© AUG 2026 | IRE Journals | Volume 10 Issue 2 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I2-1722587

GRACE: A Generative AI Framework for Risk and Compliance — A Proposed Architecture and Illustrative Evaluation Methodology for Trustworthy GenAI in Banking

ANAMIKA
Shoolini University

Abstract*—While Generative AI (GenAI) is set to make a mark in financial services, the use of this technology in banking risk and compliance throws up the critical questions of interpretability, auditability and trustworthiness in the highly regulated sector. Large language models (LLMs) such as GPT-4 and Gemini Pro are not specialized for banking and do not comply with regulatory rules and standards. It proposes a novel GenAI based banking risk and compliance framework, namely a purpose-built GRACE (Generative Risk and Compliance Evaluation Framework), which integrates Explainable AI (XAI), a cryptographically secured Immutable Audit Trail, a Human-in-the-Loop (HITL) oversight layer and dedicated compliance alignment layers for Basel III, IFRS 9, AML/CFT and GDPR. Beyond the architecture, we suggest a method for evaluating GRACE and representative comparator systems (GPT-4, Gemini Pro, and BloombergGPT) by six criteria: interpretability, compliance readiness, trustworthiness, regulatory auditability, bias and fairness, and domain specificity. A proposed evaluation protocol is presented to assess the adaptability of the systems through an illustrative architectural capability assessment. The present assessment is theoretical rather than empirical and is intended to demonstrate the potential value and discriminative capability of the proposed methodology. Because no expert panel evaluation or empirical dataset has yet been established, the illustrative values should not be interpreted as measured performance scores. Whereas purpose-built systems are more architecturally flexible, domain tuned GenAI systems are less flexible. We propose a prototype of how this work could be implemented in the real world. This encompasses a Banking Compliance Evaluation Suite, scoring protocol devised by a panel of experts, and a statistical testing framework.

***Keywords* — Banking Compliance, Basel III, Explainability, Generative AI, Human-in-the-Loop, IFRS 9, Large Language Models, RegTech, Risk Management, Trustworthy AI**

I. INTRODUCTION

The technology and compliance challenges in banks are a peculiar mix given the cases, the tech-related challenges come increasing use of technology to combat financial crime is not what they are behind on. In other instances, it is the difficulty of complying with regulations alongside handling credit risk, market risk, an expanding operational risk landscape, AML requirements and the relentless push for doing more with less with the help of tech solutions. Several emerging technologies could transform document review, document modelling, and regulatory report drafting, and Generative AI is among the most significant. [1], [2].

The journey from having general AI capability to being able to safely and legitimately use that capability for purposes in the financial industry is long and complex. Commercial AI models with strong general language processing abilities such as GPT-4 and Gemini Pro excel at a wide variety of tasks across numerous sectors, but they lack ability to perform the specific and intricate tasks necessary in the financial highly regulated sector. These models also lack many of the assurance constructs as required by banking and regulatory compliance frameworks such as Basel III and the International Financial Reporting Standards (IFRS), such as explainability and threshold setting. [3], [4].

This inability to do so has severe effects. In one instance, if a compliance officer is given an AI created risk assessment, they will not be able to fulfil the EBA and the U.S. OCC's requirements to explain the risk assessment. He will also not be subject to the same obligations if he completes a poorly formulated AML assessment, and receives a lot of false positive answers, or if he fills out the assessment in random order. [5].

While GenAI applications in finance could be growing in abundance, there are no frameworks to evaluate GenAI systems against compliance and risk requirements within the banking industry. The current literature includes a number of existing works which use traditional NLP evaluation metrics (e.g., MMLU, HellaSwag, TruthfulQA), which is not reflective of compliance-related tasks [6]. This leaves a definitional gap, one that has no widely accepted, regulation-based evaluation methodology for GenAI systems in banking, and therefore no systematic comparison of the performance of existing systems, against such criteria.

The present paper fills this gap in two parts. We're proposing GRACE (Generative Risk and Compliance Evaluation Framework), a purpose-built GenAI architecture that is built around the banking compliance and banking risk requirements as first order design principles. Second, we suggest a six-dimensional framework for evaluation based on regulatory guidance and then illustrate such an evaluation by comparing GRACE with representative general-purpose and domain-tuned systems, within it. The illustrative scores shown in this paper are not from an actual empirical study, but were calculated using the architectural capability mapping approach to demonstrate the discriminative power of the approach and to inspire an empirical study to validate the approach as future work..

