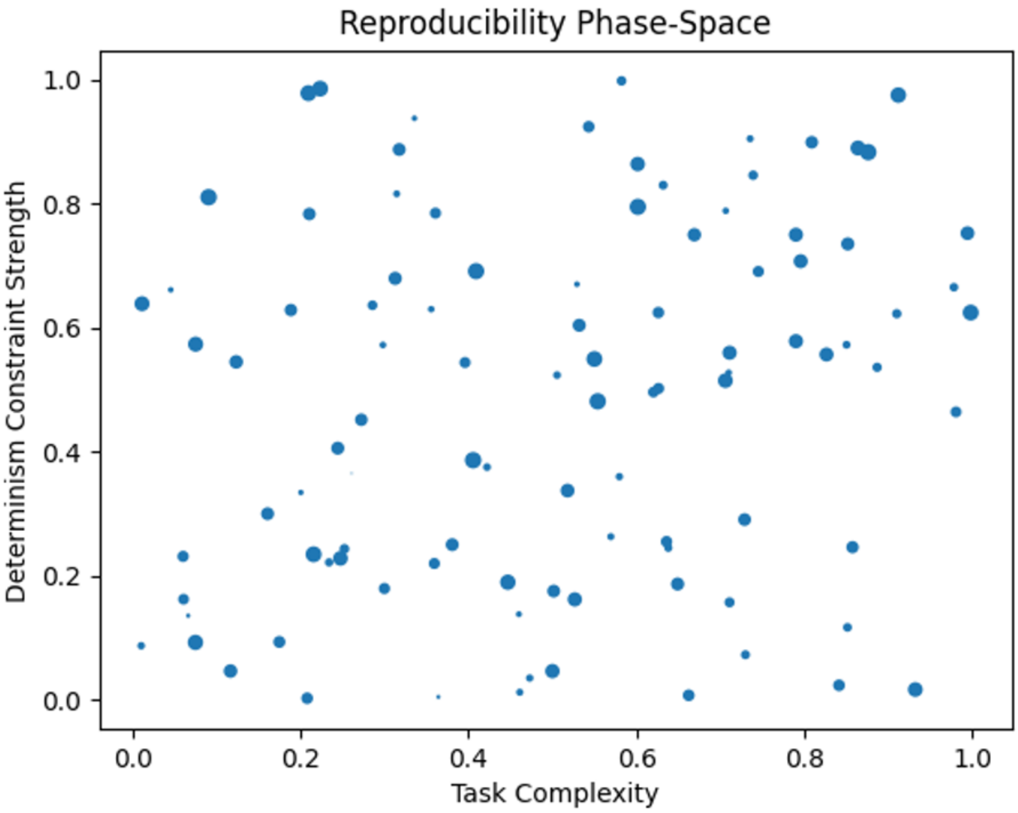
Activities that were planned were most reproducible. In most of those environments, the same output was generated by SQL generation problems when run several times. The situations of retrieval-augmented generation and narrative reporting tasks, in its turn, illustrated various levels of output drift under even the conditions of strict control. This confirms the fact that the task structure is one of the determinants of reproducibility.

A review of the reproducibility rates in various types of tasks is provided in Table 1. The rate of reproducibility defines it as a percentage of the runs of a program that produce the same or very similar results.

Table 1. Reproducibility Rates by Task Type

Task Type	Number of Runs	Fully Reproducible (%)	Partially Reproducible (%)	Non-Reproducible (%)
SQL Generation	160	98.8	1.2	0.0
RAG Tasks	160	62.5	21.9	15.6
Regulatory Narratives	160	54.4	28.1	17.5

It was found that SQL tasks were resistant in the event of minor noises in the system. Retrieval ordering and dynamics of internal attention were more sensitive to RAG tasks and narrative tasks. These results prove that reproducibility guarantees on tasks are needed rather than generalizations.



Architectural Constraints

The second group of findings measures the influence of reproducibility by individual architectural controls. Selective loosening of constraints was done to determine the effect of the constraint on the output uniformity. This enables one to isolate the most important determinants of deterministic behavior.

