RETRIEVER OR REASONER? DECOMPOSING RETRIEVAL-AUGMENTED GENERATION PERFORMANCE IN EXTERNAL AUDIT SUPERVISION

2026

BANCO DE ESPAÑA
Eurosistema

Documentos Ocasionales
N.º 2613

Andrés Alonso-Robisco, José Manuel Carbó, Carlos José García, Jorge Quintana and Javier Tarancón

RETRIEVER OR REASONER? DECOMPOSING RETRIEVAL-AUGMENTED GENERATION PERFORMANCE IN EXTERNAL AUDIT SUPERVISION

RETRIEVER OR REASONER? DECOMPOSING RETRIEVAL-AUGMENTED GENERATION PERFORMANCE IN EXTERNAL AUDIT SUPERVISION (*)

Andrés Alonso-Robisco (**)

BANCO DE ESPAÑA

José Manuel Carbó (***)

BANCO DE ESPAÑA

Carlos José García (****)

BANCO DE ESPAÑA

Jorge Quintana (*****)

BANCO DE ESPAÑA

Javier Tarancón (******)

BANCO DE ESPAÑA

(*) This work was developed in the context of a Technical Support Instrument (TSI) project funded by the European Union and managed by the European Commission's Reform and Investment Task Force (SG REFORM). The authors thank Cristina Pacella, Piotr Nowak, Ana Hernáez and Ángeles Cano for their support.
(**) andres.alonso@bde.es
(***) jose.carbo@bde.es
(****) Carlosjose.garcia@bde.es
(*****) Jorge.quintana@bde.es
(******) Javier.tarancon@bde.es

Documentos Ocasionales. N.º 2613

July 2026

https://doi.org/10.53479/43986

ISSN: 1696-2230 (online edition)

Abstract

Supervisory reviews of external audit reports require structured evidence extraction and judgement under strict confidentiality and governance constraints. We evaluate retrieval-augmented generation pipelines for this task, comparing lexical, semantic, hybrid, and oracle retrieval across on-premise open-weight models (Llama 3B, Mistral 7B, Llama 70B) and proprietary cloud models (Kimi, Claude Sonnet 4.6). Using 20 bank audit reports and a standardized central bank template of 30 questions, we score operational correctness against supervisor-provided ground truth with an independent LLM-as-judge, complemented by expert back-to-back checks. The paired design holds each question report pair fixed, separating gains driven by retrieval quality from those driven by model capability, and identifying when they operate as complements rather than substitutes. Semantic retrieval yields a sizeable and statistically robust uplift within fixed models. Under symmetric strong retrieval, Llama 70B becomes statistically indistinguishable from the best cloud benchmark, while smaller on-premise models remain constrained on higher complexity judgement questions. The results point to a capacity threshold for practical substitution and provide evidence to guide deployment trade-offs in regulated settings.

Keywords: large language models, retrieval-augmented generation, financial supervision, audit.

JEL classification: M42, G28, C88.

Resumen

La revisión supervisora de los informes de auditoría externa exige extraer evidencia de forma estructurada y emitir juicios expertos, todo ello bajo estrictas exigencias de confidencialidad y gobernanza. En este documento evaluamos el uso de los sistemas de generación aumentada por recuperación para esta tarea. Es una técnica que combina un modelo de lenguaje con un módulo que recupera previamente los fragmentos relevantes de los documentos, de modo que las respuestas se generan a partir de evidencia localizada en el propio informe y no únicamente del conocimiento interno del modelo. Comparamos cuatro estrategias de recuperación —léxica, semántica, híbrida y oráculo— aplicadas tanto a los modelos de pesos abiertos desplegables en infraestructura interna (Llama 3B, Mistral 7B y Llama 70B) como a los modelos propietarios accesibles en la nube (Kimi y Claude Sonnet 4.6). El ejercicio se realiza sobre 20 informes de auditoría bancaria y un cuestionario estandarizado de un banco central con 30 preguntas. La calidad de las respuestas se evalúa frente a una referencia elaborada por supervisores, utilizando un modelo extenso de lenguaje independiente como juez y complementando esa evaluación con revisiones cruzadas por parte de los expertos. El diseño emparejado fija cada combinación de pregunta e informe, lo que permite distinguir las mejoras atribuibles a la recuperación de las que provienen de la capacidad del propio modelo, e identificar en qué casos ambas dimensiones actúan como complementarias y no como sustitutivas. Los resultados muestran que la recuperación semántica produce, manteniendo fijo el modelo, una mejora sustancial y estadísticamente robusta. Cuando todos los sistemas operan con una recuperación fuerte y simétrica, Llama 70B alcanza un rendimiento estadísticamente indistinguible del mejor modelo en la nube, mientras que los modelos abiertos de menor tamaño siguen mostrando limitaciones en las preguntas que exigen un mayor juicio interpretativo. En conjunto, los resultados apuntan a la existencia de un umbral de capacidad a partir del cual los modelos internos pueden sustituir, en la práctica, a las alternativas en la nube, y ofrecen evidencia útil para orientar las decisiones de despliegue en entornos regulados.

Palabras clave: modelos grandes de lenguaje, generación aumentada por recuperación, supervisión financiera, auditoría.

Códigos JEL: M42, G28, C88.

Contents

1 Introduction

Textual data have long played a central role in accounting, economics, and finance. Financial statements, audit reports, regulatory filings, and supervisory communications contain rich information that is only partially captured by traditional structured variables. Over the past decade, the literature on text as data has formalised methods to extract systematic evidence from such documents and has shown their value for measurement, classification, and inference in settings where structured data are incomplete or unavailable (Gentzkow et al., 2019; Ash and Hansen, 2023). These approaches are now well established and form part of the standard empirical toolkit of economists.

Recent advances in generative artificial intelligence (GenAI), and in particular large language models (LLMs), extend this paradigm. Beyond enabling quantitative text analysis, LLMs allow natural language interaction with documents, analytical pipelines, and professional workflows. In accounting and auditing, they are viewed as productivity enhancing tools that can support documentation, report drafting, and analytical tasks that are traditionally time intensive and costly (Kokina and Davenport, 2017; Austin et al., 2021; Fedyk et al., 2022). This potential is especially relevant in audit markets characterised by fee pressure, growing data volumes, and increasing documentation requirements.

Productivity gains, however, are economically meaningful only if the outputs produced by LLMs are reliable and fit for professional use. Prior research highlights several risks associated with their deployment in accounting contexts, including hallucinated content, lack of transparency, and the risk of overreliance on automated outputs (Bommasani et al., 2021; Vasarhelyi et al., 2023; Street and Wilck 2023; Huang et al. 2024). From an audit perspective, these risks relate directly to core concepts such as the sufficiency and appropriateness of audit evidence, professional scepticism, and accountability for judgement (Commerford et al., 2022; IAASB, 2009). Recent evidence from external auditing reinforces this point: auditors perceive LLMs as valuable for routine tasks, yet external general purpose systems struggle with audit specific deliverables and raise concerns around confidentiality, liability, and hallucinations (Fotoh and Mugwira, 2025). As a result, evaluating the quality of content generated by LLMs is not a secondary concern, but a prerequisite for their responsible use.

