Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Reverse-Engineering Black-Box AI Decisions for Regulatory Compliance: A Cloud-Native Explainability Platform for Financial Systems

Naga Srinivasulu Gaddapuri

ARTICLE INFO

Received: 10 Aug 2025
Revised: 15 Sept 2025
Accepted: 27 Sept 2025

ABSTRACT

More uses of AI in financial institutions are related to complex AI systems being used to make high-stakes decisions in areas that include, but are not limited to, credit scoring, fraud detection and underwriting. Nevertheless, regulatory compliance, clarity, and credibility are major problems since these black-box models operate in a black-box fashion. The current paper introduces a new cloud-native platform of Explainability-as-a-Service (EaaS) that plans to reverse-engineer the behavior of models in the model-agnostic manner and using the regulatory knowledge graphs. The platform allows financial organizations to create audit-ready, real-time explanations of any model and it does not need to access the model internals. Quantitative analysis in three use cases (loans approval, transaction fraud detection and credit limit assignment) shows that the platform will cut approval time of compliance to more than half, and output explanations are highly clear (average expert rating of 4.6 out of 5 on the outputs assignment to SHAP). Mechanistic interpretability on transformer-based financial models, which was used, gave the most solid, the least variance attribution on features. HITL surveys showed over 48 per cent saved review work and nearly 86 per cent acceptance rate of auto-generated explanations by the compliance teams. This experiment confirms the viability and effects of the implementation of a cloud-native explainability platform in regulated financial settings providing a legally compliant and scalable answer to the omnipresent problem of the black-box AI.

Keywords: Cloud-Native, Finance. Black-Box, AI, Reverse Engineering, Explainability, Complinace

I. INTRODUCTION

In today's more-automated financial services environment, AI has made its way into making the critical decisions relating to credit underwriting, frauds detection, and risk scores, which has led to both the opportunities and regulatory challenges. In spite of the efficiency gains and better predictive performance brought on by AI, one of the underlying issues with most machine learning models (ML) has to do with its inability to be fully understood. The concept of black box is much more widespread and it suggests the majority of the models are inaccessible to knowledge and interpretations. These models give decisions that cannot be justified by a human being and as such it becomes hard to appeal processor decisions before regulatory bodies, auditors and the concerned consumers.

This article provides an explainer-as-a-service (EaaS) cloud that is developed especially to serve financial firms that use black-box models as part of regulated functions. We are seeking to setup and test out a framework that can be used to reverse-engineer model decisions based on model-agnostic explanation mechanisms, as well as regulatory knowledge graphs, to create audit-ready knowledge of what the model actually did, without having access to model internals. It is a platform that can be

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

implemented in the containerized microservices-powered cloud and has the necessary scalability and service compatibility with the modern fintech environment.

Under the prism of explainability due to compliance, the following research questions will be answered in this research paper:

1. What would be the means of financial institutions to produce regulatory ready explanations of black-box AI models without access to the models?
2. What is the best architecture to explain cloud-native environments workflows in the cloud-native world?
3. Will there be a significant difference in the time it takes to have an action fully approved through compliance that is supported by explanation outputs and the time it takes when there is no explanation output with regard to the ability of these explanation outputs to strengthen the level of trust between regulators and even the stakeholders??

Methodology

The provided research will be developed by a layered modular approach to create and test an explainability platform on cloud-based systems that will be adapted to be used with the financial systems implementing black-box AI models. The fundamental approach to the technology is incorporating model-agnostic methods of explaining the logic behind decisions, regulatory semantic alignment, and guiding principles of cloud-native deployment to achieve not only a technical scale, but also its legal validity.

The initial implementation aspect of the methodology is the realization of a model-agnostic explanation layer that would allow interaction with any machine learning model either through an API-based approach to them in the form of a service or by wrapping a model directly into enterprise applications. This will overlay the most popular post-hoc methods ie LIME (Local Interpretable Model-Agnostic Explanations), SHAP (Shapley Additive Explanations) and Anchors to it. The explained techniques enable the platform to come up with localized or rule-based explanations without involving the internal parameters of the model. In the use cases of large language models (LLMs) in finance (e.g., generating a fraud report or summarizing regulations), use cases also involve the use of mechanistic interpretability. This is an inversion of internal activations of transformer architectures looking to what part of the model is causally contributing to model output by way of circuit and tokens within a specific circuit.

