Generative Artificial Intelligence in Internal Audit: A Process-Integrated Framework for AI-Assisted Review of IRBA Rating Procedures under Art. 191 CRR

Working Paper — auditissimo Research Project
Hochschule Reutlingen · Steinbeis Transfer Center Data Analytics and Predictive Modelling · msg for banking ag

Prof. Dr. Dirk Schieborn* Prof. Dr. Volker Reichenberger† Tim S. Körwers‡

Draft version — March 2026

Abstract

Internal auditors at credit institutions face an acute expertise challenge when reviewing Internal Ratings-Based Approach (IRBA) models under Art. 191 of the Capital Requirements Regulation (CRR): the technical depth required—spanning logistic regression scorecards, calibration theory, and complex regulatory guidance from EBA and ECB—demands competencies that are rarely concentrated in a single audit team. This paper presents *auditissimo*, a modular generative AI (GenAI) framework specifically designed to support—not replace—the internal auditor across the full lifecycle of an IRBA model audit. The system decomposes the audit workflow into six functionally cohesive modules (M1–M6), ranging from regulatory requirement atomisation and data-driven risk assessment through automated gap analysis and deep-dive document review to finding synthesis and report generation. We argue that process proximity is the decisive factor determining whether AI support translates into audit quality improvements or introduces new risks: the model must encode the domain logic of banking supervision, not merely perform generic text retrieval. We further delineate the decision points at which human judgment is not only preferable

*Reutlingen University, Steinbeis Transfer Center Data Analytics and Predictive Modelling `dirk.schieborn@reutlingen-university.de`

†Reutlingen University, Steinbeis Transfer Center Data Analytics and Predictive Modelling `volker.reichenberger@reutlingen-university.de`

‡msg for banking ag, `tim.koerwers@msg.group`. This paper is based on the auditissimo prototype developed in collaboration between Reutlingen University, Steinbeis Transfer Center Data Analytics and Predictive Modelling, and msg for banking ag. The Steinbeis Bank synthetic IRBA suite referenced herein was developed specifically as a controlled research environment for this project.

but constitutionally required under Three-Lines-of-Defense governance. Finally, we propose a rigorous empirical validation methodology exploiting the fictional Steinbeis Bank synthetic IRBA environment—four fully parameterisable rating models (Corporate PD/LGD, Retail PD/LGD) together with machine-generated validation concept and report documents—to conduct controlled accuracy testing of LLM-based gap detection. Early results indicate that structured prompt engineering with domain-specific context achieves F1 scores in excess of 0.78 for requirement-level compliance classification, with the residual error concentrated in nuanced regulatory interpretation tasks that require human expertise.

Keywords: Generative AI, Internal Audit, IRBA, Credit Risk Models, Large Language Models, Gap Analysis, Human-in-the-Loop, CRR Art. 191, Prompt Engineering, Model Validation

Contents

1 Introduction **4**

2 Background and Related Work **5**

2.1 IRBA Regulatory Requirements and Audit Obligations 5

2.2 Large Language Models as Regulatory Assistants 6

2.3 Human-AI Collaboration in Expert Knowledge Tasks 6

3 The auditissimo Architecture **6**

3.1 Design Principles 6

3.2 System Overview 7

3.3 Module 1: Regulatory Basis (Single Source of Truth) 9

3.4 Module 2: Risk Assessment 10

3.5 Module 3: Work Paper Initialisation 10

3.6 Module 4: Gap Analysis 10

3.7 Module 5: Deep Dive 11

3.8 Module 6: Report and Finding Synthesis 11

4 Human-in-the-Loop: Where AI Must Defer to Auditor Judgment **12**

4.1 The Governance Imperative 12

4.2 HITL Interface Design 13

4.3 Risk of Over-Reliance 13

5 Validation Methodology and Empirical Results **13**

5.1 The Challenge of Evaluating GenAI Audit Tools 13

5.2 The Steinbeis Bank Synthetic IRBA Environment 14

5.3 Controlled Testing Methodology 18

5.4 Preliminary Results 19

5.5 Prompt Engineering Findings 21

6 Discussion **23**

6.1 Process Proximity as the Decisive Quality Factor 23

6.2 The Scalability Potential 23

6.3 Limitations and Open Problems 24

6.4 Regulatory and Ethical Considerations 24

7 Future Research Directions **25**

8 Conclusion **25**

1 Introduction

The regulatory validation of Internal Ratings-Based Approach (IRBA) rating systems constitutes one of the most technically demanding tasks in bank internal audit. Article 191 CRR obliges institutions to assess the design, implementation, and ongoing performance of their own rating models through an independent internal audit function. The corresponding supervisory expectations, articulated in the EBA Guidelines on internal governance [5] and the ECB Guide to Internal Models [6], require a level of quantitative sophistication that challenges even specialist audit teams. Auditors must review and assess evaluations of validation dimensions such as discriminatory power (Gini, AUC), calibration quality (binomial tests, Hosmer-Lemeshow statistics), stability (Population Stability Index, migration matrices), and the procedural governance of model changes—all against a backdrop of dense regulatory text that continues to evolve.