These concerns shape how LLMs are deployed in practice. Rather than relying exclusively on fully external, general-purpose models, audit firms and public institutions often favour controlled architectures that combine language models with curated internal document repositories. Retrieval-augmented generation (RAG) frameworks operationalise this approach by retrieving potentially relevant passages before generation, thereby grounding outputs in institution-specific regulations, reports, and historical documentation (Lewis et al., 2020). By construction, RAG can improve factual accuracy and reduce hallucinations, while allowing institutions to flexibly combine retrievers and generators depending on task complexity and governance constraints. This modularity supports confidentiality, reduces dependence on external providers, and facilitates auditable deployments (Eilifsen et al., 2020). From a technical perspective, it builds on advances in domain-specific representations tailored to financial and accounting language (Devlin et al., 2019; Yang et al., 2020).

While RAG architectures address key institutional and governance constraints, they introduce a distinct challenge for evaluation. In these systems, output quality depends on two separable components: the ability of the retriever to identify and supply relevant information, and the ability of the language model to reason over that information and generate coherent, complete outputs. In professional accounting and audit settings, failures in retrieval and failures in reasoning have different implications for reliability, accountability, and professional

judgement. Moreover, the relative importance of these components may vary across tasks of different complexity: routine supervisory queries may depend primarily on access to the correct contextual information, whereas analytically demanding questions may require advanced reasoning capabilities even when relevant information is available. Disentangling these margins is therefore essential for assessing the feasibility of internally deployed systems.

Most empirical work on LLM applications in accounting and auditing evaluates end-to-end output quality (Gu et al., 2024) or provides conceptual assessments of opportunities and risks (Vasarhelyi et al., 2023; Fotoh and Mugwira, 2025). Yet evidence that decomposes both margins in realistic audit and supervisory tasks remains limited. By contrast, the RAG evaluation literature more explicitly diagnoses retrieval and generation components and develops metrics tailored to the modular structure of RAG systems (Ru et al., 2024; Krishna et al., 2025; Es et al., 2023). Prior studies often rely on expert judgement by experienced auditors to assess accuracy or completeness, an approach that provides valuable insights but is costly and difficult to scale (Gu et al., 2024). As LLM-based systems become more deeply integrated into audit and supervisory workflows, there is a growing need for evaluation frameworks that are both systematic and scalable, while remaining anchored in professional standards of judgement.

Against these backdrops, this paper advances the accounting and auditing literature along two dimensions, and it does so in a setting that is directly relevant for supervisory technology (SupTech) in central banks, where innovative analytical tools are applied to supervision under strict governance constraints (Castri et al., 2019).[1] First, we implement an experimental design that explicitly separates retrieval and reasoning margins in a supervisory RAG pipeline applied to an auditing and supervisory setting. We benchmark multiple retrieval strategies (lexical BM25, embedding-based Semantic retrieval, hybrid retrieval, and an oracle retriever that supplies the relevant context by construction) across a range of generation models spanning on-premise open-weight systems (Llama 3B, Mistral 7B, Llama 70B) and proprietary cloud models (Kimi, Claude Sonnet). The design identifies the retrieval contribution by holding the model fixed and varying retrieval, and it isolates the model contribution by holding retrieval fixed and varying the model under symmetric retrieval. This decomposition clarifies when improvements in retrieval can narrow performance gaps and when model capability remains binding.

Our results point to a capacity threshold for practical substitution. Within-model comparisons show that moving from BM25 to Semantic retrieval increases accuracy by about 6.2–6.3 percentage points for both Kimi and Llama 70B. Under symmetric Semantic retrieval, Kimi substantially outperforms smaller on-premise models (Llama 3B and Mistral 7B), while Llama 70B becomes statistically indistinguishable from Kimi in overall accuracy. We further document that prompt shifts can materially affect evaluation outcomes, and that this sensitivity is strongly model dependent: under a prompt challenge that reduces prompt tailoring, Kimi exhibits a small performance drop, whereas several on-premise configurations experience substantially larger degradations.

Second, we introduce a scalable evaluation strategy that uses an independent LLM as a judge to score operational correctness against a ground truth provided by supervisors, complemented by back-to-back evaluations from human experts as a robustness check. To support inference in a setting with many ties, we complement

1 An external evaluation by leading international SupTech experts, followed by a formal action plan (Packard and Prenio, 2023; Banco de España, 2024), shaped the SupTech agenda at the Banco de España. In response to these recommendations, in 2025 the Banco de España, in collaboration with the European Commission, delivered an advanced training programme in generative artificial intelligence that combined capacity building with the development of proof-of-concept tools, including the system evaluated in this paper (see also Puig and Tarancón, 2026).

aggregated accuracy with paired question-level comparisons and exact tests that focus on discordant outcomes. This design enables systematic measurement at scale while maintaining a clear link to established standards of professional judgement in accounting and auditing.

Overall, by providing a structured decomposition of retrieval and reasoning effects across tasks of varying complexity, our analysis offers guidance not only on model performance, but on the strategic design of artificial intelligence systems under institutional and regulatory constraints. This contribution is also directly aligned with the SupTech vision of transforming supervisory data into actionable knowledge that helps supervisors better serve citizens (Banco de España, 2025).[2]

The remainder of the paper is organised as follows. Section 2 describes the institutional setting and the supervisory task. Section 3 presents the experimental design and evaluation framework. Section 4 reports the main results. Section 5 discusses paired comparisons and robustness analyses. Section 6 concludes.

2 Supervisors draw on confidential microdata (such as inspection reports), large structured registries (such as central credit registers, including AnaCredit), and unstructured sources (such as news, audit documentation, and other narrative material). A key opportunity lies in linking unstructured signals to structured supervisory information; for example, Alonso-Robisco et al. (2025) develop early warning indicators for firms using news text. This mix of sensitive sources and accountability requirements makes confidentiality, traceability, and deployment governance first-order constraints when designing and evaluating RAG systems for supervision.

2 Institutional setting: supervisory inspection of external audit reports

External audit reports play a central role in the supervision of banks, also subject to specific requirements that apply to public-interest entities in the European Union (European Parliament and the Council of the European Union 2014). Audits of financial institutions are typically conducted on an annual basis and cover a broad set of financial statements, internal controls, and risk disclosures. Given the size and complexity of banking organisations, audit reports are lengthy documents that integrate accounting judgements, risk assessments, and compliance with regulatory standards. Supervisors rely on these reports as a key input to assess the quality of financial reporting, the adequacy of internal controls, and the overall reliability of information used for prudential oversight (DeFond and Zhang, 2014; Eilifsen et al., 2020).

In banking supervision, the audit report is not treated as a binary certification, but as a source of detailed evidence. Central banks and supervisory authorities routinely perform structured inspections of audit reports to evaluate whether the external audit has been conducted with sufficient depth, independence, and professional scepticism. This process is particularly important in the banking sector, where weak audit quality can amplify information asymmetries and undermine market discipline (Nier and Baumann, 2006). As a result, supervisory authorities complement the statutory audit opinion with their own internal assessments of audit quality, using standardised procedures that can be applied consistently across institutions and over time. This use of external audit outputs as supervisory inputs aligns with the broader expectation that supervisors rely on multiple sources of information and verification to assess prudential soundness (Basel Committee on Banking Supervision, 2024).