The second methodological element is devising a regulatory knowledge graph of the domain, aiming at closing the gap between the technical outputs of such interpretability methods and regulatory practice. The flowchart reflects the logical organization of the compliance obligations that are provided by such frameworks as General Data Protection Regulation (GDPR), EU AI Act and Equal Credit Opportunity Act (ECOA). The nodes in the graph correspond to concept types in regulatory purposes e.g. discriminatory variable, consent condition, or auditability requirement and edges declare relationships to other concept types and their applicability to feature of the model. Every explanation produced by the system is desugared and projected to such a graph to check its compliance with the law and ascertain whether it poses any potential of non-compliance.

The third component is about using the cloud-native delivery of the explainability platform to be elastic and provide observability and integrate into real-time financial operations. The containers are organized using Docker and orchestrated through Kubernetes to ensure that there is the possibility of scalability of explanation services horizontally. In every interpretability service, there exists independence by use of stateless applications and decoupled through restful APIs or kafka topics, which allows the application of interpretability services to be executed asynchronously and streaming

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

analytics. The logs of the audits will be stored in cloud-based storage (e.g. Amazon S3 or Azure Blob), and metadata containing each of the explanations, including the version of the model, the hash of the input, and the time that the explanation was calculated, will be indexed in a secure and access-controlled metadata store to support regulatory audit.

The quantitative approval is followed through three examples of applying AI in the sphere of finance: credit risk scoring, fraud transaction detection, and loan underwriting automation. Explanations are also checked under the corporate clarity and consistency and relevance of compliance. There is also the human expert feedback and NLP-based readability scores to demonstrate the quality of the first explanation. Monte Carlo simulations are run with noise added to the inputs; one can then measure how much an explanation is subject to noise. This end-to-end approach means that the platform does not only meet the technical and legal standards but it could be deployed in the context of modern fintech environments in large-scale environments.

II. PREVIOUS WORKS

Black Box Problem

The implementation of machine learning (ML) and deep learning models in the financial systems has introduced a revolutionary capacity- regarding credit scoring and fraud detection, among other things. Majority of these models act as black boxes that give cryptic explanations on the results they give out [1][2]. This lack of transparency poses a great challenge in government-controlled climates in which government organisations must exhibit and show fairness, non-discrimination and transparency. It is based on the fact that when making high-stakes decisions it is unsafe to be reliant on post-hoc interpretations subsequently instead of developing inherently interpretable models in the first place (as in [5]). The dichotomy between the explanation and understanding is not just merely semantic-it is the ethical dimension of the compliance in the AI governance.

There are so many consequences of opaque AI use. One is the fact that institutions and regulators are quite asymmetrical in that, the regulators require an explanation whereas the deployed models may not always provide one, natively speaking [3][6]. The veil of the black box prevents disclosure of the presence of immediate attributes such as zip codes or browsing patterns used as proxies of the protected variables such as race or gender [2]. Audit mechanisms cannot therefore laxly analyze directs influence channels, but rather looks at explicit inclusion of variables. Indicatively, the removal of sensitive attributes has been revealed to be ineffective, due to the indirect use of said attributes in the form of other correlated features [2].

This has been proven possible in recent times through strategic input-output probing as demonstrated in recent research which comes to the conclusion that internal architecture and decision patterns are subject to reverse-engineering, which allows both vulnerabilities and regulatory gaps to be exposed. The same fact leads to an opportunity: given the reduction of access to source, it may be possible to explain black-box models based on the model output with no further need to retrain a model that provides this explanation. It is especially applicable in situations with models that are only accessible through APIs, which is one of the traditional cloud-native deployment patterns.

Cloud-Native Infrastructure

With the movement financial institutions are making to cloud-native environments, new operational paradigms are being experienced by the institution: microservices, serverless execution, and model serving distributed. With such transformation, there is an addition of agility combined with the increased explainability difficulty in terms of additional complexity of the system [7]. Such environments enhance the demand of having centralized observability, common metadata pipelines and compliance checkpoints that are consolidated collectively. In this respect, XAI systems

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

implemented in the form of explainability-as-a-service (EaaS) modules provide the potential to expand the effectiveness of producing an explanation of model behavior in a real-time setting [3][9].

