An approach to integrate generative artificial intelligence into automated credit decision-making processes

Lorenzo Quirini (Monte dei Paschi di Siena), Leandro Guerra (Experian), Giulio Mariani (Experian)

Abstract: This paper introduces an operational architecture for integrating generative artificial intelligence into automated credit decision-making systems within the banking sector. The proposed framework combines a conventional Automated Decision-Making System (ADMS), responsible for the quantitative assessment of applicants through risk, affordability, and profitability indicators, with a Retrieval-Augmented Generation (RAG) layer that interprets, validates, and contextualizes each decision, considering internal policies and external regulatory obligations. While the ADMS produces structured outcomes - approval, rejection, or referral - based on scoring models and rule-based thresholds, the RAG component evaluates their consistency with legal and ethical standards, including the General Data Protection Regulation (GDPR), the Artificial Intelligence Act, and responsible lending principles. The generative layer produces a structured explanation grounded in versioned, traceable knowledge bases, ensuring transparency, normative alignment, and resistance to hallucination. Beyond post-hoc justification, the architecture enables a feedback-oriented ecosystem in which insights from interpretive reasoning can inform adjustments to the scoring process itself. The result is a hybrid decision infrastructure that is not only statistically rigorous but also auditable, adaptable, and aligned with the future of compliant, human-centered credit automation.

1. Introduction

In the context of credit origination targeted at households, financial institutions increasingly rely on automated decision-making systems (ADMS) to evaluate applications for products such as mortgages, personal loans, and overdrafts. These systems, rooted in structured data and algorithmic logic, aim to deliver timely, scalable, and risk-aligned assessments of credit requests. The operational outcome is typically synthesized into three possible decisions: approval (green), rejection (red), or the need for further analysis (yellow), depending on the risk, affordability, and profitability profile of the applicant.

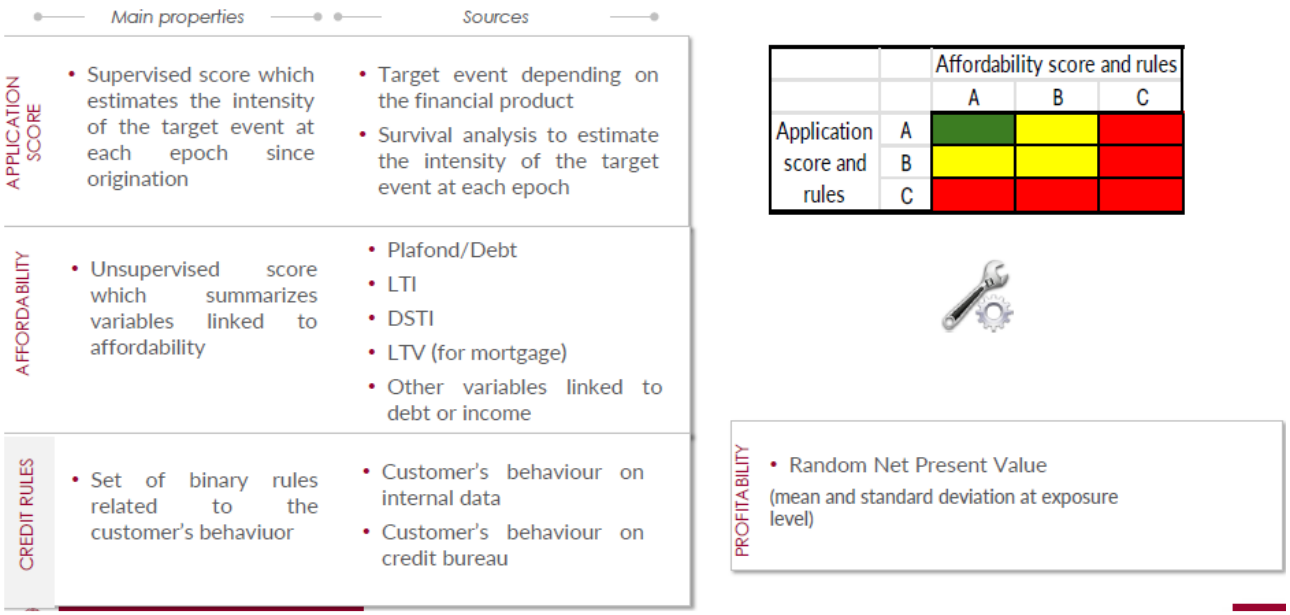

Figure 1. Example of ADMS operational output space (approve, reject, manual review).