These supervisory inspections are typically operationalised through evaluation templates. Supervisors are required to answer a series of questions based on the content of the external audit report and related documentation. The questions explicitly cover Key Audit Matters, consistent with the international auditing standard that formalised their communication in the auditor's report (IAASB, 2015). All of them comprise factual elements, such as the identification of the audit firm or the signing auditor, as well as more substantive issues related to the scope of the audit, the treatment of key risks, and the clarity of judgements disclosed in the report. The objective is not only to verify compliance, but also to assess whether the audit provides sufficient and appropriate evidence to support supervisory conclusions. This setting is well suited for evaluating retrieval-augmented generation (RAG) systems because it combines high document length, repeated query structures, and professionally meaningful ground truth.

This supervisory workflow operates under strict confidentiality and governance constraints that shape feasible deployment choices. The General Data Protection Regulation imposes requirements on data minimization, purpose limitation, and cross-border data transfers (Regulation (EU) 2016/679 of the European Parliament and of the Council). In Europe, the European AI Act introduces obligations concerning risk management, transparency, documentation, and human oversight for high-risk Artificial Intelligence (AI) systems (European Parliament and the Council of the European Union, 2024). In the United States, federal initiatives such as the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence and the U.S. National Institute of Standards and Technology (NIST) AI Risk Management Framework emphasise governance controls, documentation standards, and accountability (The White House 2023; National Institute of Standards and Technology 2023). At the international level, the OECD AI Principles provide a shared reference for responsible AI governance (Organisation for Economic Co-operation and Development, 2019).

Table 1

Full set of questions used by the central bank to inspect external audit reports of banks (a)

ID (b)	Question (template wording)	Complexity (c)
1	Audit firm signing the individual financial statements audit report	1
1.a	Conditional question. If the answer to Question 1 is "Other": identify the audit firm	2
2	Signing auditor of the individual financial statements audit report: registration number (ROAC)	1
3	Signing auditor of the individual financial statements audit report: full name	1
4	Whether the audit was performed under a joint audit arrangement	1
4.1	Conditional question. If joint audit applies: identify the co-auditing firm	2
4.1.a	Conditional question. If the answer to Question 4.1 is "Other": identify the co-auditing firm signing the individual financial statements audit report	2
4.2	Conditional question. If joint audit applies: co-auditor registration number (ROAC)	2
4.3	Conditional question. If joint audit applies: co-auditor full name	2
6	Whether the report refers to individual or consolidated financial statements	1
7	Date of the audit report	1
12	Number of Key Audit Matters disclosed in the audit report	2
13.a	Key Audit Matter: Impairment of financial assets	3
13.b	Key Audit Matter: Provisions and contingent liabilities	3
13.c	Key Audit Matter: Credit risk	3
13.d	Key Audit Matter: Revenue recognition	3
13.e	Key Audit Matter: Fair value measurement	3
13.f	Key Audit Matter: Impairment of goodwill	3
13.g	Key Audit Matter: Related party transactions	3
13.h	Key Audit Matter: Actuarial liabilities	3
13.i	Key Audit Matter: Classification of financial instruments	3
13.j	Key Audit Matter: Internal control environment and information systems	4
13.k	Key Audit Matter: Regulatory requirements and capital adequacy	4
13.l	Key Audit Matter: Recoverability of deferred tax assets	3
13.m	Key Audit Matter: Recoverability of tax credits	3
13.n	Key Audit Matter: Hedge accounting	3
13.o	Key Audit Matter: Expected Credit Loss (ECL) estimation	4
13.p	Key Audit Matter: Going concern assessment	4
13.q	Other Key Audit Matters	4
13.q.a	Conditional question. If other matters apply: specify the issue	4

SOURCE: Banco de España supervisory inspection template for external audit reports.

a Question complexity classification.
b The ID column reproduces the original numbering of the supervisory template; sub-IDs (1.a, 4.1, etc.) are conditional follow-up questions activated by the answer to the parent question.
c Complexity follows the classification reported in Table 2.

The template used in this study reflects actual supervisory practice. From an analytical perspective, the questions posed in supervisory inspection templates vary substantially in their level of complexity. Table 1 categorises these questions into four levels according to the cognitive demands they impose on a language model. Level 1 questions are objective and factual, typically requiring a binary or highly constrained answer, such as yes or no, or the extraction of a specific identifier. Level 2 questions require retrieval of relevant information and correct aggregation, but involve limited interpretation. Level 3 questions require contextual understanding and the integration of multiple pieces of information to form a coherent assessment. Finally, Level 4 questions are judgemental, context dependent, and open ended, requiring an evaluation of audit quality that closely mirrors professional supervisory judgement.

Table 2

Complexity levels of supervisory questions and corresponding cognitive demands on language models

Level	Description
Level 1	Objective and factual questions that require direct information extraction from the audit report. Answers are binary or highly constrained, such as yes or no, dates, names, or identifiers.
Level 2	Questions that require information extraction combined with minor reasoning or conditional logic, such as resolving dependencies across questions or aggregating clearly defined information.
Level 3	Contextual questions that require soft reasoning and synthesis across multiple sections of the audit report. Answers involve interpreting disclosures and integrating information, but do not require fully open-ended professional judgement.
Level 4	Judgemental and context-dependent questions that closely mirror professional supervisory or audit judgement. Answering these questions requires evaluating relevance, completeness, or quality of disclosures and applying supervisory criteria rather than extracting explicitly stated facts.

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

This classification is directly relevant for the evaluation of RAG systems in supervisory settings and is summarised in Table 2. Lower complexity questions place relatively more weight on the retrieval component because the relevant information is typically explicit in the report and errors often reflect missing or incorrect access to the appropriate section. Higher complexity questions place greater weight on model capability because they require interpretation, synthesis, and the application of supervisory criteria, even when relevant information is available. Distinguishing between these types of questions allows us to characterise where automated systems support professional judgement and where their limitations remain most pronounced.

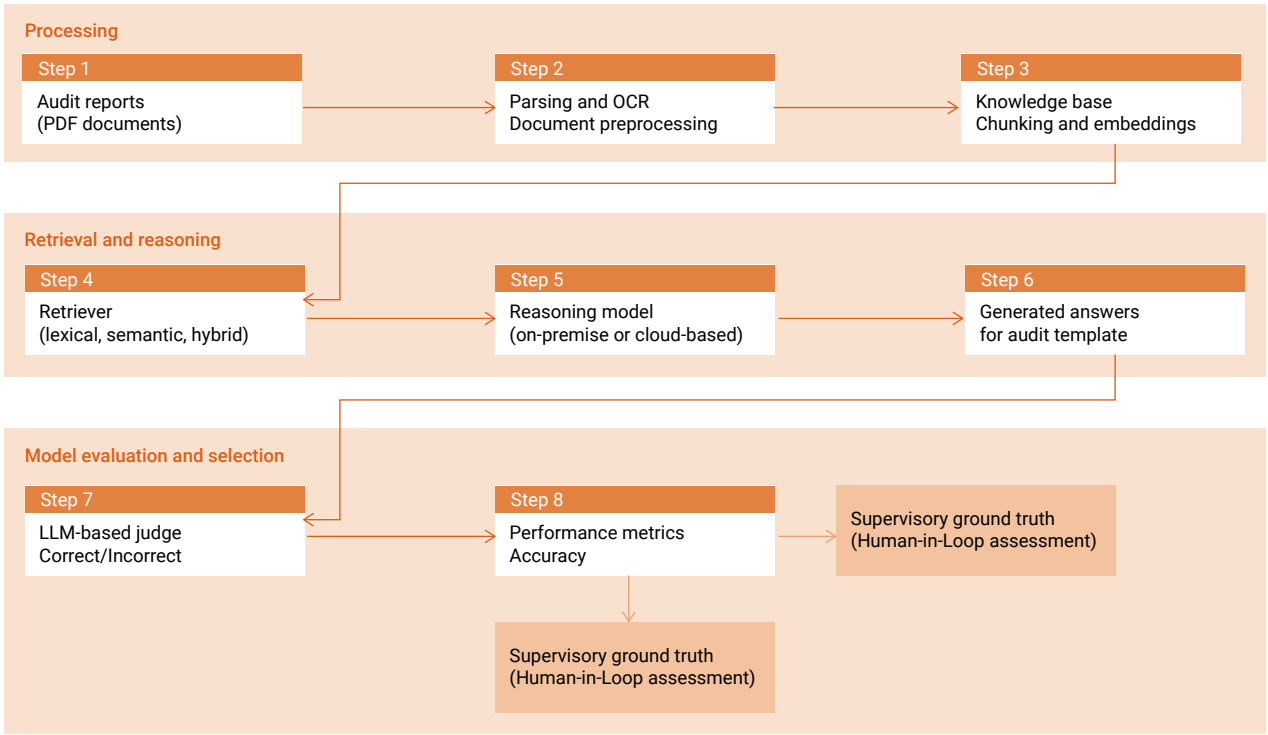
3 Experimental design and evaluation framework