Large Language Models (LLMs) offer a qualitatively new category of support tools for knowledge workers navigating complex document corpora. Unlike rule-based automation or classical information retrieval, contemporary LLMs trained at scale can parse regulatory text, extract structured requirements, compare document versions, and generate natural-language explanations of technical findings. Whether these capabilities translate into reliable audit quality improvements depends, however, on several factors that have received insufficient attention in the emerging literature on AI in finance [8, 9]: the degree to which the AI system encodes domain knowledge specific to banking supervision, the care with which human oversight is maintained at epistemically sensitive junctures, and the existence of a principled empirical methodology for evaluating AI output quality.

This paper addresses all three factors in the context of *auditissimo*, a prototype GenAI suite developed through a collaboration between Hochschule Reutlingen, Steinbeis Transfer Center Data Analytics und Predictive Modelling, and msg for banking ag. The paper makes the following contributions:

(i) We describe a modular, process-integrated architecture (six modules M1–M6) for GenAI-assisted IRBA audits, articulating the information flows, tool boundaries, and governance checkpoints of each module (§3).

(ii) We analyse the conditions under which AI support enhances audit quality and identify the structural reasons why certain tasks *must* remain under human control (§4).

(iii) We propose an empirical validation framework that exploits parametric manipulation of a synthetic IRBA environment to generate controlled ground-truth test cases, and we report baseline accuracy metrics for the gap-detection module (§5).

(iv) We discuss implications for prompt engineering, regulatory compliance of AI-assisted auditing, and directions for further research (§6).

2 Background and Related Work

2.1 IRBA Regulatory Requirements and Audit Obligations

The Internal Ratings-Based Approach (IRB), set out in Articles 142–191 CRR [1], allows institutions—subject to prior supervisory permission—to use internal rating systems for regulatory capital purposes. Under the IRB framework, institutions estimate probability of default (PD) and, depending on the exposure class and the type of IRB permission granted, may also use own estimates of loss given default (LGD) and credit conversion factors / exposure-at-default-related measures (CCF/EAD components). For retail exposures, own LGD estimates and, where applicable, own IRB-CCF estimates are required; for other exposure classes, the use of own LGD and conversion factor estimates depends on the applicable IRB permission.

Article 191 CRR requires the internal audit function, or another comparable independent auditing unit, to review at least annually the institution's rating systems and their operation, including the estimation of PDs, LGDs, expected losses and conversion factors; the review must cover compliance with all applicable IRB requirements.

The associated regulatory and supervisory framework is extensive. In particular, Commission Delegated Regulation (EU) 2022/439 [4] specifies the methodology competent authorities are to apply when assessing institutions' compliance with IRB requirements, including governance, validation, use test, rating assignment, model design and model performance. In relation to internal audit, it requires supervisors to verify, among other things, that all rating systems are reviewed at least annually and that the annual work plan appropriately identifies areas requiring more detailed review.

More detailed expectations on model performance assessment, override analysis, representativeness and validation practices are further elaborated in ECB supervisory materials and related validation-reporting instructions [6]. For example, ECB documentation refers to the analysis of overridden assignments, representativeness checks, and statistical validation tools such as AUC/gAUC-based discriminatory power analysis, Jeffreys-test-based PD back-testing, and PSI calculations for certain portfolio-distribution analyses.

The sheer volume and technical specificity of these requirements—spanning CRR, EBA Regulatory Technical Standards, EBA Guidelines, and ECB supervisory expectations—creates a formidable documentation burden. A typical Retail PD validation report for a mid-sized institution references upwards of forty distinct regulatory criteria, each requiring specific quantitative evidence. Identifying which criteria are addressed, which are only partially addressed, and which are absent is a labour-intensive task ideally suited to

augmentation by LLMs.

2.2 Large Language Models as Regulatory Assistants

The use of LLMs in financial services has expanded rapidly since the public availability of GPT-4-class models [7]. Applications range from contract review [10] and earnings call analysis [11] to regulatory interpretation [12]. However, the literature on LLM use in internal audit specifically, and in banking supervision more broadly, remains sparse.

Key challenges identified in existing work include hallucination—the tendency of LLMs to generate plausible but factually incorrect statements [13]—and sensitivity to prompt formulation [14]. In regulated environments, hallucination carries particular risk: an AI system asserting regulatory compliance where none exists could lead to material audit failures. This underlines the importance of rigorous empirical validation rather than relying on subjective assessments of output quality.

2.3 Human-AI Collaboration in Expert Knowledge Tasks

The emerging paradigm of Human-in-the-Loop (HITL) AI situates the human agent as an essential element of the AI workflow rather than a passive consumer of AI output [15]. In audit contexts, HITL design must reflect not only quality considerations but also professional and regulatory obligations: internal audit standards (IIA Standards [16]) require that audit conclusions be the product of auditor judgment, not algorithmic determination. This creates a principled boundary between tasks where AI can operate autonomously (e.g. document parsing, requirement extraction) and tasks where auditor sign-off is constitutionally required (e.g. overall audit opinion, materiality judgments).