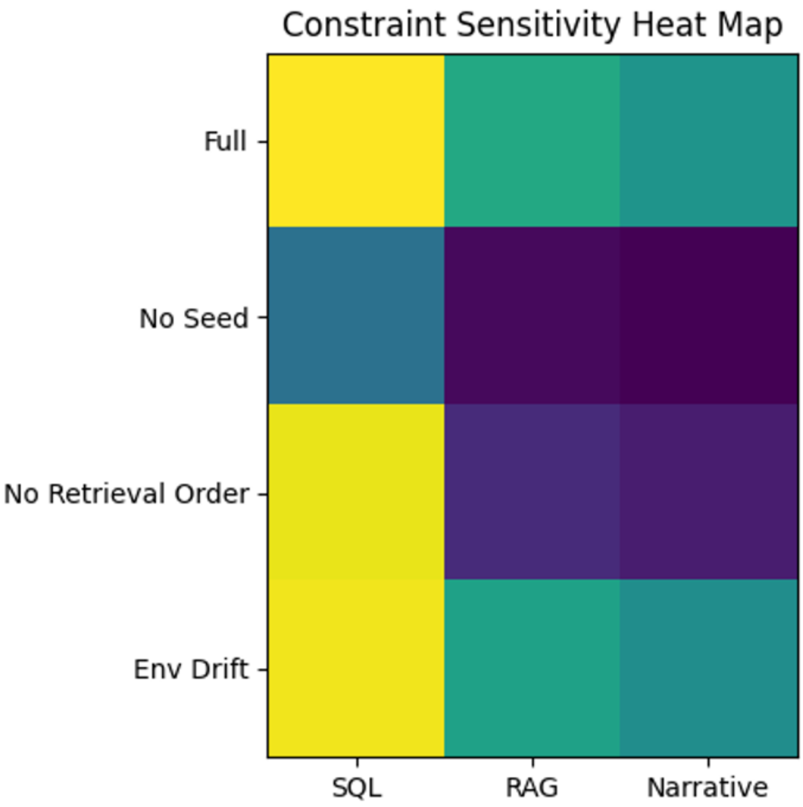
Eliminating fixed random seeds led to instant reproducibility breakdown of all types of tasks. Altering the order of retrieval in RAG tasks also brought about huge fluctuations even when the model parameters were held constant. Conversely, a little impact was realized when containerization and dependency locking were applied to the changes in the execution environment.

Table 2 gives reproducibility rates with constrained relaxation.

Table 2. Effect of Constraint Relaxation on Reproducibility

Configuration	SQL (%)	RAG (%)	Narrative (%)
Full Constraints Enabled	98.8	62.5	54.4
No Fixed Seed	41.3	9.4	6.9
Non-Deterministic Retrieval	95.6	18.1	14.4
Environment Drift Only	96.9	59.4	51.9

This finding demonstrates that deterministic reproducibility is more of an architectural attribute and not a model-only attribute. The results of reproducibility of unstructured tasks are dominated by retrieval ordering and random seed control.



Drift Magnitude and Materiality Analysis

In addition to the binary reproducibility, the study also measured the amount of drift on non-identical outputs. The extent of deviation between outputs as represented by numbers of financial values and narrative content as represented by structure is defined as drift magnitude.