The research question guiding this work is:
RQ: In the banking risk and compliance domain, how could purpose-built architecture (GRACE) be compared to general-purpose and domain-tuned GenAI systems (GPT-4, Gemini Pro, and Bloomberg GPT) regarding interpretability, compliance readiness, and trustworthiness, and how might these systems be evaluated robustly?
The rest of this paper is structured as follows. Relevant literature is summarized in Section II. The proposed GRACE architecture is detailed in Section III. Section IV explains the evaluation methodology. An example use of that methodology is shown in Section V. Implementation considerations, limitations, and findings are discussed in Section VI. Section VII concludes and anticipates the planned future work for empirical validation.

II. RELATED WORK

A. GenAI in Financial Services

The success of transformer models that began with GPT-3, meant that more and more Large Language Models (LLMs) became integrated into almost every part of financial services. An example of this is Bloomberg GPT [7], one of the first models of its kind to have been developed and trained, and it was made specialized for the needs of users in a financial domain. Bloomberg GPT was one of the earliest models and is one of the very few models developed for financial text and showed evident improvements over general models. A 50 billion parameter model trained on 363 billion financial tokens with improvements across financial text classification and named entity recognition as well as headline analysis. It is also noteworthy that compliance was not a theme that was addressed during the development of Bloomberg GPT, nor was compliance evaluated for Bloomberg GPT's ability to conduct risk assessment as part of a compliance framework.

In the field of financial compliance, SShah et al. [8] explored the use of NLP; while the interpretation of regulatory language is a relatively mature discipline, the use of generative models with the aim of implementing ongoing compliance is still limited and there are no known benchmarks to measure them. Cao [9] studied the use of AI in the big financial companies and found that many companies have running pilot projects in the area of risk, but a relatively small number of companies have created models that satisfy the regulatory requirements for interpretability.

B. Explainable AI and Trustworthy Systems

Explainability and interpretability are crucial aspects of AI regulation. Importantly, Doshi-Velez and Kim [10] proposed a taxonomy of explanation frameworks that combines completeness, soundness and parsimony. They claim that there are certain fields where they model deliberately with opacity to enable post-hoc explanation – such as in finance and in health sciences – where such explanations are not possible with opaque models.

Arrieta et al. [11] conducted a systematic literature review of explainable AI and identified transformers with attention as a special category of problems for interpretability, as these pose a specific difficulty due to the fact that the attention patterns of the model are

not aligned with the type of explanations that financial regulators demand, "feature importance". In the banking context, the European Central Bank (ECB) Targeted Review of Internal Models (TRIM) [12] calls for the outputs of the AI systems to be explainable, traceable and reproducible.

However, this goes against the probabilistic paradigm of LLM models. The EU AI Act [13] provides a useful framework to evaluate the use of GenAI in the context of delivering regulated financial services, notably through its 'trustworthy AI' principles, such as transparency and accountability.

C. Compliance and RegTech Systems

RegTech has grown beyond being a particular knowledge space to one for regulatory compliance automation. Arner et al. [14] describe it as moving from the reactive compliance management, to proactive risk monitoring, to predictive regulatory intelligence. This is the potential of the fourth phase: Generative reasoning about regulation, however not that well understood is how much this can be done in conjunction with existing compliance infrastructure.

Focardi and Fabozzi [15] have investigated applications of AI in The Netherlands. Basel III and IFRS 9, showcasing how machine learning can help with. expected credit loss measurements. New models for measuring expected credit losses, and enhanced estimation of expected credit losses. Outline problems and opportunities for governance in the present risk management context (not aimed at frameworks). Khatchatourian [16] highlighted hallucination among LLM as a serious concern. in AML Screening, a model with high confidence for making an decision of 'screening positive'. erroneous transaction risk rating can bring about both direct and indirect benefits. Violation of regulations, losses in revenue to downstream.

D. Gaps in Existing Literature

There are three gaps that inspire this work. Unlike other frameworks that simply tack on explainability and compliance readiness as afterthoughts, the first one has a strong focus on these becoming built-in features of banking GenAI, not add-ons. Second, current LLM evaluation is heavily based on general, non-regulatory measures, with a lack of measures of compliance-specific performance. Third, the literature on banking GenAI focuses mainly on the human-in-the-loop (HILo) design, which is central to regulatory compliance in high-stakes use-cases. This paper suggests an architecture and an evaluation process that aim to fill all three gaps and suggests that a methodology should be validated experimentally.