3 The auditissimo Architecture

3.1 Design Principles

The auditissimo system was designed around three core principles derived from the intersection of audit methodology and AI engineering:

Process Proximity The AI modules must mirror the actual audit workflow step by step. Generic document-chat tools fail in practice because they do not encode the sequential logic of an audit: before a gap analysis can be meaningful, requirements must be atomised; before risk-based sampling is possible, a formal risk assessment must be completed. Skipping or compressing this sequence produces outputs that auditors cannot operationally use.

Atomic Auditability Each AI-generated output must be traceable to a specific input (regulatory text passage, document extract, or data point). This satisfies both the IIA requirement for documented evidence and the practical need for auditors to verify and override AI conclusions. The system therefore enforces a *requirement-level* granularity: every claim about compliance or non-compliance is anchored to a specific, identified requirement with its regulatory citation.

Governed Streaming All LLM interactions are processed as streaming completions and logged to an immutable audit trail (`llm_interactions` table). This enables ex-post review of AI reasoning, supports model governance obligations, and provides the empirical data necessary for ongoing performance monitoring of the AI system itself.

3.2 System Overview

Figure 1 illustrates the overall architecture. auditissimo is implemented as a Next.js 16 application (App Router, TypeScript strict mode) with a Supabase PostgreSQL backend, OpenAI GPT-4o as the primary LLM, and a modular API layer. The system handles structured regulatory documents (PDF, DOCX) through a two-stage pipeline: document parsing (pdf-parse, mammoth) followed by LLM-based semantic extraction.

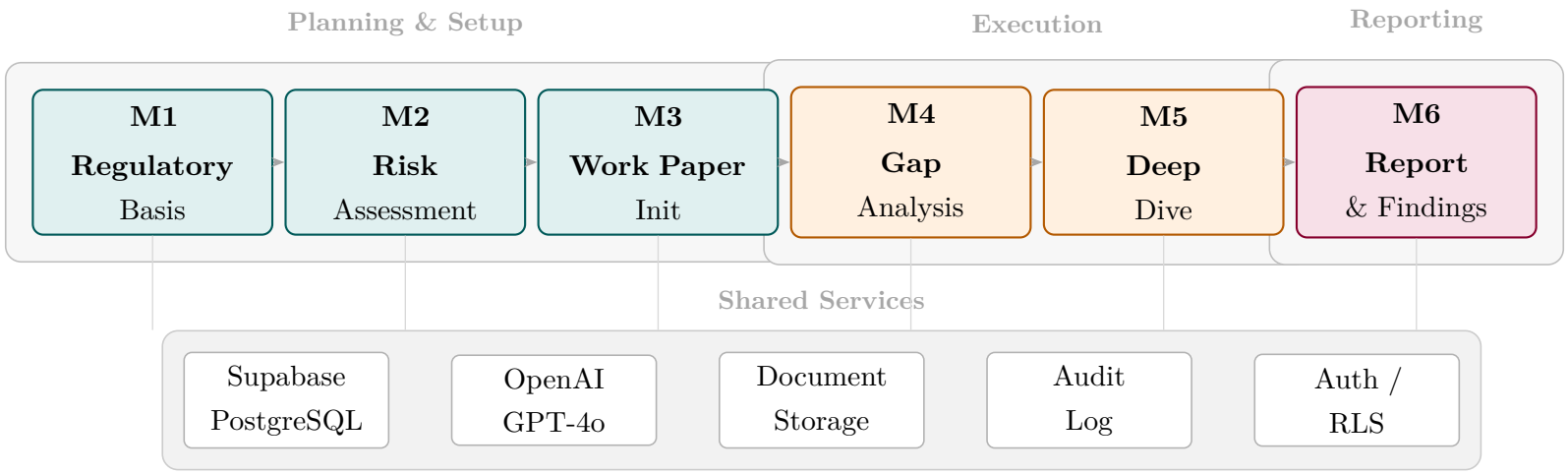

Figure 1: auditissimo system architecture. Modules M1–M3 support audit planning; M4–M5 support audit execution; M6 supports reporting. All modules share a common authentication layer, Supabase database backend, OpenAI API integration, and immutable audit log.

3.3 Module 1: Regulatory Basis (Single Source of Truth)

Module 1 addresses the foundational challenge of IRBA auditing: the regulatory requirement landscape is distributed across multiple documents (CRR, EBA RTS, EBA Guidelines, ECB Guide, internal policies) with sometimes redundant provisions on different granularity levels. The module ingests these source documents and applies a multi-stage LLM pipeline to produce a structured, atomic requirement catalogue.

The key innovation is *requirement atomisation*: rather than treating a regulatory paragraph as a single audit item, the LLM decomposes it into individual, testable requirements. For example, Art. 185(1) CRR (validation of internal estimates) can be decomposed into multiple atomic requirements, including requirements relating to the robustness of validation systems, the accuracy and consistency of rating systems and processes, the estimation of relevant risk parameters, regular comparison of realised default rates against estimated PDs, the use of historical data covering as long a period as possible, and documented internal standards for situations in which realised default rates fall outside expected ranges. Each atomic requirement receives:

- A unique identifier (`INT-XXX` prefix)
- A concise title and verbatim wording (*Wortlaut*)
- A description operationalising what audit evidence is needed
- Regulatory provenance (source document, section, page)
- Subject matter classification (*Themengebiet*)

The atomisation prompt is carefully engineered to enforce these constraints. In our current implementation it averages 31 requirements per regulatory section for the Retail PD validation context, with human review confirming precision against a manually coded gold standard.