While the efficiency and consistency of ADMS have proven indispensable, these systems remain primarily designed to evaluate quantitative metrics. They lack embedded mechanisms to assess whether a given decision is fully aligned with broader ethical, regulatory, and institutional principles. As the regulatory landscape becomes more dynamic, with evolving requirements around fairness, explainability, and data privacy, traditional rule-based approaches struggle to maintain interpretability and normative coherence. Compliance with frameworks such as the General Data Protection Regulation (GDPR), the forthcoming European Artificial Intelligence Act, and internal codes of conduct increasingly demands a layer of reasoning that goes beyond pure statistical scoring.

This paper introduces a hybrid decision architecture that retains the operational backbone of a conventional ADMS while enhancing it through the integration of a generative artificial intelligence (GenAI) module. Specifically, we explore a Retrieval-Augmented Generation (RAG) approach, whereby the outputs of the ADMS are transmitted to a generative model tasked with evaluating their consistency with regulatory and institutional norms and producing a structured justification. The generative layer does not alter the original decision logic but complements it by adding a second axis of interpretability. In doing so, we aim to bridge the gap between statistical rigour and normative transparency, creating a decision framework that is both auditable and adaptive.

2. The Quantitative Core: Automated Credit Decision Systems

At the heart of the proposed framework lies an Automated Decision-Making System (ADMS), designed to operate across a variety of regulatory and credit environments while maintaining a consistent internal logic based on risk, affordability, and profitability. The ADMS is structured to deliver a discrete outcome for each credit application - approval, rejection, or referral for further analysis - based on a combination of statistical scores, eligibility rules, and profitability metrics. These decisions are not made in isolation; rather, they are the result of a modular architecture that integrates supervised and unsupervised models, credit policy constraints, and dynamic product-specific thresholds.

The quantitative core of the ADMS consists of three analytical components. The first is a supervised score that estimates the probability and intensity of a target adverse event (such as default or early delinquency) over time. This model is trained using historical loan performance data and adapted to the specific characteristics of the product in question, whether it be a personal loan, mortgage, or credit line. The second component is an affordability block, which aggregates unsupervised scores and rule-based evaluations to assess whether the applicant has sufficient financial capacity to absorb the requested obligation. Variables such as the debt-to-income ratio (DTI), loan-to-income (LTI), and, for mortgage, loan-to-value (LTV), are used alongside behavioural clustering to generate a synthetic affordability index. The third component incorporates a set of binary rules derived from applicant behaviour, including their past interactions with the bank and credit bureau information, to flag deviations from acceptable usage patterns or historical precedents.