The experimental setting is grounded in a realistic supervisory use case: the inspection of external audit reports of banks by a central bank. The dataset consists of publicly available audit reports corresponding to 20 supervised entities. These reports are lengthy and heterogeneous documents that integrate financial statements, audit opinions, and detailed disclosures on key audit matters. For each report, experienced financial supervisors complete a standardised inspection template comprising 30 questions (see Table 1) that cover factual information, audit scope, and the treatment of material risks. These completed templates constitute the supervisory ground truth used throughout the analysis and reflect expert professional judgement.

The objective of the study is not to train or fine-tune language models, but to evaluate the operational performance of retrieval-augmented generation (RAG) systems in supporting this supervisory task. The ground truth provided by supervisors is therefore used exclusively as a reference to assess whether automated outputs are correct, rather than as training data. This design mirrors a practical deployment scenario in which a system is expected to assist supervisors by pre-filling inspection templates, while final responsibility and judgement remain with human experts. Figure 1 provides an overview of the end-to-end retrieval, generation, and evaluation workflow underlying the experimental design.

We then proceed to evaluate a standard RAG pipeline that processes audit reports and generates answers to supervisory questions. Audit reports are first parsed and converted into text, addressing the heterogeneity of PDF formats that include scanned pages, tables, and mixed layouts. Document preprocessing refers to this conversion

Figure 1
RAG and evaluation workflow for supervisory inspection of external audit reports



SOURCE: Banco de España.

Table 3
Retrieval strategies considered in the RAG pipeline

Retrieval method	Description
BM25 (lexical)	Sparse lexical retrieval based on term frequency and inverse document frequency. It retrieves document chunks that share exact or near-exact tokens with the query and serves as a strong baseline for keyword-driven information needs, though it can miss relevant passages when terminology differs or synonyms are used.
Semantic (dense)	Embedding-based retrieval using dense vector representations of queries and document chunks. It captures semantic similarity beyond exact keyword overlap and is particularly suited for paraphrased or conceptually related queries, though it may miss exact or rare term matches.
Hybrid (BM25 + Semantic)	Combination of lexical and dense retrieval scores, typically through linear weighting or rank fusion (Cormack et al. 2009). This approach balances precision on exact terms with Semantic recall and is often more robust across heterogeneous query types.
Oracle retrieval	An upper-bound benchmark used for comparison, in which the retriever is forced to return the ground-truth relevant chunk(s) when available. By eliminating retrieval noise, it isolates generation errors and serves as a reference for the best performance a RAG system can achieve.

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

of heterogeneous PDF audit reports into machine-readable text, including OCR for scanned pages, layout cleaning, segmentation, and normalisation before chunking. The processed text is segmented into chunks and indexed in a vector database. For each question, a retriever selects relevant text segments that are provided as context to the generation model. In all baseline configurations, the retriever returns the top k ranked chunks, which we pass verbatim to the generator as contextual input. In our setting, chunks correspond to segmented portions of the audit report produced during preprocessing, typically at paragraph level. We set k = 5 throughout as a balanced choice that increases contextual coverage while limiting noise and context-length saturation. For the Oracle condition, we provide the supervisor-identified relevant paragraph by construction, which we denote as k = –1. We consider multiple retrieval strategies, including lexical retrieval based on term frequency (Robertson and Zaragoza, 2009), Semantic retrieval using contextual embeddings (Karpukhin et al., 2020; Reimers and Gurevych, 2019), hybrid approaches that combine lexical and Semantic signals (Wang et al., 2021), and an oracle retriever that supplies the relevant context by construction. Table 3 summarises the main characteristics and intended role of each retrieval strategy in the evaluation.

On the generation side, we evaluate different language models that span a range of deployment and reasoning capabilities, from smaller, on-premise models designed to be cost-effective and privacy preserving, to larger, cloud-based models with more advanced reasoning capacity. Prompt templates are tailored to each supervisory question and include role definition, task instructions, valid output formats, and explicit constraints on the use of retrieved context. This design reflects the need for precise and auditable outputs in professional accounting and supervisory environments (Eilifsen et al., 2020; Vasarhelyi et al., 2023). Table 4 summarises the main characteristics of the generation models considered in the evaluation.

For each audit report and each question in the supervisory template, the RAG system generates an answer conditioned on the retrieved context. All generation models are queried using a common instruction template that defines the professional role of the model, constrains the admissible output space, and enforces exclusive reliance on the retrieved evidence. The prompt frames the model as an auditor at the Banco de España and requires answers

Table 4

Generation models considered in the RAG pipeline

Generation model	Model(s) used	Description
On-premise open-weight LLM	Llama 3B; Mistral 7B; Llama 70B	Deployed in a controlled on-premise environment. This setup supports strong data governance, auditability, and reproducibility, which are critical in supervisory contexts, at the cost of higher infrastructure and maintenance requirements (Khatri and Brown 2010).
Cloud-based proprietary LLM	Kimi (K2 Instruct); Claude Sonnet 4.6	Commercial accessed through cloud APIs. These models typically offer stronger general-purpose performance and advanced reasoning capabilities, but involve external data processing and raise additional considerations regarding confidentiality, data protection, and regulatory compliance (European Parliament and the Council of the European Union 2016; European Banking Authority 2019).

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

to be clear, concise, and grounded in current regulation (see Figure 2) (Banco de España, 2017). The trace includes the supporting excerpt from the document together with page number and block identifiers, thereby enabling auditability and ex post verification of the reasoning process.

We deliberately hold the main prompt template fixed across models and retrieval strategies in the benchmark to reduce confounding from prompt tuning and to reflect an institutional constraint: supervisory deployments typically rely on stable, compliance oriented instruction templates rather than model specific prompt optimisation.

Figure 2

Prompt template used for all generation models in the supervisory RAG pipeline. The prompt enforces role definition, constrained outputs, exclusive reliance on retrieved context, and structured traceability requirements

Prompt used for all generation models

You are an auditor at the Banco de España, expert in report drafting and audit analysis.