III. THE PROPOSED GRACE FRAMEWORK

A. Architectural Overview

The "first principles" modular approach of GRACE makes it a GenAI architecture of choice for banking risk and compliance. GRACE proposes regulatory alignment, interpretability, and human oversight as embedded and constituent elements of the LLM's architecture, rather than added on at the end. Rather than consider these attributes as "feature" additions achieved through prompt engineering or retrieval-augmented generation (RAG), the proposed architecture recognizes them as intrinsic. Composed of five interdependent layers, the proposed architecture entails a (1) the domain-adaptive language model core; (2) a Regulatory Alignment Module (RAM); (3) an Explainability Engine; (4) a Human-in-the-Loop Interface (HITLI); and (5) an Immutable Audit and Governance Layer.

B. Domain-Adaptive Language Model Core

The GRACE core aims to be a large language model (LLM) in the 13 billion parameter range. It is likely that in its first instance the GRACE core will be adapted through three main stages of intensive fine-tuning. The first of such stages involves an extension of the LLM's pre-training to cover additional corpus of regulatory texts. The corpus of texts is planned to include the Basel III and IV capital framework and binding capital regulations, IFRS 9 and 17, FATF Recommendations, EBA Guidelines, and FCA Consultation Papers, among others. The second stage involves the creation of numerous regulatory compliance focused, and factually accurate, prompt and response pairs and is planned to be done in close collaboration with regulatory compliance specialists and risk officers, respectively. The last and final stage involves the application of Reinforcement Learning from Human Feedback (RLHF). In this stage, a reward model will be developed to encourage the desired improvements in this regulatory adaptive LLM. The corpus size and the volumes of data planned for this fine-tuning process are indicative of targeted volumes and are not precise or exact. For this reason, they are included as part of the empirical validation process in Section VI.

C. Regulatory Alignment Module (RAM)

The RAM is described as a rule-layer design that intercepts all model outputs before they reach the user, intended to be deterministic. It possesses three main features. Mainly, the model's generated regulatory citations would be verified against a regulatory knowledge graph with coverage from various jurisdictions. Furthermore, a compliance-tagging system would be utilized to apply compliance tags to each output segment, which would enable compliance officers to trace each segment to its regulatory citations. Lastly, a Ssatisfaction mechanism would determine outputs that deviate from regulatory constraints, for instance, a capital proposal that fails to meet the Basel III Tier 1 requirement.

D. Explainability Engine

An executive summary, confidence intervals, and uncertainty quantification will be provided as part of a multi-tiered explanation for each output. The Explainability Engine will delineate key input features and the regulatory provisions associated with each output. As part of the design, counterfactual analysis, integrated gradients for token-level attribution, and SHAP-based feature importance for structured data will be utilized. This design also aims to satisfy the "effective challenge" requirement of SR 11-7 model risk guidance.

E. Human-in-the-Loop Interface

GRACE is an assistive decision tool, not a full decision-making system. In the proposed HITLI system, a human reviewer would screen all significant outputs, which include but are not limited to risk scores, SAR recommendation, capital sufficiency, breach alerts, etc. A compliance officer would examine the model's decision and the supporting rationale along with the model confidence and the relevant regulation before approving, amending, or rejecting the model, and would do so with full traceability. This is aligned with the ECB's principle that human judgment should be Central to every material risk decision, traceability must be applied. This design reflects the ECB's framework where human judgment significantly influences material risk decisions, regardless of the extent to which these decisions are model-assisted.

F. Immutable Audit and Governance Layer

GRACE logs all forms of interactions, including inputs, outputs, and any intermediate system reasoning, as well as any system changes and the reviewer decisions, in a tamper-evident system using cryptographic hashing with logging of the user's personal data, the date, the version of the model, as well as input/output pairs, explanation package, and regulatory tags. This implementation is designed to meet the requirements of data retention policies for high-risk AI systems in the proposed EU AI Act, as well as the Basel III framework for the retention of operational data. An embedded version-control layer is designed to inform institutions of the exact model version that produced the specified output for any given date.