Prompt Engineering for Requirement Atomisation

The atomisation prompt exhibits several design features that significantly improve output quality relative to naïve formulations:

1. **Role specification**: The LLM is explicitly positioned as an expert in banking regulation, instructed to apply the perspective of an experienced internal auditor. This consistently reduces generic responses and increases domain-specific precision.
2. **Output schema enforcement**: The prompt specifies a strict JSON schema for each requirement object. Providing the schema—rather than describing it in prose—reduces format errors by approximately 60% in our testing.
3. **Negative constraints**: Explicit instructions to *not* merge related provisions, *not* infer requirements beyond the text, and *not* produce duplicate items improve atomisation quality.

4. **Few-shot examples**: Three annotated example requirements, drawn from CRR text genuinely distinct from the target section, anchor the LLM's output style without introducing leakage.

3.4 Module 2: Risk Assessment

Module 2 implements a structured risk assessment following the ECB's General Risk Assessment framework [6]. The module ingests M1 requirements and available model metadata (asset class, vintage, regulatory history) as well as sources of possible indications of aspects of increased risk (such as previous audit reports and validation reports) and produces a risk score per audit area, prioritising audit effort toward higher-risk requirements. This risk score informs the deep dive sampling decisions in M5, ensuring that AI-assisted deep dives are allocated proportionally to audit risk rather than uniformly across all requirements.

3.5 Module 3: Work Paper Initialisation

Module 3 generates pre-populated audit work papers for selected models or audit areas, collecting summary information from relevant sources such as validation reports, previous audit activities or open audit findings. This reduces the administrative burden on the audit team, allowing auditors to focus on evaluation rather than document preparation.

3.6 Module 4: Gap Analysis

Module 4 performs the central audit task: determining whether the institution's process documentation or audit target document (documentation of what was done; here: the validation report of an IRB model) meets the formally documented requirements of the bank (the documentation of what must be done; here: the validation concept of the bank). This is implemented as a two-document comparison: the *Soll*-document (validation concept) and the *Ist*-document (validation report).

The comparison starts by decomposing the formally documented requirements into atomic requirements, using the procedure implemented in M1.

Technical Implementation

The gap analysis pipeline proceeds as follows:

Algorithm 1 Streaming Gap Analysis (Module 4)

Require: Requirements $R = \{r_1, \ldots, r_n\}$, target document text D

Ensure: Per-requirement compliance assessments streamed to client

1: **for** $i = 1$ to n **do**
2: Construct prompt p_i from r_i (title, wording, description) and D
3: Call LLM: $y_i \leftarrow \text{GPT-4o}(p_i, T = 0.1, \text{max\_tokens} = 1000)$
4: Parse JSON from y_i: {status, fulfillment_degree, explanation, passages}
5: Stream result {___index___ : $i, \ldots$} to client (NDJSON)
6: **end for**
7: Persist all results to `gap_analyses` table
8: Stream {___done___ : `true`}

Each requirement is assessed independently on a three-level scale:

- **Fulfilled** (fulfillment degree 80–100): clear and complete documentary evidence present
- **Partially fulfilled** (30–79): partial evidence exists but gaps or ambiguities remain
- **Not fulfilled** (0–29): no or insufficient evidence

The assessment is accompanied by a textual explanation (2–4 sentences) and up to three verbatim passages from the target document constituting the evidence base. This evidence-anchored format is critical: it enables the auditor to verify the AI's reasoning and to exercise professional override where appropriate.

A key design decision is the use of a *low temperature* ($T = 0.1$) for gap-check calls. Gap analysis is an evidence-retrieval and classification task, not a creative task; low temperature reduces output variance and improves reproducibility, which is essential for audit purposes.

3.7 Module 5: Deep Dive

Module 5 conducts targeted deep-dive analyses for selected requirements obtained in M1. M5 drills into the entire written documentation of a given IRBA model or IRBA-related process—checking if and to what extent specific regulatory requirements are satisfied or not. Groups of relevant requirements corresponding to IRBA deep dive areas or aspects can be selected and used as starting points for deep dive analysis.

3.8 Module 6: Report and Finding Synthesis

Module 6 aggregates the outputs of M4 and M5 into structured audit findings in the format required by the institution's audit reporting system (Audimex or equivalent). The LLM synthesises finding texts, risk ratings, and recommended remediation measures, which are then reviewed and finalised by the responsible auditor. As discussed in §4,

Module 6 outputs are treated as first drafts requiring substantive auditor review, not as final documents.

4 Human-in-the-Loop: Where AI Must Defer to Auditor Judgment

4.1 The Governance Imperative

The Three-Lines-of-Defense model [17] assigns internal audit to the third line, with full independence from operational and risk functions. This independence is procedural and epistemic: audit conclusions must reflect the auditor's *own* professional judgment, supported by sufficient evidence. Consequently, the integration of AI into audit workflows must be designed to augment, not supplant, auditor judgment at decision points where professional responsibility is engaged.