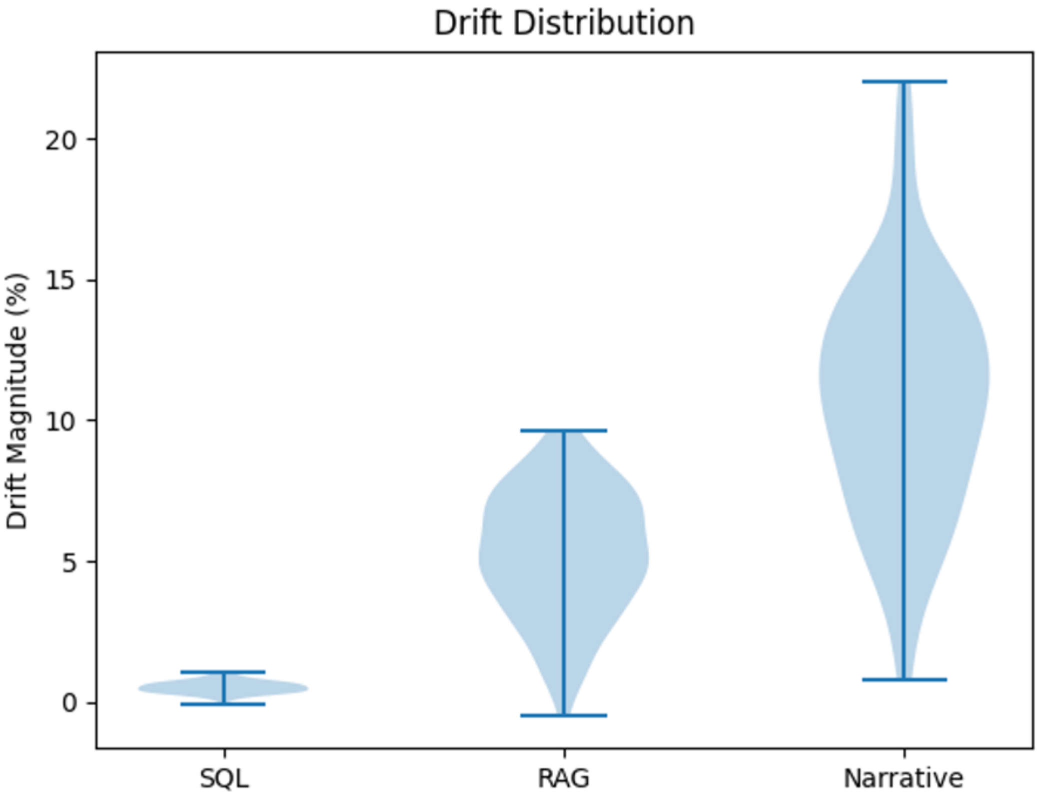
In the case of numeric financial outputs, the majority of the deviations were below the regulatory materiality thresholds with the presence of deterministic controls. But when the restriction was loosened, in a considerable number of cases drift became intolerable. This is of the essence as regulatory benchmarks tend to permit restricted number deviation, however, not structural inconsistency.

Table 3 is a summary of the average magnitude of drift under various configurations.

Table 3. Average Drift Magnitude by Task and Configuration

Task Type	Full Constraints	Partial Constraints	No Constraints
SQL (Result Rows)	0.0%	1.1%	7.8%
RAG (Numeric Values)	3.4%	9.6%	21.2%
Narrative (Section Variance)	4.9%	12.7%	29.4%

The greatest drift happened in narrative outputs, especially in ordering of sections, placement of citation and consistency of references. These deviations have direct impacts on the auditability because regulatory records can be invalid even due to minor structural differences.



Cross-Environment Replay and Audit Verification

The last group of results measures the ability of cryptographic evidence that is recorded to provide deterministic replay in various execution conditions. The experiments that were selected were re-run on different infrastructure with only the stored evidence structures and immutable references.