The question to be answered is: ["question", as provided in the input JSON].

Answers to the audit questions must be clear and concise, based on the definitions of the concepts in Annex 9 of Circular 4, published in the Spanish Official Gazette (BOE), No. 296, Wednesday 6 December 2017, Section I, p. 119987.

If no answer is available, respond with "No Information".

Respond ONLY using the information contained in the provided context.

The valid values for the answer are: [valid values].

The context is: [content].

Format the output as valid JSON, escaping double quotation marks when necessary, with two fields:

— the answer under the key "answer"
— the traceability under the key "trace"

The trace must not exceed 500 characters and must include metadata containing the original text that justifies the answer, the page number, and the specific block from which the text is extracted.

The response must contain ONLY parseable JSON using json.loads() in Python.

Example of a valid response:
{"answer": "Yes", "trace": "The answer is yes because the document states that a key audit matter exists."}

SOURCE: Banco de España.

Figure 3

Prompt used to instruct Kimi as an independent LLM-based judge for binary evaluation of generated answers against supervisory ground truth

Prompt used for Kimi as LLM-based judge

You are a senior AI quality assurance evaluator. Your task is to assess whether the model has answered the question correctly.

You must return 0 if the answer is incorrect and 1 if the answer is correct.

Expected answer:
[ground truth]

Model-provided answer:
[model truth]

When evaluating the answer, take into account the following rules:

1) *Punctuation such as '.' and ',' is not important.*
2) *Ignore differences between uppercase and lowercase letters.*
3) *Singular and plural forms are considered equivalent.*
   For example, if the provided answer is "Individuals" and the expected answer is "Individual", the answer is correct and the output must be 1.
4) *If the expected answer is "No" and the provided answer is "No", the answer is correct and the output must be 1.*
5) *If the expected answer is "Not applicable" and the provided answer is "Not applicable", the answer is correct and the output must be 1.*
6) *Ignore minor errors such as a single incorrect letter or number when the answer consists of more than one word.*

The answer is valid EXCLUSIVELY if it is either 0 or 1.

SOURCE: Banco de España.

This choice should be interpreted accordingly. A single prompt may interact differently with heterogeneous models, particularly smaller on-premise models, which can be more sensitive to formatting requirements and instruction granularity. Therefore, the reported performance levels do not necessarily represent each model's best-case outcome. To quantify this sensitivity, Appendix A reports a prompt-shift robustness exercise that evaluates how accuracy changes when we reduce prompt tailoring while keeping the retrieval configuration fixed. Appendix B reproduces the full New prompt.

The generated answer is then evaluated against the supervisory ground truth using an independent large language model (LLM) acting as a judge. In our setup, we use Kimi as the judging model. As shown in Figure 3, the judge assesses whether the generated answer is correct with respect to the expert provided reference and assigns a binary label indicating correctness or incorrectness. The supervisory ground truth is used as an external professional benchmark for the task, based on assessments provided by central-bank stakeholders involved in the project.[3] This approach follows recent work that uses language models to evaluate model outputs at scale when direct human evaluation is costly or impractical, while remaining anchored in a reference defined by experts (Gu et al., 2024; Zheng et al., 2023; Liu et al., 2023).

Importantly, evaluation is based on operational correctness rather than textual similarity. Answers are not compared at the token or string level, but assessed according to whether they convey the correct information required by the supervisory template. This is particularly relevant in accounting settings, where the same information can be

3 These answers are used only for evaluation, not for model training, fine-tuning, or prompt optimisation. This separation limits the risk that the benchmark reflects training-contamination bias rather than genuine task performance.

expressed using different wording without affecting its substantive meaning. By reducing each question to a binary correctness outcome, the evaluation framework allows results to be aggregated across questions, reports, retrieval strategies, and generation models.

Therefore, model performance is summarised using standard classification metrics computed over binary correctness labels. Accuracy is reported as the primary metric and captures the proportion of supervisory questions that are correctly answered. Performance is analysed separately for each combination of generation model, retrieval strategy, and retrieval depth, and further examined by question complexity, as defined in Table 2. This enables a structured decomposition of retrieval and model contributions under comparable conditions, although retrieval and generation can interact in practice. To support inference in a setting with many tied outcomes, we complement aggregated accuracy with paired question-level comparisons and exact tests that focus on discordant pairs.

Overall, the evaluation framework is designed to assess whether RAG systems can reliably support supervisory inspection of audit reports under realistic institutional constraints. Rather than focusing on learning or generalisation in a statistical sense, the analysis measures whether such systems can automate routine components of the inspection process with sufficient accuracy to generate meaningful productivity gains for supervisors, while preserving the role of expert judgement in the final evaluation.

4 Aggregated results

We focus on two central questions. First, what level of accuracy can be achieved, on average, when answering supervisory questions based on external audit reports? Because average accuracy can mask errors in specific inspections or question types, Section 5 complements aggregate results with complexity-level patterns and paired comparisons of discordant outcomes. Second, how do retrieval quality and model capacity jointly shape performance in a supervisory retrieval-augmented generation pipeline?

Performance is evaluated using accuracy, defined as the proportion of supervisory questions that are correctly answered according to the LLM-based judge. Results are aggregated across audit reports and questions, and reported by generation model and retrieval strategy for a comparable setting with retrieval depth k = 5 (as defined in Section 3).[4] Table 5 summarises the main results for standard retrieval strategies.

Table 5 points to two descriptive patterns. First, retrieval quality materially affects average performance. Across models, Semantic retrieval improves accuracy relative to lexical BM25, with particularly large gains for stronger models. Second, under strong retrieval (Semantic), model capacity remains a key determinant of performance. Kimi achieves the highest aggregated accuracy (0.873), closely followed by Llama 70B (0.870). Smaller on-premise models remain materially below this level even with Semantic retrieval (0.797 for Mistral 7B and 0.753 for Llama 3B). These gaps are economically meaningful given the operational objective of pre-filling supervisory templates with minimal expert correction.

To clarify where these differences arise, charts 1–5 report accuracy by retrieval strategy and question type. Question types correspond to the four complexity levels introduced in Section 2. Three features stand out. First, Types 1–2 exhibit high accuracy for most models even under BM25, indicating that a large share of routine factual extraction can be automated reliably when the relevant information is explicit in the report. Second, performance drops sharply for Type 4 under non-oracle retrieval, and the magnitude of this drop varies substantially across models. This pattern is consistent with Type 4 questions requiring interpretation and supervisory judgement, which

Table 5
Aggregated accuracy by generation model and retrieval strategy (a) (b)

Generation model	BM25	Semantic	Hybrid	Oracle
Llama 3B	0.733	0.753	0.725	0.892
Mistral 7B	0.742	0.797	0.768	0.940
Llama 70B	0.807	0.870	0.852	0.963
Kimi	0.812	0.873	0.870	0.945
Claude Sonnet 4.6	0.752	0.795	0.783	0.950

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

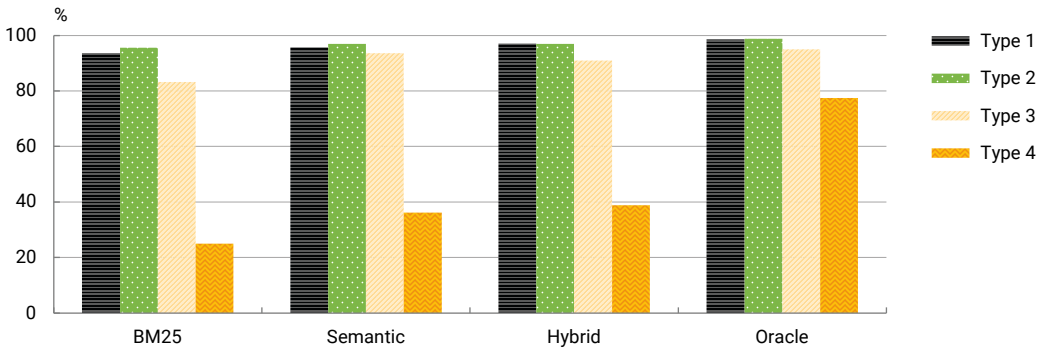
a Baseline retrieval uses k = 5 chunks; Oracle uses the gold paragraph (k = -1).
b Accuracy is computed against supervisory ground truth using an independent LLM-based judge (Kimi).