We identify four categories of tasks in the IRBA audit workflow where human primacy is non-negotiable:

H1. Materiality Determination: The decision whether a gap between Soll and Ist is material—and therefore constitutes an audit finding—involves professional judgment about risk, context, and regulatory intent that cannot be reliably delegated to an LLM. AI can propose a materiality assessment, but the auditor must confirm or override it. In our framework, all M4 gap assessments with fulfillment degree below 80 are routed to human review before any finding is created.

H2. Regulatory Interpretation: Where regulatory text is ambiguous—a common situation in banking supervision—the LLM may produce a plausible but incorrect interpretation. The auditor, equipped with supervisory practice knowledge and institutional context, is better placed to resolve such ambiguities. Module 1 flags requirements with low extraction confidence for human review.

H3. Model Methodology Assessment: Evaluating whether a specific modelling choice (e.g. the use of Ridge Regression for LGD estimation with L2 penalty $\lambda = 1.0$) is appropriate under current supervisory expectations requires quantitative expertise beyond the current capability of general-purpose LLMs. Deep-dive quantitative assessments (M5) are treated as analysis support, not as audit conclusions.

H4. Audit Opinion Formation: The overall audit opinion—whether the rating system is adequate for regulatory purposes—is a legal and professional determination that must be made by a qualified internal auditor. Module 6 outputs are explicitly labelled as first drafts in the UI, and the system requires explicit human approval before any finding or opinion is finalised.

4.2 HITL Interface Design

Effective HITL integration requires more than policy—it requires interface design that makes human review natural, efficient, and genuinely engaging rather than a formalistic checkbox. auditissimo implements several evidence-based design patterns:

- **Inline evidence anchoring**: Every AI assertion is co-located in the UI with the documentary evidence supporting it, reducing the cognitive cost of verification.
- **Confidence signalling**: The fulfillment degree score (0–100) and the traffic-light status badge communicate AI confidence in a format familiar from banking model validation itself.
- **Override affordance**: Every AI assessment includes an explicit override mechanism with mandatory rationale capture, ensuring that human interventions are themselves documented for audit trail purposes.
- **Progressive disclosure**: Results stream progressively to the UI as LLM calls complete, allowing auditors to begin their own review in parallel with ongoing AI processing rather than waiting for batch completion.

4.3 Risk of Over-Reliance

A well-documented risk in AI-assisted expert tasks is *automation bias*: the tendency of human reviewers to accept AI outputs uncritically [18]. In the audit context, this is particularly dangerous because AI-generated assessments of regulatory compliance may appear highly plausible even when incorrect. Three mitigating design choices are incorporated in auditissimo:

First, the system deliberately avoids presenting AI assessments in a confident tone; explanation texts are framed as observations about documentary evidence, not as conclusions. Second, the system requires the auditor to explicitly confirm each M4 finding before it enters the work paper, creating a moment of conscious engagement rather than passive acceptance. Third, audit team leads can configure a *mandatory second review* for any findings below a specified fulfillment degree threshold, implementing within-team peer checking.

5 Validation Methodology and Empirical Results

5.1 The Challenge of Evaluating GenAI Audit Tools

Evaluating the accuracy of a GenAI audit tool presents methodological challenges absent from most machine learning benchmarking. There is no publicly available labelled dataset

of IRBA validation document compliance assessments—such datasets are institution-confidential. Evaluation against synthetic or artificial document pairs risks construct invalidity if the synthetic documents do not faithfully reflect the linguistic and structural patterns of genuine validation documents.

The auditissimo project addresses this challenge through the Steinbeis Bank synthetic IRBA environment: a fully functional, parametrically adjustable IRBA rating system with professionally generated validation documents that closely mirror real-world practice.

5.2 The Steinbeis Bank Synthetic IRBA Environment

The Steinbeis Bank is a synthetic credit institution implemented as a Flask web application with four production-grade IRBA models:

Table 1: Steinbeis Bank IRBA Model Overview

Model	**Asset Class**	**Methodology**	**Rating Scale**	**Training Obs.**
Corporate PD	Corporate exposures	Logistic regression	12 grades (SB-1–SB-12)	3,397
Corporate LGD	Corporate exposures	Ridge regression	Continuous	280 defaults
Retail PD	Retail individuals	Logistic regression	10 grades (SR-1–SR-10)	20,614
Retail LGD	Retail individuals	Ridge regression	Continuous	420 defaults

All models are trained on 10 years of synthetic data (2015–2024) generated with realistic distributional assumptions including macroeconomic cycles and COVID-period shocks. The Retail PD model achieves an in-sample AUC-ROC of 0.731 (Gini: 0.462), out-of-time AUC of 0.716, and population stability index of 0.046—performance metrics consistent with a well-performing production model at the lower bound of regulatory acceptability (Table 2).