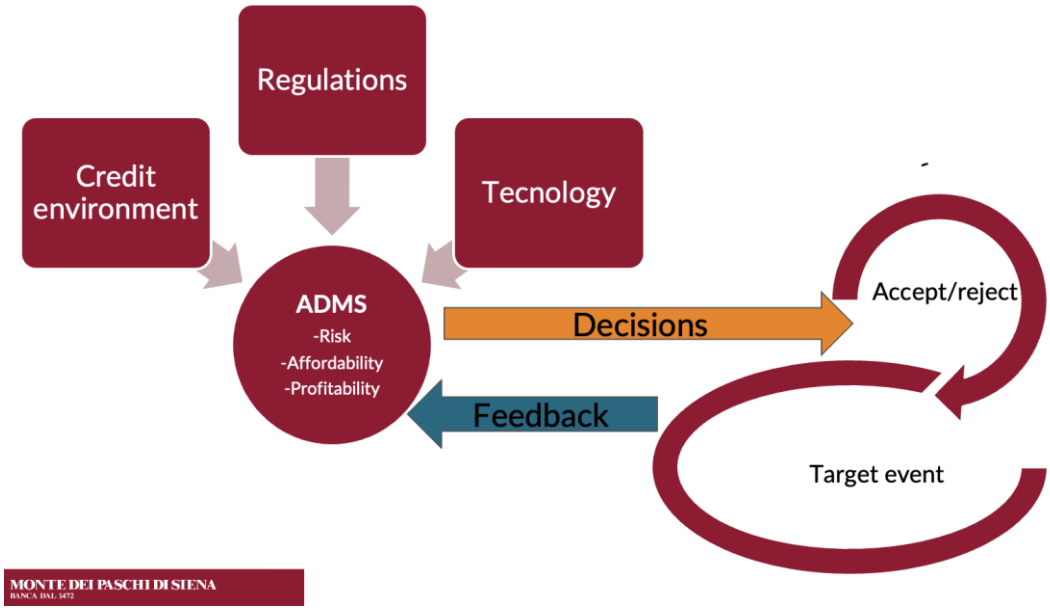

Figure 2. Architecture of the ADMS decision logic, combining supervised and unsupervised scoring with rule-based assessments.

In addition to this multi-dimensional scoring architecture, the system also computes the expected profitability of the operation through a Random Net Present Value (rNPV) moments like mean and standard deviation. This calculation incorporates stochastic projections of repayment flows, expected recovery values in case of default, and cost of capital assumptions based on macroeconomic scenarios.

Since the rNPV depends, among other factors, on the probabilities of installment payments over time, these probabilities can be adjusted in function of the customer's affordability to help prevent over-indebtedness.

The final decision output is then mapped into one of the three outcome classes, green (approve), red (reject), or yellow (manual review), depending on the interaction of risk, affordability, and profitability signals.

This architecture enables consistent, scalable, and auditable decisions. However, it remains fundamentally quantitative. While the ADMS can express whether a credit operation satisfies predefined financial constraints, it does not inherently address whether the outcome is consistent with regulatory expectations, internal governance norms, or ethical lending practices. The result, although statistically sound, remains silent on the broader context in which the decision will be interpreted and eventually scrutinized.

3. The interpretive layer: generative AI and the role of RAG

To expand the decision-making process beyond purely quantitative reasoning, a second analytical layer has been developed using Generative Artificial Intelligence, implemented through a Retrieval-Augmented Generation (RAG) framework. Designed to contextualize the output of the Automated Decision-Making System (ADMS), this layer synthesizes structured legal, regulatory, and institutional knowledge to assess the normative consistency of each credit decision in real time.

The RAG system, developed within the ExperianGPT platform, combines the linguistic capabilities of a large language model with a retrieval engine capable of accessing curated, versioned knowledge bases. These repositories include national and European regulatory documents such as the Italian Consolidated Banking Law, the Code of Conduct for Credit Information Systems, the General Data Protection Regulation (GDPR), and the forthcoming Artificial Intelligence Act. In addition, internal credit policies, decision protocols, and institutional conduct rules are incorporated into the same framework. All documents are managed within a sandboxed infrastructure that enforces ICT segregation, source validation, and full traceability.