4 We focus on k=5 as a balanced configuration that provides contextual coverage while limiting noise and context-length saturation.

Chart 1

Accuracy by retrieval strategy and question type (a) (b) (c)

Model: Kimi



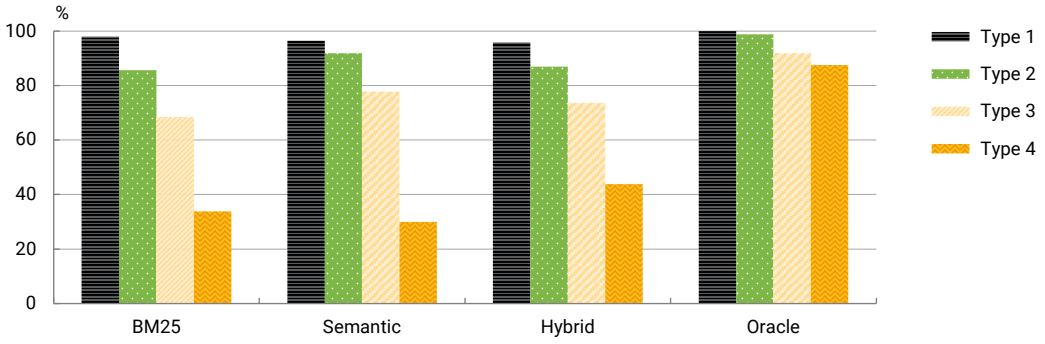
SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

a Accuracy is the share of supervisory questions answered correctly according to the LLM-based judge.
b k denotes the number of retrieved chunks at generation time; ORACLE uses the gold paragraph (k = −1).
c Question types correspond to the four complexity levels defined in Section 2.

Chart 2

Accuracy by retrieval strategy and question type (a) (b) (c)

Model: Claude Sonnet



SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

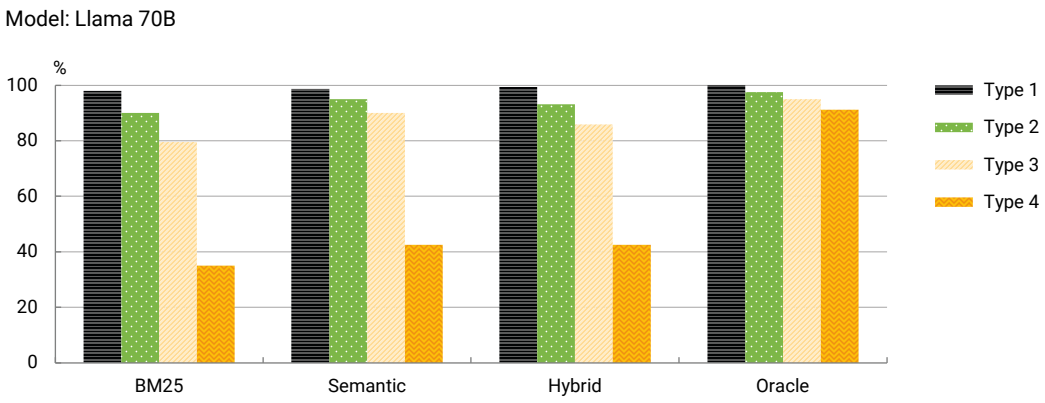
a Accuracy is the share of supervisory questions answered correctly according to the LLM-based judge.
b k denotes the number of retrieved chunks at generation time; ORACLE uses the gold paragraph (k = −1).
c Question types correspond to the four complexity levels defined in Section 2.

places greater weight on model capability even when retrieval provides relevant context. Third, the relative benefit of retrieval improvements is concentrated in Types 3–4: Semantic retrieval raises performance compared to BM25, while hybrid retrieval is not systematically superior to Semantic retrieval in this setting and can be slightly worse for some models.

The charts also report an oracle benchmark that supplies the gold paragraph by construction. Oracle retrieval compresses performance differences for low-complexity questions and substantially lifts Type 4 accuracy, yet it does not eliminate cross-model gaps. This suggests that both retrieval quality and model capability contribute

Chart 3

Accuracy by retrieval strategy and question type (a) (b) (c)

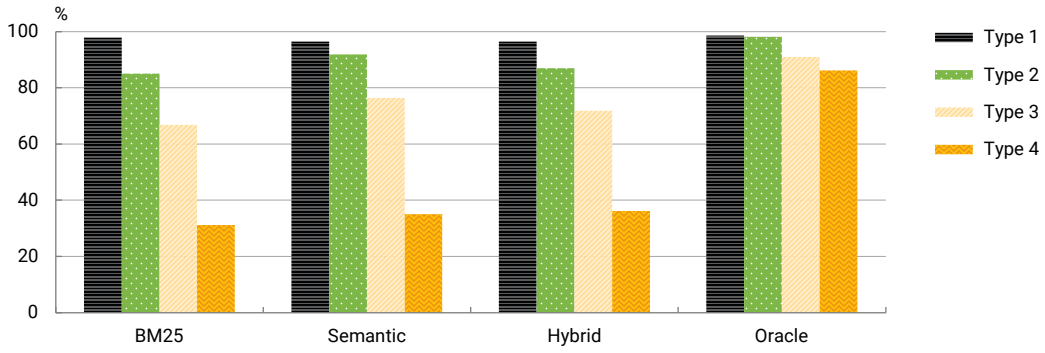


SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

a Accuracy is the share of supervisory questions answered correctly according to the LLM-based judge.
b k denotes the number of retrieved chunks at generation time; ORACLE uses the gold paragraph (k = −1).
c Question types correspond to the four complexity levels defined in Section 2.

Chart 4

Accuracy by retrieval strategy and question type (a) (b) (c)



SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

a Accuracy is the share of supervisory questions answered correctly according to the LLM-based judge.
b k denotes the number of retrieved chunks at generation time; ORACLE uses the gold paragraph (k = −1).
c Question types correspond to the four complexity levels defined in Section 2.