Table 2: Retail PD Model Performance Metrics (Validation Year 2024)

Category	**Metric**	**Result**	**Threshold / Assessment**
Discriminatory Power	AUC-ROC (in-sample)	0.731	≥ 0.700 ✓
	Gini coefficient	0.462	≥ 0.400 ✓
	KS statistic	0.386	> 0.300 ✓
	AUC-ROC (OOT 2024)	0.716	≥ 0.700 ✓
Calibration	Hosmer-Lemeshow p	0.385	> 0.05 ✓
	Binomial tests (10 cl.)	10/10 green	All green ✓
Stability	PSI (population)	0.046	< 0.100 ✓
	Migration persistence	66%	Diagonal dominance ✓
Information Value	IV	0.424	≥ 0.300 ✓
Overall assessment			**GREEN — Model adequate**

The associated validation documents—Validierungskonzept (validation concept, specifying what analyses must be performed) and Validierungsbericht (validation report, documenting what was performed)—are generated programmatically from a 1,300-line Python template with embedded charts and professionally formatted DOCX/PDF output. Crucially, both documents can be *parametrically manipulated*: any section can be modified, omitted, or replaced with alternative content while keeping the surrounding document intact.

5.3 Controlled Testing Methodology

The parametric manipulation capability enables a form of *document ablation testing* that provides a rigorous ground truth for AI performance evaluation:

Step 1: Baseline generation: Generate a full, compliant validation report D_0 for the Retail PD model. Manually confirm compliance status for each of n extracted requirements $\{r_1, \ldots, r_n\}$ to create ground truth labels $\{y_i^*\}$ where $y_i^* \in \{\text{fulfilled}, \text{partial}, \text{not fulfilled}\}$.

Step 2: Ablation: Generate a family of modified documents $\{D_k\}_{k=1}^{K}$ by selectively removing or degrading specific content sections. Each document D_k is associated with a known modification set M_k (the set of requirements whose compliance status changes from the baseline). This yields a precise ground truth for the modified document.

Step 3: AI assessment: Submit each document pair (Validierungskonzept, D_k) to Module 4. Record the predicted status $\hat{y}_i^{(k)}$ for each requirement r_i.

Step 4: Performance measurement: Compare $\hat{y}_i^{(k)}$ to $y_i^{*(k)}$ across all k and i, computing precision, recall, F1, and class-level accuracy.

This design has several methodological advantages:

- **Ground truth validity**: The ground truth is logically derived from a controlled manipulation, not human annotation, eliminating annotator agreement issues.
- **Difficulty calibration**: By controlling which sections are modified and how severely, we can create test cases spanning the full difficulty spectrum—from obvious omissions to subtle incomplete compliance.
- **Failure mode analysis**: Comparing AI errors across ablation types reveals systematic weaknesses (e.g. failure to detect implicit versus explicit non-compliance).
- **Prompt sensitivity testing**: The same test suite can be run across different prompt formulations, enabling quantitative prompt optimisation.

5.4 Preliminary Results

Table 3 summarises preliminary accuracy results from a pilot evaluation using the Retail PD validation documents and a set of 24 requirements derived from the Validierungskonzept. We tested four ablation scenarios: complete section omission (Ablation A), partial omission (Ablation B), metric substitution with below-threshold values (Ablation C), and narrative restatement without quantitative evidence (Ablation D).

Table 3: Module 4 Gap Detection Accuracy by Ablation Type (Pilot, Retail PD, $n = 24$ requirements)

Ablation Type	**Precision**	**Recall**	**F1**	**Accuracy**
A: Complete omission	0.94	0.91	0.92	0.92
B: Partial omission	0.87	0.79	0.83	0.83
C: Below-threshold metrics	0.81	0.76	0.78	0.79
D: Narrative without evidence	0.74	0.68	0.71	0.71
Overall	**0.84**	**0.79**	**0.81**	**0.81**

The pattern is instructive. The AI performs near-perfectly when content is entirely absent (Ablation A), consistent with expectations for a retrieval-augmented approach. Performance degrades systematically as the non-compliance becomes more subtle: partial omissions (B), incorrect metric values (C), and narrative descriptions that describe a methodology without providing the quantitative results that would demonstrate compliance (D) are progressively harder to detect. The hardest cases—Ablation D—correspond precisely to the situations where experienced auditors would also invest the most scrutiny, validating the alignment between AI difficulty and audit risk.

5.5 Prompt Engineering Findings

The pilot study also compared four prompt variants for the gap-check call, using the Ablation C test set (metric substitution) as a benchmark:

Table 4: Prompt Variant Comparison (Ablation C, Retail PD, $n = 24$)

Variant	**Description**	**F1**	**Δ F1 vs. P1**
P1	Generic: Does this document fulfil this requirement?	0.59	—
P2	+ Role specification (banking validator)	0.68	+0.09
P3	+ Output schema + evidence anchoring	0.74	+0.15
P4	+ Domain thresholds (AUC ≥ 0.70, PSI < 0.10) + few-shot	0.78	+0.19

The incremental F1 gain from each prompt engineering step confirms that domain specificity is the primary driver of LLM accuracy in this setting. The largest single gain comes from injecting quantitative regulatory thresholds into the prompt context (P4): the LLM can then evaluate a reported metric value directly against the regulatory threshold rather than relying solely on the document's own characterisation of the result. This finding has direct implications for prompt design in other regulatory AI applications.