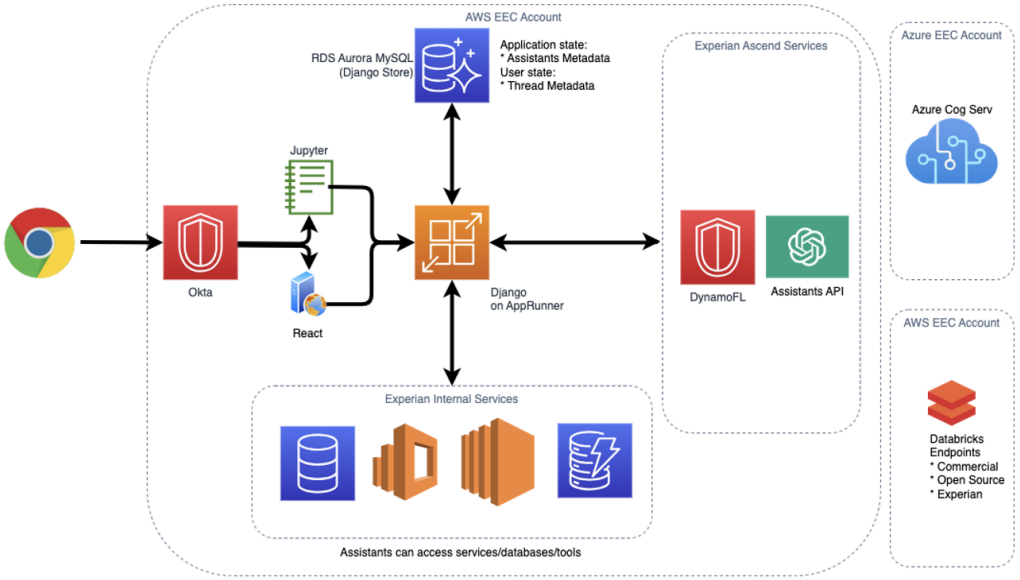

Figure 3. ExperianGPT architecture based on Retrieval-Augmented Generation, integrating ADMS outputs with normative reasoning.

Once the ADMS issues its output and supporting indicators, this information is transmitted directly to the RAG layer via a secure system-to-system connection. The model then performs two key tasks. First, it conducts a normative validation of the decision, confirming whether the proposed action aligns with applicable regulatory and internal standards or flagging elements that require revision or escalation. This assessment is grounded entirely in retrieved content; no output is generated without a reference to documented knowledge. Second, the system constructs a structured explanation of the decision rationale, which may take the form of a tabular synthesis or a natural language justification. These outputs incorporate the relevant financial attributes of the case, reference the underlying legal or policy sources applied, and are formatted to support both compliance workflows and credit analyst review.

By embedding this layer into the decision pipeline, the institution establishes a "four-eyes" model supported by artificial intelligence, where human analysts are equipped with a consistent, traceable narrative to complement and critically evaluate the ADMS outcome. For instance, in a case where the installment-to-income ratio is high and profitability is negative, the system may generate an explanation such as: "Given the applicant's 33% installment burden and a projected rNPV of -€640, the credit request fails to meet the institution's sustainability and profitability thresholds defined in Policy Section 2.3." This output not only improves transparency but also helps bridge the operational divide between model-generated risk scores and human judgment, transforming credit decisions into fully contextualised and explainable outcomes.

4. Ensuring compliance, fairness, and ethical integrity

One of the defining advantages of incorporating a generative layer into credit decision-making lies in its ability to bring regulatory and ethical alignment directly into the operational core of the process. In traditional architectures, compliance is often addressed through after-the-fact audits or embedded in rigid rule sets, making it ill-suited to respond to ambiguous cases, normative evolution, or edge scenarios that fall outside codified logic. By contrast, the RAG layer operates synchronously with each individual decision, enabling contextual and normative checks to occur in real time, and not as retrospective control mechanisms.

This dynamic alignment operates across several critical dimensions. The first is fairness—ensuring that decisions do not produce direct or indirect discrimination based on sensitive attributes or correlated proxies. Because the RAG system grounds its output in verifiable legal and policy documents, it can identify when a rule or score application may produce disparate impact, even if unintentionally. The second-dimension concerns privacy: by referencing data lineage and variable usage policies stored in the institutional knowledge base, the system can flag uses of data that exceed what is permitted under GDPR or internal governance. This is particularly important when behavioural data or inferred variables are involved, as these may introduce the risk of profiling or a lack of informed consent.

A third dimension of normative integrity involves the institution's obligation to avoid encouraging or enabling over-indebtedness. While profitability indicators such as rNPV can highlight negative expected value, they may not always capture ethical boundaries, such as issuing loans that are technically viable but likely to place the applicant in financial distress. The RAG layer addresses this

by incorporating responsible lending guidelines and internal caps on financial stress indicators, allowing for qualitative reasoning that goes beyond pure expected value optimization.