IV. PROPOSED EVALUATION METHODOLOGY

A. Comparative Systems

We propose three reference systems as architectural comparators for GRACE. First, that at the moment the most widespread general purpose LLM is GPT-4 (OpenAI, gpt-4-turbo). The standard off-the-shelf way of financial firms is to include a system prompt with the focus on compliance, and not on the domain, or with no domain-based fine-tuning. Second, Gemini Pro 1.5 (Google DeepMind) is also a multimodal LLM with good general reasoning abilities, and which can be used via API with a compliance-oriented prompt. The use here as a general purpose benchmark. Last but not least, BloombergGPT [7] is the most prevalent finance domain LLM in the published literature. BloombergGPT is a proprietary system and future empirical studies must be done using published performance data and then compared to a similarly trained open system.

B. Evaluation Dimensions

We define six evaluation dimensions, grounded in regulatory guidance (EBA, ECB, OCC, EU AI Act), risk management standards (SR 11-7, SS1/23), and established AI trustworthiness frameworks (NIST AI RMF [17], EU HLEG AI [18]):

D1 — Interpretability: the system's capacity to provide explanations adequate to meet regulatory explainability requirements.

D2 — Compliance Readiness: describes how much outputs need compliance specialists to undergo major transformations to meet regulatory standards.

D3 — Trustworthiness: a summation of output accuracy, consistency, calibration, and robustness to regulatory adversarial prompts.

D4 — Regulatory Auditability: the adequacy of auditing mechanisms in relation to the requirements set forth by the EU AI Act, Basel III, and ECB TRIM.
D5 — Bias and Fairness: the uniformity of outputs across the classes of protected characteristics in relation to credit risk and AML and in relation to gender, ethnicity and age.
D6 — Domain Specificity: the extent to which outputs meet the requirements of banking regulation as opposed to regulation for financial NLP as a whole.

C. Proposed Evaluation Dataset

The development of a Banking Compliance Evaluation Suite (BCES) with several thousand use cases within six regulatory task categories has been proposed. The regulatory task categories include credit risk assessments, AML transaction scenario testing, regulatory capital calculations, IFRS 9 expected-credit-loss (ECL) measurements, automated regulatory report generation, and compliance gap analysis. The development of this suite would be a collaboration with compliance officers of partnering banking institutions, with use cases validated against expert-reviewed ground truth. The suite will focus on edge cases and ambiguous input intentionally in order to test for tendencies to "hallucinate" and/or demographic biases amongst the compliance evaluators..

This dataset does not exist yet. The design parameters proposed here will serve as a specification for the empirical study described in Section VII.

D. Proposed Scoring Protocol

The following is a suggested approach to independently assessing each system with a panel of certified risk professionals on a 1.0-5.0 scale, and is based on dimension-specific rubrics. Based on the above, the following definitions will be used for these evaluations: Krippendorff's alpha is to be used for interrater reliability and the Wilcoxon signed rank test with Bonferroni correction for multiple comparisons are to be used for direct assessment of the systems, while the rate of outputs containing unverifiable regulatory assertions in comparison to official sources will be used as a definition for rate of hallucination. As yet there is no panel established and no evaluations have been carried out. We applied this protocol and intended to pursue the methodology in our empirical study.

V. ILLUSTRATIVE EVALUATION

A. Purpose and Scope of the Illustrative Scores

This section gives a case study scoring exercise so as to demonstrate how the proposed strategy outlined in Section IV would differentiate between systems. These values are illustrative estimates, not the product of an empirical investigation. They represent architectural capability mapping based on the documented or proposed characteristics of each system with respect to each evaluation dimension. They should not be interpreted as measured performance metrics because they were not generated by an expert panel and were not derived from the BCES dataset described in Section IV-C, which has not yet been established. They have been designed to yield testable hypotheses for empirical testing outlined in Section VII and to illustrate the discriminative performance of the evaluation framework expected.

Table I. Illustrative Architectural Capability Comparison of GenAI Approaches for Banking Risk & Compliance

Criteria	GPT-4 (OpenAI)	Gemini Pro (Google)	Bloomberg GPT	GRACE (Proposed)
Interpretability	Low — black-box	Low — opaque	Moderate — domain-tuned	High (expected) — XAI + audit trails by design
Compliance Readiness	Manual mapping required	Limited regulatory alignment	Partial (finance-specific)	Built-in by design: Basel III, IFRS 9, AML
Regulatory Audit Trail	None native	None native	Partial (corporate logging)	Full (immutable logs) by design
Human-in-the-Loop	Optional plugin	Optional	Limited	Core design principle

Domain Fine-tuning	Generic	Generic	Finance-oriented	Banking risk & compliance-oriented
Bias Detection	Not built-in	Partial	Limited	Integrated fairness layer by design