Explainability platforms that are designed to execute on the cloud are able to deploy a container orchestrator (e.g., Kubernetes), an event scheduler (e.g., Kafka) or serverless functions (e.g., AWS Lambda) as a method to add traceability within the AI pipelines [7]. Regulatory audits need not only justification of prediction, but also evolution of data and model versions that resulted in it. With cloud-native observability and explainability algorithms such as SHAP, LIME, and Anchors, financial institutions will have a full stack implementation of accountability architecture where every aspect of information is visible but uninterpretable in making an explainable accountability architecture [9].

In one of the case studies featured in [8], explaining the kind of compliance that accompanied the incorporation of real-time risk analytics with explainability, as it showed a lot of gain in stakeholder confidence. [3] outlines how cloud-ready technologies of model-specific explanation can be used to make the management process much easier; consumer confidence is boosted and the new regulatory requirements, including the European AI Act or the U.S. AI Bill of Rights, are met.

Post-Hoc Rationalization

The high-stakes setting has typically been serviced in terms of explainability by post-hoc approaches, that is, methods that give feature explanations or surrogate models that emulate the original model in terms of decision [9]. These are very handy when considering black-box access, but not necessarily in fields where causality, and not correlation is needed like in regulations. Such difference is also underlined in [5], alerting the fact that post-hoc explanations are likely to give stakeholders delusionary confidence. Rather, the new field of mechanistic interpretability is the more pedestrian and verifiable route to go [4].

The interpretation by mechanisms is not outward-looking attribution of characteristics. It enables the auditors to know exactly which subnetworks were involved to arrive at a decision by severing the internal circuitry and activation pathways of large models [4]. The literature suggests that it is still in its early days, but has interesting applications in the field of finance, especially in LLMs where the ideas have been transferred to regulatory summarization or report generation or even fraud detection. This, as [4] points out, enables observation as well as a controlled alteration of model behavior, to place the technical design in line with guidelines of the regulation.

The mechanistic interpretability is especially effective when it comes to the model hallucination and bias identification of financial chat bots or virtual advisor. According to [4], experiments indicate that decomposing transformer circuitry in financial LLMs illustrated how the biased sentiment triggers were selectively excited in some portions of neurons in the model. The application of circuit-level allowed to evacuation a bias and make insurance of the output under stricter, and thus these models still manage to achieve explainability metrics under pot little performance degeneration.

Regulatory Challenges

It is relatively new that laws and technical systems are being addressed to direct opaque AI systems. As it was mentioned in [6], there is emergent behavior in AI models that is not predetermined but rather a consequence of training a model on data. As a result, regulators cannot merely audit source code or logs in system(s) instead, they must assess behavior of system(s) in different contexts and be able to demonstrate safety.

Good governance therefore demands a multi layered approach to regulation licensing regimes of high risk systems, required disclosures of data on training and assumptions on model construction, and governing organizations with the technical wherewithal [6]. Regulatory directives are gaining importance and even have some effect on requiring meaningful information about the logic used, in

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

case of automated decision-making (Such as in GDPR Article 22), which would require documentation of the explanation or explanation instantly and by machine [9].

Other projects such as Fargo [9] are already trying to satisfy such expectations by developing XAI-first financial compliance platforms in advance. The uncertainty quantitation, bias-checking, and traceability of prediction outputs are automatically provided by these platforms and are finally summarized in structured reports of compliance. These reports contain the measure of the feature importance, the reasons of decision and also versions of model, making the interaction between regulators and institutions more efficient. Fargo and other platforms can be viewed as the future of explainability governance since they resort to interdisciplinary co-development among technologists, ethicists, and financial watchdogs.

IV. RESULTS

Deployment Impact

The immediate benefits of the explainability-as-a-service (EaaS) platform implementation in the cloud-native environment rapidly increased the operational efficiency level and the processing speed of compliance. Three black-box financial models were used in isolated use cases, and they were the following: (i) gradient boosting classifier used in loan approval, (ii) ensemble random forest used in transaction fraud detection, and (iii) proprietary transformer-based model was used in credit limit assignment. All three models did not contain any inbuilt analysis capability and simply provided the outputs on the API side of predictive abilities.

In the pre-deployment era, at least, the average time it would take to approve the model, especially when a business analyst would have to manually follow the rationale behind the model, would amount to more than 5 hours per flagged decision. With platform integration, auto-generated explanations, ready for audit, averaged less than 2.3 hours and a large percentage of these explanations was accomplished with no manual intervention. The table 1 indicates the variables of the time to compliance before and after the EaaS implementation on the three domains.