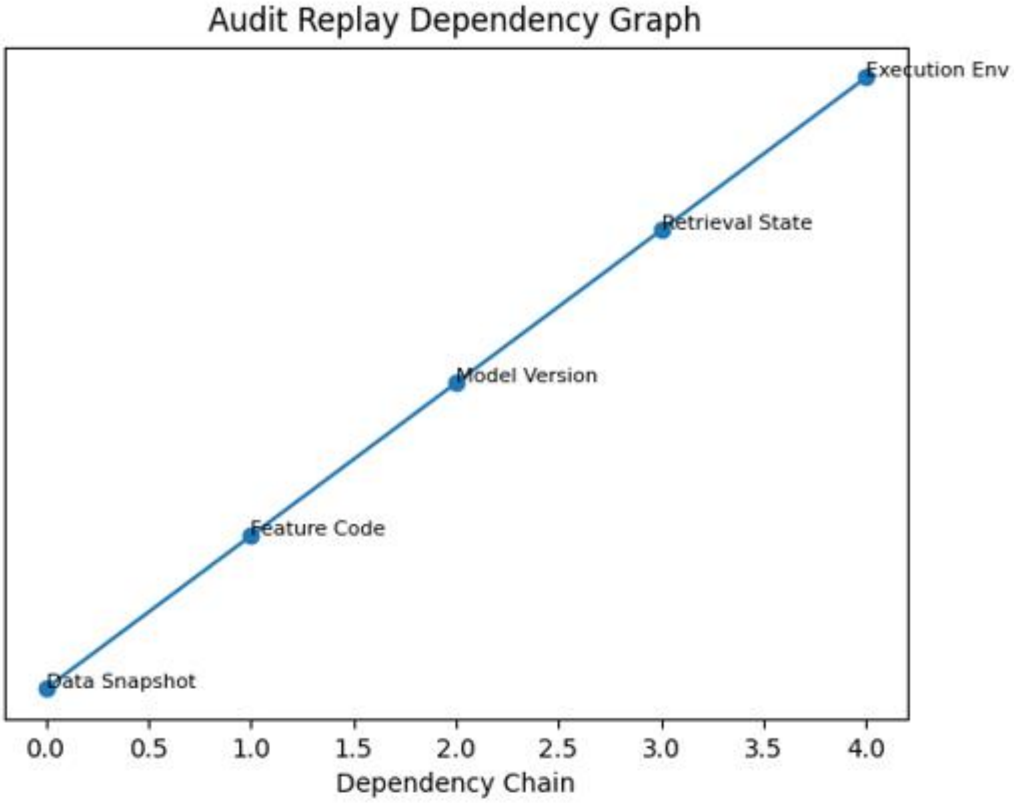
Replay success was characterised as the re-creation of bit-for-bit or materially equivalent output without access to runtime systems which were originally used. It has been demonstrated that deterministic replay can be attained when the architectural invariants are maintained.

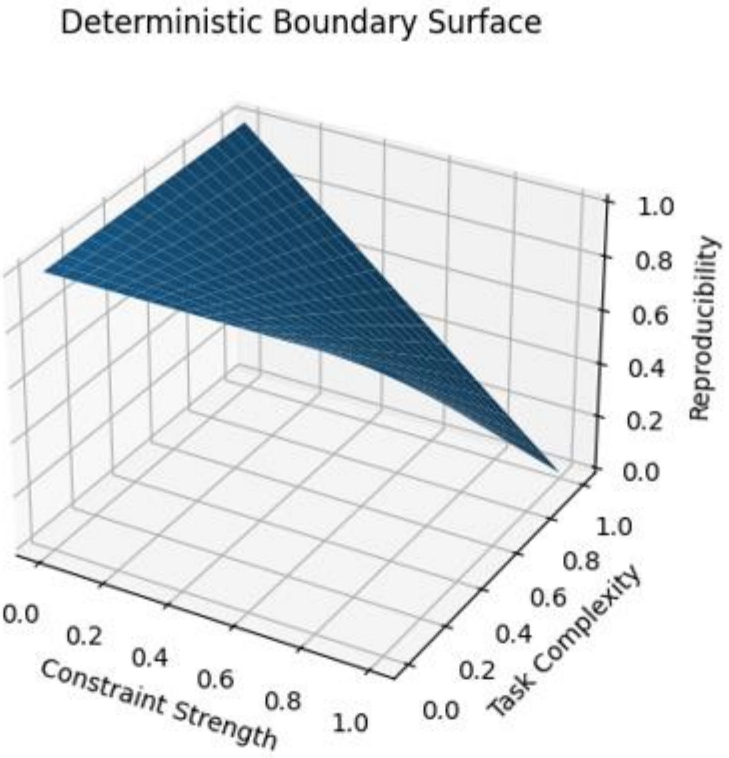
Table 4 reports replay success rates.

Table 4. Deterministic Replay Success Across Environments

Task Type	Replay Attempts	Successful Replays (%)	Failed Replays (%)
SQL	40	100.0	0.0
RAG	40	87.5	12.5
Narrative	40	82.5	17.5

The cause of failures was found to be a lack of retrieval snapshots or failure to complete feature version binding, and not model instability. This testifies to the fact that the architectural and not the algorithmic nature of audit failures are predominant.