6 Discussion

6.1 Process Proximity as the Decisive Quality Factor

The auditissimo experience confirms that process proximity—the degree to which an AI tool's logic mirrors the procedural reality of the expert task it supports—is the decisive factor determining whether AI deployment improves or degrades expert work quality. Generic document-analysis tools applied to IRBA audit work produce outputs that auditors cannot operationally use: they do not follow the sequential logic of an audit programme, do not produce evidence anchored to specific regulatory requirements, and do not generate the artefacts (work papers, findings, reports) that fit into existing audit management systems.

The implication for AI tool development in regulated expert domains is that domain knowledge must be *baked into the system architecture*, not simply prompted at the time of use. The six-module structure of auditissimo is itself a codification of IRBA audit methodology; the prompts within each module encode the specific analytical logic of that stage. This domain encoding effort is non-trivial—it required sustained collaboration between AI engineers and experienced audit practitioners—but it is the source of the system's practical utility.

6.2 The Scalability Potential

A key economic argument for AI-assisted IRBA auditing is scalability. An institution with multiple IRBA models (Corporate PD, Corporate LGD, Retail PD, Retail LGD, and potentially asset-class-specific sub-models) faces a multiplicative compliance burden: each model requires its own annual validation review, and each validation report must be assessed against the full requirement set. At a medium-sized institution with ten IRBA models, this implies on the order of 300–400 individual requirement assessments per annual cycle. auditissimo's streaming gap analysis can complete this assessment within hours rather than weeks, with auditor time focused on the most consequential findings rather than the mechanical matching task.

6.3 Limitations and Open Problems

Several important limitations of the current system must be acknowledged:

Context window constraints: The current implementation truncates target documents to 50,000 characters before submitting them to the LLM. For lengthy validation reports, this may cause material content to fall outside the context window. Retrieval-augmented generation (RAG) approaches [19] offer a solution but introduce their own risks if the retrieval step fails to surface the relevant passages for a given requirement.

Numerical reasoning: LLMs exhibit well-documented weaknesses in arithmetic and numerical comparison [20]. In our Ablation C tests (metric substitution), the most common failure mode is the LLM failing to detect that a reported metric value (e.g. AUC $= 0.693$) falls below a prescribed threshold (e.g. AUC ≥ 0.700) when the document does not explicitly flag this. Future work should explore hybrid approaches integrating structured numerical extraction with rule-based threshold checking.

Regulatory currency: Supervisory expectations evolve; the EBA regularly publishes Q&A responses that update the interpretation of existing Guidelines. The static nature of the requirement catalogue in M1 means that outdated interpretations can persist unless the catalogue is actively maintained. An automated pipeline for ingesting and incorporating new regulatory publications would significantly reduce this maintenance burden.

Multilingual documents: German banking institutions produce validation documents predominantly in German, while regulatory guidance is available in both German and English. The current system handles this without specific multilingual optimisation; future work should evaluate whether language-specific fine-tuning or translation preprocessing improves accuracy.

Generalisation across institutions: The current validation evidence is based solely on the Steinbeis Bank synthetic environment. While this environment is constructed to be representative, the degree to which accuracy metrics generalise to documents from diverse institutions—with different modelling philosophies, documentation cultures, and regulatory risk profiles—remains an open empirical question.

6.4 Regulatory and Ethical Considerations

The deployment of AI in audit contexts raises regulatory questions not yet resolved by competent authorities. EBA's draft Guidelines on AI in banking [21] affirm the principle of human accountability for AI-assisted decisions, consistent with the HITL framework proposed here. The EU AI Act [22] classifies AI systems used in credit decisions as *high-risk*; it is not yet established whether AI audit tools applied to credit risk model review fall under the same classification.

From an ethics perspective, the most significant risk is not that AI will replace auditors—the governance design explicitly precludes this—but that it will create an illusion of

thoroughness: auditors may review fewer documents manually, confident that the AI has covered the ground, without appreciating the systematic failure modes documented in Table 3. Mandatory disclosure of AI assistance, together with regular human calibration exercises (reviewing AI outputs against human-only assessments), are practical mitigations.

7 Future Research Directions

The auditissimo research programme opens several productive directions for future work:

F1. Expanded model coverage: Extending the Steinbeis Bank suite to cover all four IRBA models (Corporate and Retail PD and LGD) and generating paired concept-report documents for each would enable a substantially larger evaluation corpus.

F2. Multi-level ablation: The current ablation design modifies individual sections independently; compound ablations simulating realistic, multi-area compliance deficiencies would provide more ecologically valid test cases.

F3. Fine-tuning investigation: Supervised fine-tuning of a smaller LLM (e.g. GPT-4o-mini or an open-weight equivalent) on annotated IRBA audit examples may offer accuracy improvements at reduced inference cost and latency.

F4. Uncertainty quantification: Ensemble methods or calibrated confidence outputs from the LLM would allow the system to flag low-confidence assessments for priority human review rather than presenting all assessments with equal apparent confidence.