Crucially, these checks are not applied at the portfolio level or as batch processes; they are embedded at the micro-decision level, ensuring that each operation is evaluated not only for financial soundness but also for normative coherence. This produces a dual benefit: on the one hand, it strengthens institutional robustness by pre-empting decisions that could later be deemed non-compliant or ethically problematic; on the other, it fosters public and supervisory trust by ensuring that the institution can explain and justify its actions in terms of both policy and principle. In this model, compliance is not a detached obligation, and it is operationalized as a native property of the credit decision-making pipeline itself.

5. Controlling hallucinations and ensuring trustworthy generation

The adoption of generative AI in regulated domains such as credit decision-making presents a fundamental challenge: the risk of hallucination, wherein the model produces outputs that are linguistically coherent but factually incorrect, legally unfounded, or untraceable to authoritative sources. In settings governed by legal obligations—such as GDPR, the AI Act, and banking supervision frameworks—this risk is not merely reputational; it has direct implications for institutional accountability, auditability, and regulatory compliance. For this reason, the architecture implemented in our RAG system is deliberately constrained to maximize interpretability and minimize uncontrolled generation.

The system is built around a retrieval-first paradigm. In this configuration, the language model does not generate content from latent associations alone; instead, it is prompted to generate only when grounded in retrieved materials from a curated, versioned knowledge base. These documents include national and European regulatory texts, internal credit policy manuals, model governance guidelines, and prior decision precedents. All content is indexed and version-controlled, allowing for explicit traceability of sources referenced during the generative process. In practical terms, this means that every output produced by the system can be mapped to a discrete set of documents, with metadata such as version number, date of last review, and internal custodian available for verification.

In instances where the system is unable to locate sufficient or reliable material to ground a response, whether due to gaps in documentation or ambiguity in the query, the model is programmed to abstain. Instead of extrapolating or speculating, it flags the case for human review, invoking a fallback mechanism that maintains integrity at the cost of completeness. This ensures that the model operates not as a creative engine, but as a disciplined interpreter—a tool that constructs meaning within the strict epistemic boundaries of validated institutional knowledge.

To maintain operational relevance, the system is also periodically fine-tuned using anonymized examples of real credit decisions, drawn from recent operational history. These examples are used not to train new decision logic, but to calibrate the model's ability to match institutional tone, formatting conventions, and explanatory standards. Importantly, all data used in this process is fully anonymized and pre-processed to eliminate personally identifiable information, ensuring compliance with data protection regulations. The result is a generative system that evolves with the institution, but remains anchored in a framework of epistemic discipline, legal validity, and operational trust.

6. Toward a feedback-driven decision ecosystem

The current implementation of the system operates in a forward-only configuration: the Automated Decision-Making System (ADMS) performs the primary credit evaluation, and the Retrieval-Augmented Generation (RAG) layer acts as a post-decisional interpreter, verifying, justifying, or refining the outcome. While this unidirectional architecture already introduces significant gains in explainability, compliance, and operational trust, it also opens a clear trajectory toward a more ambitious objective: the construction of a feedback-driven ecosystem in which generative and predictive components do not merely coexist but evolve together.