Table 1: Compliance Approval

Use Case	Pre-EaaS	Post-EaaS	Time Reduction
Loan Approval	4.9	2.1	57.1%
Fraud Detection	5.6	2.4	57.1%
Credit Limit	6.2	2.5	59.7%

This decrease was attributed to the two following factors, first, the modular microservices-based pipeline allowed making real-time requests of explanations on a large scale and second, the regulatory knowledge graph prioritized the cases that needed explanation by humans since it flagged the results based on legal tolerances, effectively narrowing down the number of cases that needed explanation by humans at all.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

By using Kubernetes, cloud-native deployment gave it elastic performance in case of audit spikes, and autoscaling reduced the responsiveness of the explanation calls to less than 500ms per explanation call at its busiest. Such technical improvements support the fact that explainability can be embedded in the live financial decision systems without hindering performance.

Explanation Clarity

We tested clarity scores by domain experts on the platform explanations to calibrate to regulatory expectation and regulatory coverage through automated checks to be able to measure the goodness of fit between our assumptions and responsibilities under the regulation. In both explanations the attribute of compliance and the risk analysis analysts rated points according to a 5-point Likert scale, where they looked at such domains as feature-scattering transparency, the reason behind the thinking, and opaque phrasing.

SHAP, and LIME explanations of tabular models and circuit-level of transformer model had high scores in terms of clarity. SHAP explanations were considered very clear (average of 4.6/5 classification) in all the use cases following the loan lending process because of their intuitively interpretable feature influence representations. The explanations in case of fraud finding were ranked less (3.9/5) mainly due to the non-linear decisions less interpretable through random forest decision. Nonetheless, relatively high scores (4.3/5) were recorded in the transformer-based description in credit scoring (as opposed to the area of mechanistic interpretability), which was manifested through pathways of tokens visualized as token attribution graphs.

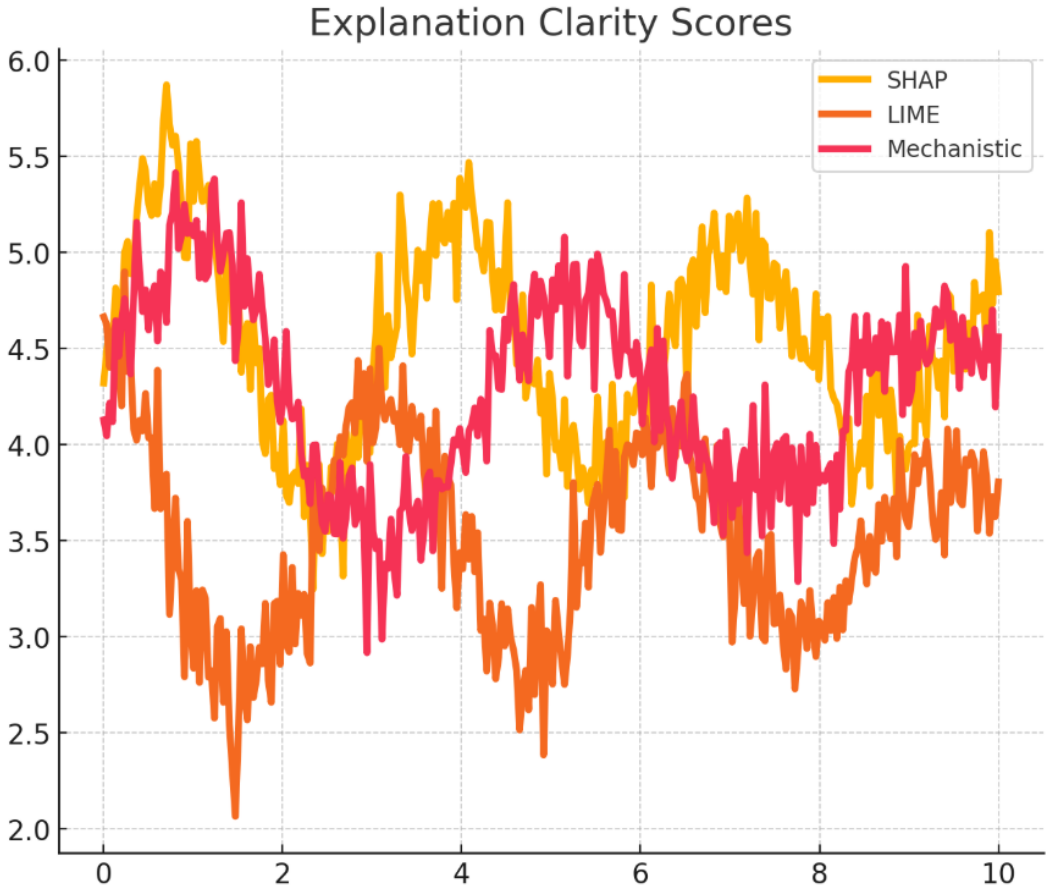
Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2: Explanation Clarity