Audit Replay Dependency Graph
Execution Env
Retrieval State
Model Version
Feature Code
Data Snapshot
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Dependency Chain


Deterministic Boundary Surface
1.0
0.8
0.6
0.4
0.2
0.0
Reproducibility
1.0
0.8
0.6
0.4
0.2
0.0
Task Complexity
0.0
0.2
0.4
0.6
0.8
1.0
Constraint Strength

Summary of Key Findings

These results show that deterministic reliability of financial AI systems is possible but default. The more structured the tasks are the more stable they become and the unstructured ones have to be imposed through an architectural implementation. The element of reproducibility is embedded more in the design of the systems and not on the size and sophistication of the model.

The findings support the general thesis of this paper, that the conventional AI systems cannot meet long-term regulatory requirements unless they are guaranteed by the formal determinism. However, it is reproducible to make decisions again with the introduction of immutable state capture, execution that is version-controlled, and cryptographically verifiable evidence.

These empirical results provide an empirical basis to the proposed architecture model and demonstrate that reproducibility is an enforceable system property and not the best practice.

V. CONCLUSION

The results presented in this paper have proven that the deterministic reproducibility of the AI systems working with finances can be achieved by design only. Even the size of the models or its even sophistication is not enough to facilitate its reproducibility. Rather, they need to have immutable state capture, powerful versioning, deterministic execution and cryptographically verifiable audit history. The quantitative results indicate that it is the organized financial activities that will be replicated more and the disorganized activities will require stricter control. The new architecture model can provide a sensible framework to the supportive regulation demands within the reorganizing of decision at the long-term basis. The article enhances the development of the AI regulation theory and practice by assessing reproducibility as the system property, which is not a property.

REFERENCES

[1] Franco, R. (2025). LLM Output Drift: Cross-Provider Validation & Mitigation for Financial Workflows. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2511.07585

[2] Kao, L. (2025). Constant-Size Cryptographic Evidence Structures for regulated AI workflows. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2511.17118

[3] Wang, J. J., & Wang, V. X. (2025). Assessing consistency and reproducibility in the outputs of large language models: evidence across diverse finance and accounting tasks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2503.16974

[4] Meyman, E. (2025). Deterministic Governance as Epistemic Commitment: Why reproducibility Matters for AI accountability. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5659170

[5] Onyenahazi, O. B. (2025). Integrating artificial intelligence in financial auditing to enhance accuracy, efficiency, and regulatory compliance outcomes. International Journal of Research Publication and Reviews, 6(7), 23–44. https://doi.org/10.55248/gengpi.6.0725.2402

[6] Carter, R. E., Attia, Z. I., Lopez-Jimenez, F., & Friedman, P. A. (2019). Pragmatic considerations for fostering reproducible research in artificial intelligence. Npj Digital Medicine, 2(1), 42. https://doi.org/10.1038/s41746-019-0120-2

[7] Choowan, P., & Daovisan, H. (2025). Artificial Intelligence in Data Governance for Financial Decision-Making: A Systematic Review. Big Data and Cognitive Computing, 10(1), 8. https://doi.org/10.3390/bdcc10010008

[8] Li, D., Yu, G., Wang, X., & Liang, B. (2025). AuditableLLM: a Hash-Chain-Backed, Compliance-Aware auditable framework for large language models. Electronics, 15(1), 56. https://doi.org/10.3390/electronics15010056

[9] Han, H. (2025). Challenges of reproducible AI in biomedical data science. BMC Medical Genomics, 18(S1), 8. https://doi.org/10.1186/s12920-024-02072-6

[10] Rahrooh, A., Garlid, A. O., Bartlett, K., Coons, W., Petousis, P., Hsu, W., & Bui, A. A. (2023). Towards a framework for interoperability and reproducibility of predictive models. Journal of Biomedical Informatics, 149, 104551. https://doi.org/10.1016/j.jbi.2023.104551