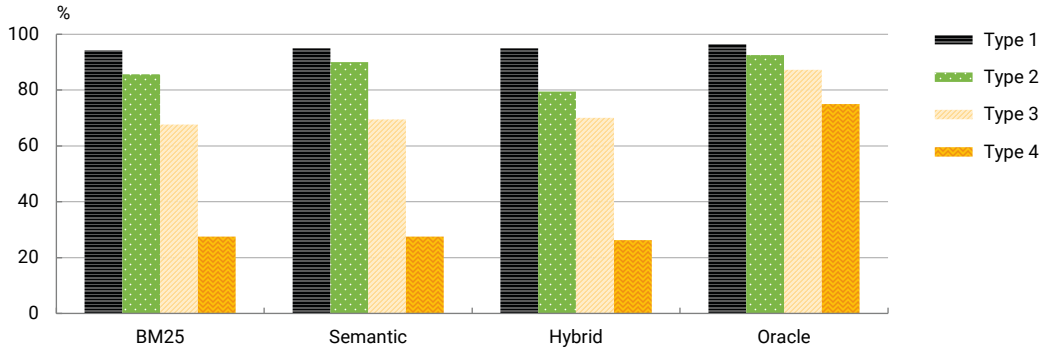
to errors, and it motivates the paired comparisons in Section 5, where we isolate (i) retrieval uplift holding the model fixed and (ii) model differences holding retrieval fixed, using exact tests on discordant pairs.

Taken together, aggregated accuracy and type-level patterns suggest that strong retrieval is necessary but not sufficient for high performance in supervisory inspection tasks. The next section examines these mechanisms through paired question-level comparisons and robustness analyses, including prompt robustness under a prompt shift.

Chart 5

Accuracy by retrieval strategy and question type (a) (b) (c)

Model: Llama 3B



SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

a Accuracy is the share of supervisory questions answered correctly according to the LLM-based judge.
b k denotes the number of retrieved chunks at generation time; ORACLE uses the gold paragraph (k = −1).
c Question types correspond to the four complexity levels defined in Section 2.

5 Robustness and deployment implications

The robustness analysis is designed to check whether the main policy conclusions survive more demanding comparisons at the question-report level. The core idea is simple: each comparison holds the underlying supervisory case fixed and asks whether performance changes when we vary the retriever, the model, or the prompt. This design is useful for deployment decisions because it distinguishes three operational margins: better evidence access, stronger reasoning capacity, and sensitivity to institutional prompt design. Appendix A reports the exact paired-test formulation and the full set of tables.

Three results are most relevant from a policy-oriented standpoint. First, retrieval quality is not a secondary engineering detail. Holding the model fixed, switching from BM25 to Semantic retrieval increases accuracy by about 6.2–6.3 percentage points for both Kimi and Llama 70B, and the gains are concentrated in discordant cases where Semantic retrieval is correct and BM25 is not. Second, model capacity still matters once retrieval is aligned: under the same Semantic retriever, Llama 70B is statistically indistinguishable from Kimi, whereas Mistral 7B and Llama 3B remain clearly weaker. Third, the practical substitution margin is present only for the large on-premise model in this benchmark: 70B+Semantic beats Kimi+BM25 in discordant pairs, while the smaller on-premise models do not deliver the same margin. Table 6 summarises this condensed robustness evidence.

This condensed evidence supports a threshold view of deployment. Retrieval improvements generate systematic gains, but they are complements rather than full substitutes for reasoning capacity. In the present benchmark, a large on-premise model with strong retrieval can match or exceed a cloud configuration with weaker retrieval, which is directly relevant for institutions facing confidentiality, auditability, and vendor-dependence constraints. By contrast, smaller on-premise models remain more exposed on higher-complexity questions and under prompt changes. The policy implication is therefore not that one architecture dominates universally, but that deployment should be matched to the complexity of the supervisory task, the quality of evidence retrieval, and the acceptable level of human review.

Table 6
Condensed robustness evidence for deployment interpretation

Margin tested	Key evidence	Operational interpretation
Retrieval quality	Semantic retrieval raises accuracy by 6.2–6.3 percentage points for Kimi and Llama 70B relative to BM25.	Evidence access is a first-order design lever for supervisory RAG systems, especially for template pre-filling and factual extraction.
Model capacity	Under Semantic retrieval, Llama 70B and Kimi are statistically indistinguishable, while 7B and 3B models remain below Kimi.	Strong retrieval narrows gaps, but does not fully offset limited reasoning capacity on more demanding supervisory questions.
Substitution margin	70B+Semantic beats Kimi+BM25 in discordant pairs; the same pattern does not generalise to smaller on-premise models.	On-premise substitution is plausible only when retrieval quality and model capacity are jointly sufficient.
Prompt robustness	Reduced prompt tailoring substantially lowers accuracy for on-premise models, while Kimi changes little and Claude improves in this exercise.	Model selection should include prompt-sensitivity checks, not only a single optimized benchmark.

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.
NOTE: Detailed paired tests are reported in Annex 1.

6 Conclusions, policy discussion and further research

This paper evaluates retrieval-augmented generation (RAG) systems for supervisory inspection of external audit reports and provides an experimental decomposition of retrieval and model contributions to operational accuracy. Using a real supervisory dataset and a standardized inspection template, we compare alternative retrieval strategies and language models that span on-premise open-weight systems and proprietary cloud models. The paired design allows us to distinguish performance changes driven by retrieval quality from those driven by model capability, and to assess where these components act as complements rather than substitutes. Methodologically, we also show how an independent LLM-as-judge protocol can scale operational scoring against supervisor-provided ground truth, complemented by expert back-to-back checks as a robustness guardrail.

Three findings emerge. First, retrieval quality is a robust and economically meaningful driver of performance. Holding the model fixed, switching from lexical BM25 to Semantic retrieval yields a statistically robust uplift of about 6.2–6.3 percentage points for both Kimi and Llama 70B, and improves performance for smaller models as well. In some configurations, this retrieval uplift offsets a substantial share of the gap associated with weaker model capability, although higher-capacity models remain advantaged on judgement-intensive questions. This result indicates that retrieval improvements can generate large marginal returns even when the generation model is already strong, and it supports a design focus on robust evidence selection as a prerequisite for reliable supervisory assistance.

Second, model capability remains binding once retrieval is aligned. Under symmetric Semantic retrieval, Kimi substantially outperforms smaller on-premise models such as Llama 3B and Mistral 7B, with gaps that are both economically meaningful and concentrated in discordant cases. By contrast, Llama 70B with Semantic retrieval becomes statistically indistinguishable from Kimi with Semantic retrieval, suggesting a capacity threshold above which on-premise systems can match cloud performance in this benchmark. This pattern clarifies the limits of retrieval-based substitution: better retrieval narrows gaps and eliminates many avoidable errors, yet it does not fully compensate for limited reasoning capability on analytically demanding supervisory questions.

Third, prompt robustness constitutes an additional operational dimension. Under a prompt shift that reduces prompt tailoring while holding retrieval fixed, Kimi exhibits a small performance drop, whereas on-premise configurations experience substantially larger degradations. This finding reinforces that governance-relevant evaluation should not rely exclusively on a single prompt formulation, particularly when systems are expected to operate under stable compliance-oriented templates that may not be optimised for each model.

These results also qualify a simple scaling-law narrative in which larger models are always the dominant lever. In supervisory RAG pipelines, performance is a joint function of model capability, retrieval quality, and task complexity. This interaction implies that institutions need not optimise solely for parameter count, especially when on-premise deployment provides value in privacy, security, auditability, and institutional risk control (Khatri and Brown 2010).