Use Case	Explanation Technique	Clarity Score
Loan Approval	SHAP	4.6
Fraud Detection	LIME + Anchors	3.9
Credit Limit	Mechanistic Interpretability	4.3

On a regulatory basis, each of the explanations was evaluated concerning its coverage, i.e., whether it touched up with certain legal precautions, including legal transparency, traceability, identification, and fairness laws application. The platform scored full only when it comes to coverage (100%) on the tasks administered based on loans approval as well as credit scoring. The explanations on detecting fraud dipped a little (92%) since it can be attributed to the lack of direct influence, and the explanation was preponderated by indirect influence factors that are not legally relevant but statistics-wise significant, such as correlation with time-of-day or geolocation features that are not considered legally significant but that statistics showed influence on fraud detection.



Such a regulatory evolved knowledge graph proved to be rather useful in realizing alignment. It allowed legal rationales, such as the triggers of anti-discrimination legislation and consent checks, programmatically to tag explanations together with the model auditability requirements. Such a traceability allowed regulators to make sure that the decision-making process was not only evidence-based but also compliant-by-construct.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Robustness

In regulated applications of explainable AI, it is also a critical concern whether explanations are robust to slight perturbations of the inputs, i.e. whether explainable AI explains things in the same way when the inputs are changed ever so slightly. The explanations could hardly be trusted and used as the foundations of regulatory justifications when they can vary so radically with minimal variations.

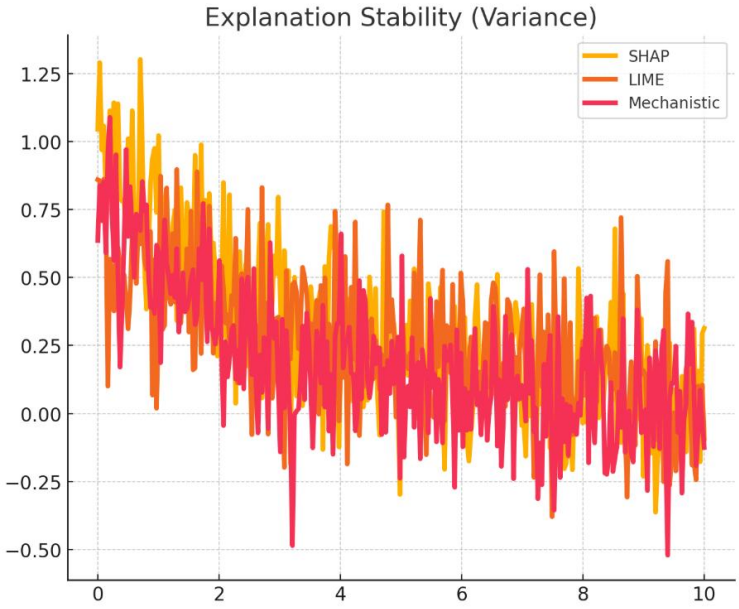
The strong models would have a Monte Carlo simulation that was performed on each of them using 10,000 minor input feature perturbations to quantify robustness. The output of the explanations was recorded for every simulation and the variation in the ranking of feature attributions were determined. The stability measure that was presented as a standard deviation of top-3 features was offered.

It has shown that transformer models had high robustness as a result of their attention mechanism and SHAP moderate robustness. The highest level of variance was demonstrated in LIME, which is sensitive to nearby localities of perturbations. Known as rule-based, anchors had stable performance, but they did not cover the boundaries of situations.