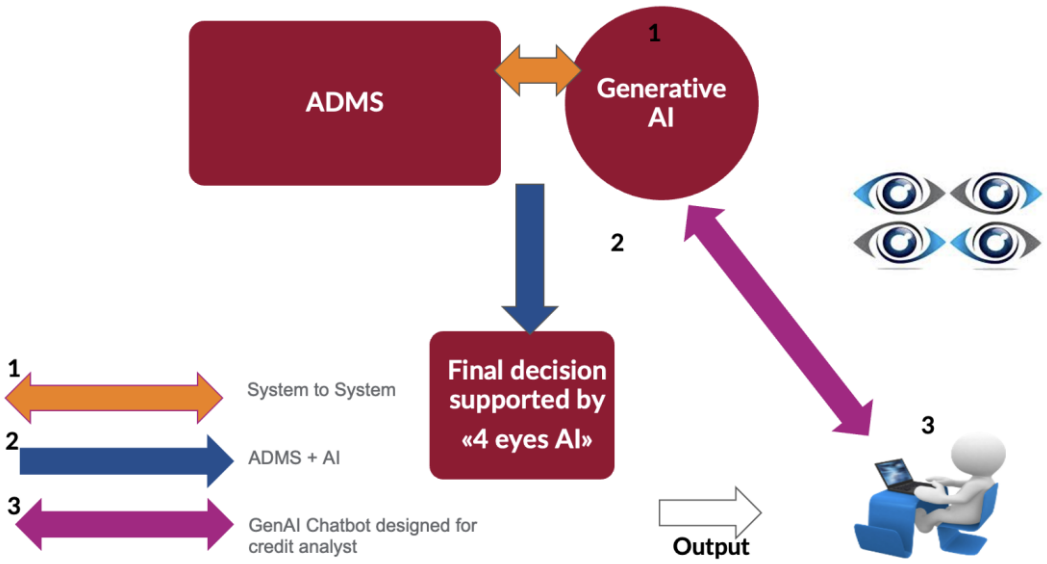

Figure 5. Feedback-driven architecture for adaptive credit decisioning through RAG-informed adjustments.

In this extended architecture, insights derived from the RAG system could be systematically reintegrated into the ADMS pipeline. For example, recurrent patterns identified through natural language justifications, such as frequent inconsistencies between scoring outputs and normative constraints, may prompt a re-examination of feature selection, rule thresholds, or override protocols within the core decision logic. Similarly, flagged mismatches between ADMS outcomes and internal credit conduct policies could serve as empirical signals for refining eligibility criteria or adjusting model calibration to better reflect institutional risk appetite and ethical boundaries.

Beyond technical recalibration, the RAG layer may also contribute to model governance. Its structured outputs could be logged and analyzed across time to detect drift in model behaviour, shifts in operational context, or the emergence of decision patterns that warrant supervisory scrutiny. This creates the foundation for a virtuous loop: operational data feeds the ADMS; ADMS outputs are interpreted by the RAG; RAG rationales generate institutional learning; and this learning, in turn, informs the evolution of both systems.

Such a configuration moves the institution from static automation to adaptive intelligence—a state in which the decision infrastructure becomes responsive not only to data but also to norms, experience, and change. It reflects a deeper maturity in the deployment of artificial intelligence in finance: not as a set of frozen rules or black-box optimizer, but as a living system capable of learning, aligning, and improving in step with its regulatory, ethical, and human context.

7. Conclusion

This paper has presented an integrated architecture for enhancing automated credit decision-making through the application of generative artificial intelligence, grounded in the concrete operational context of banking institutions subject to European regulatory frameworks. By combining a robust statistical engine with a retrieval-based generative reasoning layer, the proposed system introduces a new class of credit infrastructure, one capable of delivering decisions that are not only efficient but also interpretable, justifiable, and aligned with both institutional policy and legal obligation.

Beyond its immediate operational benefits, this architecture suggests broader implications for the future of credit governance. As financial institutions adopt AI-based decision tools at scale, the demand for systems that are transparent by design, and not merely auditable in retrospect, will become increasingly non-negotiable. In this sense, the integration of RAG into decision-making workflows may help shape future regulatory standards, setting a precedent for how automated decisions should be explained, validated, and monitored. Moreover, the ability to trace decisions back to legal sources and institutional norms opens new possibilities for how banks engage with their clients, potentially fostering a more dialogical and trust-based relationship, where credit outcomes can be meaningfully communicated and contested.

Equally important is the redefinition of the human role in this landscape. Far from removing human judgment, the architecture proposed here elevates it, by equipping analysts with structured, context-rich rationales that support critical thinking and discretionary oversight. In doing so, it positions generative AI not as a substitute for expert intervention, but as a tool for institutional alignment and cognitive augmentation.

Ultimately, the model we propose is not simply a technical enhancement, but a conceptual reframing of what it means to make credit decisions in a complex, regulated, and ethically demanding environment. It signals a shift from automation as a goal to automation as a vehicle for accountability, adaptability, and institutional intelligence.