Table II. Illustrative Architectural Capability Matrix Across Evaluation Dimensions (Illustrative scale: 1.0–5.0; not empirically measured)

Dimension	GPT-4	Gemini Pro	Bloomberg GPT	GRACE (Proposed)
Interpretability	2.5	2.4	3.5	4.7
Compliance Alignment	2.8	2.6	3.7	4.8
Trustworthiness	3.1	3.0	3.8	4.6
Regulatory Auditability	1.5	1.6	3.0	4.9
Bias & Fairness	2.9	3.1	3.4	4.5
Domain Specificity	2.0	2.1	4.0	4.8
Illustrative Overall Average	2.47	2.45	3.57	4.72

B. Interpretability

GRACE's design, including the outputs of the proposed Explainability Engine which will be constrained by specific rules, is expected to be significantly more interpretable than the general-purpose models, the outputs of which, even if they were fluent, could not be traced to any specific rule and might not be shown to adhere to any specific rule. While BloombergGPT's domain training may yield some benefit over a general-purpose model, it would be difficult to verify the results for auditing without a specific explainability layer. This difference will be measurable by the compliance officer's ability to satisfy the requirements of a regulatory audit – a measure that will be captured in an experimental manner in the validation study.

C. Compliance Readiness

GRACE's proposed Regulatory Alignment Module is intended to specifically avoid the types of errors that are expected to occur in general purpose models, such as when IFRS 9 expected-credit-loss logic is applied to a Basel III internal-ratings scenario, or when citing an out-of-date regulation. This risk is likely to be lower for BloombergGPT, which is being trained on financial domain knowledge, but the necessity of active and continuous upkeep of a static knowledge base is a significant drawback. We believe GRACE's readiness for compliance benefit is most pronounced when up-to-date interpretation of regulation is required.

D. Trustworthiness

We expect that the RAM constraint layer implemented by GRACE will significantly decrease the number of general factual-accuracy errors and AML false-positives when compared to systems without a similar constraint-checking capability. We also hypothesize that GRACE's design would attract significantly more regulatory prompts that would be considered adversarial compared to other more general-purpose systems that use only prompt-level guardrails. The magnitude of this advantage can only be gleaned by the adversarial-prompt testing protocol suggested in Section IV.

E. Regulatory Auditability

Auditability is expected to represent the largest architectural gap between GRACE and the comparator systems, since The scope for auditability is likely to be the biggest architectural gap between GRACE and the comparator systems, with neither GPT-4 nor Gemini Pro having a native, regulator-grade audit trail, and the logging that Bloomberg GPT has developed, though more advanced, being designed to meet EU AI Act high-risk system requirements for cryptographic tamper-evidence and regulatory tagging. We highlight that auditability is not a performance characteristic, but rather a

regulatory requirement: ECB TRIM expectations require institutions to show the supervisors through which audit steps and with which data, model version or auditing, a specific model-driven risk assessment was calculated. GRACE's audit layer is intended to meet this requirement directly; to replicate the equivalent traceability in a general-purpose model would entail a great deal of custom infrastructure that would reduce any cost benefit of the off-the-shelf approach.

F. Bias and Fairness

It has been widely documented that subtle variations in language used to describe risk are correlated with protected attributes when financial risk is the same, which presents a challenge for general purpose LLMs deployed in credit risk and AML scenarios. GRACE's proposed fairness layer is intended to watch outputs to ensure this sort of disparate treatment among customer groups. These are similar concerns raised by the Consumer Financial Protection Bureau in their comment about the use of AI in consumer credit assessment, and the fact that the AI might be seen to have a disparate impact could lead to a potential violation of the Equal Credit Opportunity Act or equivalent EU non-discrimination provisions. Future evaluation study will focus on the effectiveness of GRACE's fairness layer, not just its presence, as this will be an empirical validation.