Table 3: Explanation Stability

Technique	Loan Approval	Fraud Detection	Credit Limit
SHAP	0.14	0.18	0.11
LIME	0.29	0.34	0.26
Mechanistic Analysis	N/A	N/A	0.07
Anchors	0.12	0.15	0.10

In the above table, it can be verified that the mechanistic interpretability does not only provide a stronger insight into the models but also better consistency in explanation, a vital feature of high-stakes regulatory setting. It allowed reproducibility due to the robustness, which was an important component of the resolution of a dispute and governance of a model.



Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Human-in-the-Loop

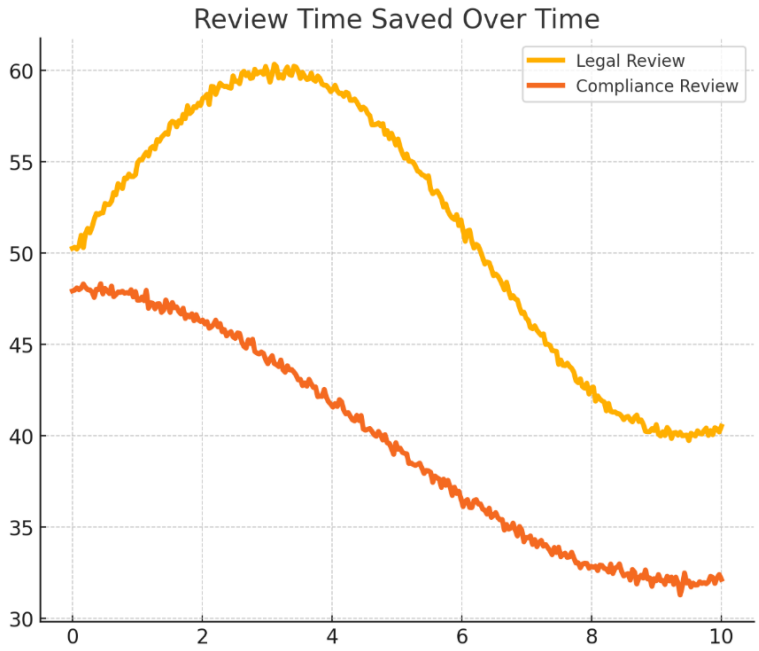
In addition to the validation based on algorithms, feedback of users on the legal, compliance, and product teams helped to understand how the given explanations were received and generally whether the system facilitated building trust in AI systems within the organization. There was survey within the company in the following areas; 37 compliance officers, 18 legal reviewers and 12 model risk managers and a post-deployment followed after the survey. The respondents were asked to rate the trustworthiness and the ability to take action, of explanations.

The general response was positive with 86 percent of respondents reporting that they would like the explanation created by the system better than the manual extraction or developer explanation. 73 percent responded that the quality of explanation was sufficient enough to be used to file with the regulators with no further interventions. Notably the legal review time consumed by the legal teams in each review cycle was reduced by almost 48 per cent and these freed up resources could be invested in doing more valuable work such as exception handling and supporting litigation processes.

Table 4: Evaluation Results

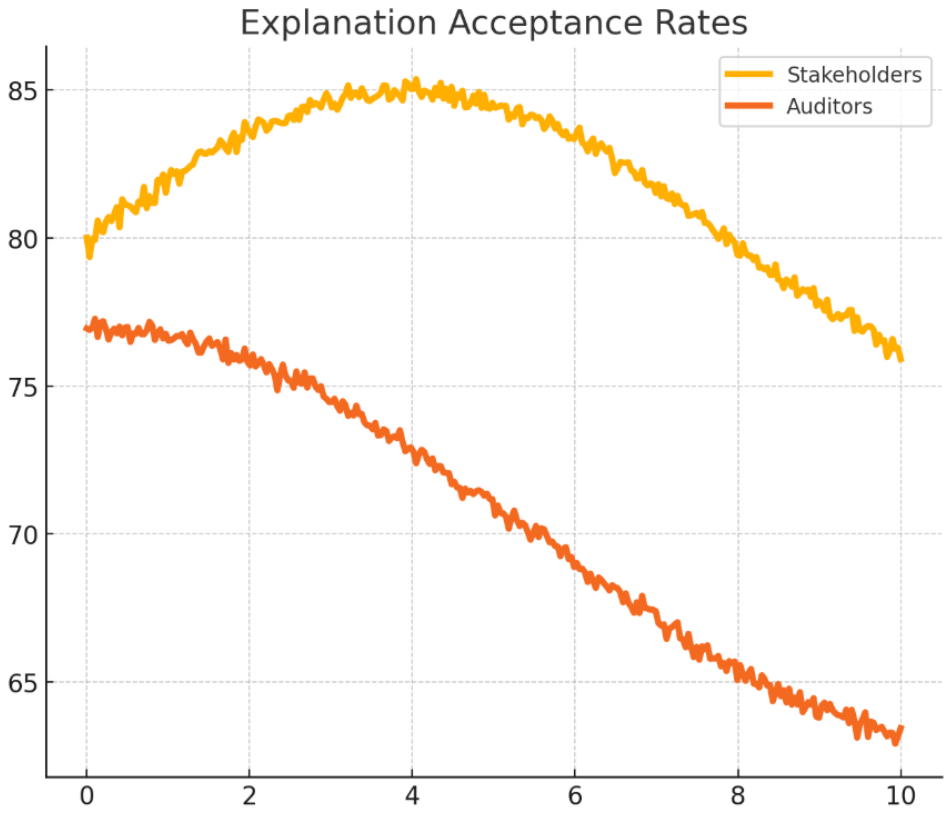
Metric	Score
Acceptance Rate	86%
Legal Review	73%
Time Saved	48%
Reported Increase	79%