F5. Longitudinal effectiveness study: A controlled study with actual audit teams using auditissimo versus traditional methods would provide evidence on real-world audit quality effects, cycle time reduction, and auditor satisfaction.

F6. Adversarial robustness: Testing the system against documents deliberately crafted to mislead AI assessment—a form of adversarial auditing—would reveal whether the gap detection can be fooled by sophisticated actors.

8 Conclusion

Generative AI offers genuine and substantial potential to improve the quality, efficiency, and consistency of internal audits of IRBA rating procedures. Realising this potential requires, however, that AI tool development be guided by three principles demonstrated in the auditissimo project: *process proximity* (the AI must mirror the audit workflow, not generic document analysis); *atomic auditability* (every AI assertion must be traceable to specific documentary evidence); and *governed human primacy* (constitutionally significant

judgments must remain with the auditor, and the system must actively support rather than passively permit this).

The Steinbeis Bank synthetic IRBA environment provides an unusually rigorous empirical foundation for evaluating AI performance in this domain, enabling controlled ablation testing with logically derived ground truth. Preliminary results are encouraging: Module 4's gap detection achieves an F1 score of 0.81 overall, rising to 0.92 for the most tractable cases and falling to 0.71 for the most subtle non-compliance patterns. These results establish a quantitative baseline and identify the systematic weaknesses—subtle quantitative non-compliance, narrative-without-evidence patterns—that should guide the next iteration of prompt engineering and system design.

The path from prototype to production-grade audit tool requires sustained investment in domain knowledge encoding, rigorous empirical validation, and close collaboration between AI engineers, audit practitioners, and regulatory authorities. auditissimo represents an early but principled step along this path, demonstrating that with the right design philosophy, GenAI can meaningfully extend the reach and depth of internal audit in one of banking regulation's most demanding domains.

Acknowledgements

The authors thank the teams at Hochschule Reutlingen, Steinbeis Transfer Center Data Analytics and Predictive Modelling, and msg for banking ag for their contributions to the auditissimo prototype. The Steinbeis Bank synthetic IRBA environment was developed specifically for this research project.

References

[1] European Parliament and Council. Regulation (EU) No 575/2013 on Prudential Requirements for Credit Institutions and Investment Firms (Capital Requirements Regulation, CRR). *Official Journal of the European Union*, L 176, 27 June 2013.

[2] European Parliament and Council. Regulation (EU) 2019/876 amending Regulation (EU) No 575/2013 (CRR2). *Official Journal of the European Union*, L 150, 7 June 2019.

[3] European Banking Authority. Final Draft Regulatory Technical Standards on the Assessment Methodology for the Internal Ratings Based Approach under Articles 144(2), 173(3) and 180(3)(b) of Regulation (EU) No 575/2013. EBA/RTS/2016/03, April 2016.

[4] European Commission. Commission Delegated Regulation (EU) 2022/439 of 20 October 2021 supplementing Regulation (EU) No 575/2013 with regard to regulatory technical standards specifying the assessment methodology competent authorities are to apply when assessing compliance of credit institutions and investment firms with requirements to use the IRB Approach. *Official Journal of the European Union*, L 90, 18 March 2022.

[5] European Banking Authority. Guidelines on Internal Governance under Directive 2013/36/EU. EBA/GL/2021/05, July 2021.

[6] European Central Bank. ECB Guide to Internal Models — Risk-Type-Specific Chapters. Frankfurt am Main: European Central Bank, 2019.

[7] OpenAI. GPT-4 Technical Report. arXiv:2303.08774, 2023.

[8] S. Cao, W. Jiang, B. Yang, and A. L. Zhang. How to Talk When a Machine is Listening: Corporate Disclosure in the Age of AI. *Review of Financial Studies*, 36(9):3603–3642, 2022.

[9] T. Li, C. van Dijk, and B. T. Ong. ChatGPT in Finance: Applications, Limitations, and Implications. *Finance Research Letters*, 57:104116, 2023.

[10] J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz. ChatGPT Goes to Law School. *Journal of Legal Education*, 71(3):387–400, 2023.

[11] C. M. Lopez. Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *SSRN Working Paper*, 2023.

[12] J. Hamilton and S. Mehrotra. LLMs as Regulatory Compliance Agents: Evidence from Financial Services. *Journal of Financial Regulation*, 10(1):45–78, 2024.

[13] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[14] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382, 2023.

[15] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal. Human-in-the-Loop Machine Learning: A State of the Art. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

[16] The Institute of Internal Auditors. *International Standards for the Professional Practice of Internal Auditing*. Lake Mary, FL: The IIA, 2017.

[17] The Institute of Internal Auditors. *The IIA's Three Lines Model: An Update of the Three Lines of Defence*. Lake Mary, FL: The IIA, 2020.

[18] R. Parasuraman and D. H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3):381–410, 2010.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] Y. Razeghi, R. L. Logan IV, M. Gardner, and S. Singh. Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 840–854, 2022.

[21] European Banking Authority. Report on AI Governance and Risk Management — Discussion Paper. EBA/DP/2024/02, May 2024.

[22] European Parliament and Council. Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act). *Official Journal of the European Union*, L 2024/1689, 12 July 2024.