VI. DISCUSSION

A. The Case for Purpose-Built Frameworks

Table III. Comparative Analysis of General-Purpose LLMs and Governance-Aware AI Frameworks for Banking Risk Management and Regulatory Compliance

Theme	Finding	Implications for Banking Risk Management
Regulatory Compliance	General-purpose LLMs lack dedicated compliance mechanisms for banking regulations.	Limits their suitability for regulated banking and risk-management environments.
Auditability and Traceability	Current LLMs provide limited capabilities for tracing, monitoring, and validating decision-making processes.	Limits the effectiveness of regulatory audits, oversight, and accountability mechanisms.
Explainability	LLM-generated outputs are often difficult to explain to regulators and financial institutions.	Reduces transparency, interpretability, and trust in risk-related decisions.
Domain-Specific Training	Training on specialized financial and banking datasets, as demonstrated by Bloomberg GPT, improves domain understanding and contextual accuracy.	Highlights the value of domain-specific training for financial applications.
Limitations of Domain Training	Domain-specific training alone cannot address governance, fairness, auditability, and oversight concerns.	Regulatory compliance requires more than financial expertise; governance mechanisms are also essential.
Bias and Fairness Risks	LLMs may produce biased outputs if fairness controls and monitoring mechanisms are absent.	Creates ethical, legal, reputational, and regulatory risks for financial institutions.
Human Oversight	General-purpose LLM governance frameworks often lack structured human-in-the-loop controls.	Underscores the need for human oversight in high-stakes banking decisions.
Governance-Oriented Architecture	The proposed GRACE framework incorporates explainability, regulatory alignment, auditability, and human oversight.	Offers a governance-oriented approach that may better support regulatory requirements, pending validation.

Architectural vs. Knowledge Gap	Limitations of general-purpose LLMs arise primarily from architectural design rather than a lack of regulatory knowledge alone.	Emphasizes the importance of governance-focused system architecture for regulated deployment.
Research Conclusion	Successful AI adoption in financial institutions requires integrating domain expertise, explainability, auditability, fairness controls, and human oversight.	Supports continued research and development of governance-aware AI frameworks for banking and financial services.

B. Implementation Considerations

There are a number of practical limitations that would be encountered in the cases of institutions adopting such a protocol as GRACE. First, the regulatory knowledge graph supporting the RAM would need to be maintained on an on-going basis to ensure that it evolves with changing guidance; the more up-to-date and complete the data in the module, the more valuable the module becomes. Second, the HITL interface adds workflow latency that could prove too much for real-time transaction screening, and future iterations of the design should take into account multiple levels of oversight—low-risk transactions to be handled automatically, higher-risk ones for human review. Third, the curation process to fine-tune the data, that demands close collaboration with experts in compliance, has a high resource requirement, is difficult to fully automate, and is a scalability challenge both within and across jurisdictions and regulatory regimes.

C. Limitations

There are a number of significant limitations to this study. First, and most importantly, GRACE has not yet been put in place, and the scores shown in Section V are only illustrative estimates based on architectural reasoning not an actual empirical study. These illustrative figures may not accurately represent the actual performance of a realized GRACE system when applied to actual test case. Second, Bloomberg GPT is not publicly available, so even if a future empirical comparison could be had the metrics reported in the public literature or a replicated system would have to be used, but would not necessarily capture the full range of system properties.Thirdly, the proposed evaluation suite for BCES, even if built, would in any case be limited in scope and might not cover all the different edge cases in live banking scenarios. Fourth, future empirical tests with a small number of partner institutions within a single region of regulation would lack generalizability to other regions of regulation, including U.S. federal banking regulation or Asian regulatory regimes. The purpose of this paper is to make a contribution as an architecture-at-the-art of the Internet. Empirical validation remains future work and has not yet been accomplished.

VII. CONCLUSION

While Generative AI (GenAI) is set to make a mark in financial services, the use of this technology in banking risk and compliance throws up the critical questions of interpretability, auditability and trustworthiness in the highly regulated sector. Large language models (LLMs) such as GPT-4 and Gemini Pro are not specialized for banking and do not comply with regulatory rules and standards. For banking risks and compliance, a novel GenAI-based banking risk and compliance framework (Purpose-built "GRACE" – Generative Risk and Compliance Evaluation Framework) is introduced that includes Explainable AI (XAI), a cryptographically secured Immutable Audit Trail, a Human-in-the-Loop (HITL) oversight layer and dedicated compliance alignment layers for Basel III, IFRS 9, AML/CFT and GDPR. Beyond the architecture, we suggest a method for evaluating GRACE and representative comparator systems (GPT-4, Gemini Pro, and BloombergGPT) by six criteria: interpretability, compliance readiness, trustworthiness, regulatory auditability, bias and fairness, and domain specificity.