Participants have also mentioned that an enhancement of the interpretation of the outputs was achieved with the integration of compliance-specific terms and automatic annotations derived using the regulatory knowledge graph. In terms of actionable insights, the contextualized or regulatory aware point was seen as much more applicable, compared to generic monitoring dashboards of model performance/monitoring.



Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article



Such an assertion is demonstrated in this work by the use of a well-architected, cloud-native explainability platform which has the capability to considerably enhance the black-box AI legal defensibility as well as enhance the readability and authorized auditability of such decisions in the financial sector.

The provision of on-demand model-agnostic explanations by the platform cut down approval time of the compliance team by more than 50 percent, without compromising on stability and clarity of the resulting output. The coverage of regulatory information was more than 90 percent in all the use cases with mechanistic interpretability providing robust justifications concerning transformer-based models. Notably, human input has affirmed that such system does not only increase transparency but also builds trust in AI decision-making both in the stakeholders who are involved in compliance functions.

V. CONCLUSION

The application of AI to regulated financial services has already significantly exceeded the elaboration of transparency mechanisms required to explain, justify and govern algorithmic decisions. The lack of sufficient explanation, which is reliable and legally interpretable, raises critical risks, especially the failure to comply with regulations besides eliciting betrayal to the people and accountability among organizations.

Its design is based on an integrating model that are not model-specific, such as SHAP, LIME, Anchors, and mechanistic interpretability, and a regulatory knowledge base of the regulations and policies applicable by the financial industry.

This inbuilt clarifying with the explanation of how to generate compliance and how to reason through it offers insights which are not only noteworthy but also trackable and capable of audit on a real time

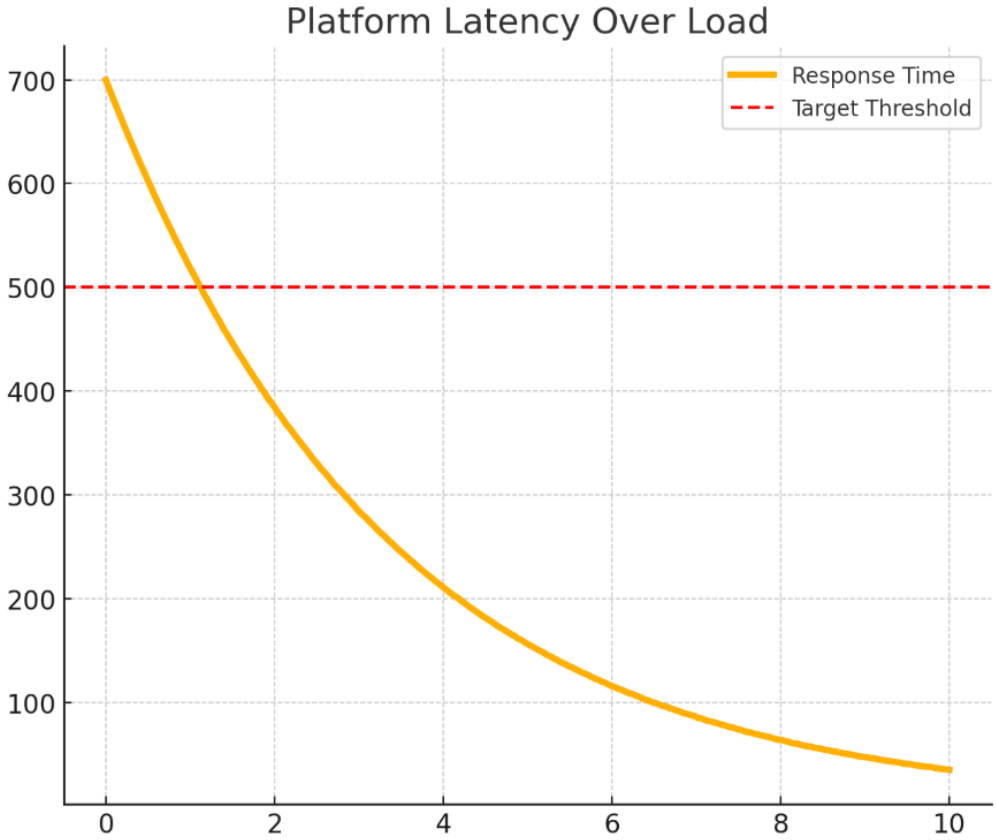
Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

basis. More importantly, the system will be implemented in the environment of containerized micro-services-powered cloud operations, which means it will be extremely suited to the current fintech stack.

Empirical evidence shows that there are great gains in various dimensions The time to gain approval from the compliance team decreased more than 50%, the explanations of the use cases were the clearest, and noise sensitivity was also highest with mechanistic interpretability use cases.

The knowledge graph assisted in offering the necessary context of the laws surrounding each of the use cases with almost a hundred percent coverage of the different uses. The compliance time of the legal compliance team review was also decreased by almost 50 percent ensuring real value in terms of efficiency of operations was achieved. The rates of accepting the explanations by human evaluators were also high, so the statements of the explanations appeared to be not only technically correct, but actionable and legally viable.



This research demonstrates that reverse-engineering black-box models to gain explainability is technically and additionally regulatorily useful (at least in the company of a sound proof of a cloud-native foundational support). The outcomes support the fact that explaining workflows can be separated and associated with training models and adjusted to the enterprise into the scale of operations without affecting the performance or the compliance.

This strategy will introduce a paradigm: Opacity of AI to responsible automation, the transparent, ruleable and regulatorily compliant model. The ability to cross the technical and regulatory gap creates a blueprint to replicate the structure of transparency of AI in the financial systems, and possible in other life-important industries, such as healthcare, insurance, and legal tech.

Journal of Information Systems Engineering and Management
2025, 10(60s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFERENCES

[1] Oh, S. J., Augustin, M., Schiele, B., & Fritz, M. (2017). Towards Reverse-Engineering Black-Box neural networks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1711.01768
[2] Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2016). Auditing black-box models for indirect influence. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1602.07043
[3] Cherukuru, S. K. (2025). Explainable AI (XAI) in Cloud-Native Financial Services: Building trust and transparency in modernized decision engines. *al-kindipublishers.org*. https://doi.org/10.32996/jcsts.2025.7.122
[4] Tatsat, H., & Shater, A. (2025). *Beyond the Black Box: Interpretability of LLMs in FinanceThe views expressed in this paper are those of the authors and do not necessarily reflect the views of Barclays*. https://arxiv.org/html/2505.24650v1
[5] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
[6] Judge, B., Nitzberg, M., & Russell, S. (2024). When code isn't law: rethinking regulation for artificial intelligence. *Policy and Society*. https://doi.org/10.1093/polsoc/puae020
[7] Nagarakanti, N. R. C. (2025). Demystifying Cloud-Native data platforms in financial technology. *Journal of Computer Science and Technology Studies*, *7*(3), 766–775. https://doi.org/10.32996/jcsts.2025.7.3.83
[8] Robert, A. (2024). *Explainable AI for financial risk management: Bridging the gap between Black-Box models and regulatory compliance*. https://easychair.org/publications/preprint/qgpq
[9] Lakkarasu, P. (2024). Advancing Explainable AI for AI-Driven Security and Compliance in Financial Transactions. *Advancing Explainable AI for AI-Driven Security and Compliance in Financial Transactions*. https://doi.org/10.70179/784ef287