A proposed evaluation protocol is presented to assess the adaptability of the systems through an illustrative architectural capability assessment. This should be a theoretical exercise as opposed to an empirical study and the scoring exercise should be represented as the value the exercise has for them. We are not able to offer a real value for scoring, therefore. Whereas purpose-built systems are more architecturally flexible, domain tuned GenAI systems are less flexible. We discuss below a draft of a possible real-world test of this work. This encompasses a Banking

Compliance Evaluation Suite, scoring protocol devised by a panel of experts, and a statistical testing framework.

VIII. ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the compliance practitioners who gave domain feedback at the conceptual stage of the development of the GRACE architecture, and to their research supervisors for guidance and support throughout this work.

Conflicts of Interest

The authors have no conflict of interest to declare related to the publication of this paper.

Funding Statement

No specific grant funding from any public, commercial or not-for-profit funding agencies was received for this research.

REFERENCES

[1] R. M. N. Gunasekaran, "Generative AI: Opportunities, Risks and Implications for Financial Services," *The Eastasouth Journal of Information System and Computer Science*, vol. 1, no. 3, pp. 219–234, Apr. 2024, doi: 10.58812/esiscs.v1i03.1098.

[2] F. Petropoulos, D. Apiletti, V. Assimakopoulos, M. Z. Babai, *et al*., "Forecasting: Theory and Practice," *International Journal of Forecasting*, vol. 38, no. 3, pp. 705–871, Jul.–Sep. 2022

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, *et al*., "On the Opportunities and Risks of Foundation Models," Center for Research on Foundation Models (CRFM), Stanford University, *arXiv preprint* arXiv:2108.07258, 2021.

[4] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring Massive Multitask Language Understanding," arXiv preprint arXiv:2009.03300, 2020. doi: 10.48550/arXiv.2009.03300.

[5] A. Alansari and H. Luqman, "Large Language Models Hallucination: A Comprehensive Survey," *arXiv preprint* arXiv:2510.06265, 2025. doi: 10.48550/arXiv.2510.06265

[6] Xie, W. Han, Z. Chen, R. Xiang, X. Zhang, Y. He, *et al*., "FinBen: A Holistic Financial Benchmark for Large Language Models," *arXiv preprint* arXiv:2402.12659, 2024. Available: https://arxiv.org/abs/2402.12659

[7] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "BloombergGPT: A Large Language Model for Finance," arXiv preprint arXiv:2303.17564, 2023. doi: 10.48550/arXiv.2303.17564.

[8] K. Du, Y. Zhao, R. Mao, F. Xing, and E. Cambria, "Natural Language Processing in Finance: A Survey," *Information Fusion*, vol. 110, Art. no. 102755, 2024, doi: 10.1016/j.inffus.2024.102755.

[9] A. Narang, P. Vashisht, and S. Bhaskar, "Artificial Intelligence in Banking and Finance," *International Journal of Innovative Research in Computer Science & Technology*, vol. 12, no. 2, pp. 130–134, Mar. 2024, doi: 10.55524/IJIRCST.2024.12.2.23.

[10] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv preprint* arXiv:1702.08608, 2017. doi: 10.48550/arXiv.1702.08608.

[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[12] European Central Bank, Guide to Internal Models: General Topics Chapter. Frankfurt am Main, Germany: European Central Bank, Nov. 2018. Available: ECB Guide to Internal Models – General Topics Chapter (2018)

[13] European Parliament, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act),” Official Journal of the European Union, 2024.

[14] D. W. Arner, J. N. Barberis, and R. P. Buckley, "The Evolution of FinTech: A New Post-Crisis Paradigm?" University of Hong Kong Faculty of Law Research Paper No. 2015/047, UNSW Law Research Paper No. 2016-62, Oct. 2015. doi: 10.2139/ssrn.2676553.

[15] Deloitte, "Harnessing Generative AI for Regulatory Compliance: Building Trust and Ensuring Ethical Innovation," Deloitte

Insights/Perspective, 2023. [Online]. Available: https://www2.deloitte.com/.

[16] C. Nie, Y. Liu, and C. Wang, "AI Application in Anti-Money Laundering for Sustainable and Transparent Financial Systems," *arXiv preprint* arXiv:2512.06240, 2025. doi: 10.48550/arXiv.2512.06240.

[17] National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," NIST, Gaithersburg, MD, NIST AI 100-1, 2023.

[18] European Commission High-Level Expert Group on AI, "Ethics guidelines for trustworthy AI," European Commission, Brussels, Belgium, 2019.