Beyond this specific use case, the results contribute to the broader debate in artificial intelligence, and in supervisory technology (SupTech) in particular, around proprietary cloud models versus open-weight models deployed on-premise. In our benchmark, a large on-premise model and a leading cloud model deliver very similar performance once retrieval is aligned: under Semantic retrieval, Kimi reaches 0.873 accuracy and Llama 70B reaches

0.870, and paired tests do not reject equality. Under weaker retrieval, the cloud model remains slightly ahead (0.812 versus 0.807 under BM25). These figures provide a concrete performance trade-off that institutions must balance against other considerations, including confidentiality constraints and geopolitical or vendor dependence.

Finally, we identify several promising lines of further research in the Suptech agenda:

- Using large language models (LLMs), and agentic workflows to link confidential supervisory reports to structured information sources such as AnaCredit. This would allow direct benchmarking across models on tasks such as default prediction, case prioritisation, or the identification of data-quality issues.

- Applying unsupervised learning techniques, including dimensionality reduction and outlier detection, to embeddings produced by domain-adapted or fine-tuned models. This could support anomaly detection and clustering of supervisory cases at scale.

- Extending the use of LLMs to the assessment of other risks that rely heavily on unstructured information, such as reputational, operational, or geopolitical risk. This direction becomes especially relevant as supervisors pursue simplification agendas and seek to reduce reporting burden while improving data usability (Banco de España, 2025).

References

Alonso-Robisco, Andrés, Andrés Azqueta-Gavaldón, José M. Carbó, José L. González, Ana I. Hernáez, José L. Herrera, Jorge Quintana and Javier Tarancón. (2025). "Empowering Financial Supervision: A SupTech Experiment Using Machine Learning in an Early Warning System". Documentos Ocasionales, 2504, Banco de España. https://repositorio.bde.es/handle/123456789/39320

Ash, Elliott, and Stephen Hansen. (2023). "Text Algorithms in Economics". *Annual Review of Economics,* 15, pp. 659-688. https://doi.org/10.1146/annurev-economics-082222-074352

Austin, Ashley A., Tina D. Carpenter, Margaret H. Christ and Christy S. Nielson. (2021). "The Data Analytics Journey: Interactions Among Auditors, Managers, Regulation, and Technology". *Contemporary Accounting Research,* 38(3), pp. 1888-1924. https://doi.org/10.1111/1911-3846.12680

Banco de España. (2017). "Circular 4/2017, de 27 de Noviembre, Del Banco de España, a entidades de crédito, sobre normas de información financiera pública y reservada, y modelos de estados financieros". *Boletín Oficial del Estado (BOE),* 296, 6 December. https://www.boe.es/eli/es/cir/2017/11/27/4

Banco de España. (2024). "Banco de España Response to the External Evaluation of the Use of Technological Innovation in the Prudential Supervisory Function. Action Plan". *Banco de España Evaluation Programme.* https://www.bde.es/f/webbe/INF/MenuHorizontal/SobreElBanco/programa_evaluaciones/ActionPlan.pdf

Banco de España. (2025). *Plan Estratégico 2030.* https://www.bde.es/f/webbe/INF/MenuHorizontal/SobreElBanco/mision/plan_estrategico/PLAN_ESTRATEGICO_BDE_INGLES_version_def.pdf

Basel Committee on Banking Supervision. (2024). *Core Principles for effective banking supervision.* Bank for International Settlements. https://www.bis.org/bcbs/publ/d573.pdf

Biden, Joseph R. (2023). "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence". Presidential Actions, October 30, University of Nebraska - Lincoln. https://digitalcommons.unl.edu/context/scholcom/article/1265/viewcontent/Biden_AI_2023_10_30.pdf

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, … and Percy Liang. (2021). "On the Opportunities and Risks of Foundation Models". *Computer Science – Machine Learning.* https://doi.org/10.48550/arXiv.2108.07258

Castri, Simone di, Stefan Hohl, Arend Kulenkampff and Jermy Prenio. (2019). "The suptech generations". FSI Insights, 19, Financial Stability Institute – Bank for International Settlements. https://www.bis.org/fsi/publ/insights19.htm

Commerford, Benjamin P., Sean A. Dennis, Jennifer R. Joe and Jenny W. Ulla. (2021). "Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence". *Journal of Accounting Research,* 60(1), pp. 171-201. https://doi.org/10.1111/1475-679X.12407

Cormack, Gordon V., Charles L. A. Clarke and Stefan Buettcher. (2009). *Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods.* SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM), New York, 19 July, pp. 758-759. https://doi.org/10.1145/1571941.1572114

DeFond, Mark, and Jieying Zhang. (2014). "A Review of Archival Auditing Research". *Journal of Accounting and Economics,* 58(2-3), pp. 275-326. https://doi.org/10.1016/j.jacceco.2014.09.002

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. (2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". *Computer Science – Computation and Language,* 04805. https://doi.org/10.48550/arXiv.1810.04805

Eilifsen, Aasmund, Finn Kinserdal, William F. Messier and Thomas E. McKee. (2020). "An Exploratory Study into the Use of Audit Data Analytics on Audit Engagements". *Accounting Horizons,* 34(4), pp. 75-103. https://doi.org/10.2308/HORIZONS-19-121

Es, Shahul, Jithin James, Luis Espinosa-Anke and Steven Schockaert. (2023). "Ragas: Automated Evaluation of Retrieval Augmented Generation". *Computer Science – Computation and Language,* 15217. https://doi.org/10.48550/arXiv.2309.15217

European Banking Authority. (2019). *EBA Guidelines on Outsourcing Arrangements EBA/GL/2019/02.* https://www.bde.es/f/webbde/INF/MenuHorizontal/Normativa/guias/EBA-GL-2019_02_EN.pdf

European Parliament and the Council of the European Union. (2014). *Regulation (EU) No 537/2014 of the European Parliament and of the Council of 16 April 2014 on specific requirements regarding statutory audit of public-interest entities and repealing Commission Decision 2005/909/EC (Text with EEA relevance).* https://eur-lex.europa.eu/eli/reg/2014/537/oj/eng

European Parliament and the Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).* https://data.europa.eu/eli/reg/2016/679/oj

European Parliament and the Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance).* https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Fedyk, Anastassia, James Hodson, Natalya Khimich and Tatiana Fedyk. (2022). "Is Artificial Intelligence Improving the Audit Process?". *Review of Accounting Studies,* 27, pp. 938-985. https://doi.org/10.1007/s11142-022-09697-x

Fotoh, Lazarus Elad, and Tatenda Mugwira. (2025). "Exploring Large Language Models in external audits: Implications and ethical considerations". *International Journal of Accounting Information Systems,* 56(S/100748). https://doi.org/10.1016/j.accinf.2025.100748

Gentzkow, Matthew, Bryan Kelly and Matt Taddy. (2019). "Text as Data". *Journal of Economic Literature,* 57(3), pp. 535-574. https://doi.org/10.1257/jel.20181020

Gu, Hanchi, Marco Schreyer, Kevin Moffitt and Miklos Vasarhelyi. (2024). "Artificial intelligence co-piloted auditing". *International Journal of Accounting Information Systems,* 54(S/100698). https://doi.org/10.1016/j.accinf.2024.100698

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin and Ting Liu. (2024). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". *ACM Transactions on Information Systems,* 43(2), pp. 1-55. https://doi.org/10.1145/3703155

Information Technology Laboratory. (2023). *AI Risk Management Framework (AI RMF 1.0).* National Institute of Standards and Technology. https://www.nist.gov/itl/ai-risk-management-framework

International Auditing and Assurance Standards Board (IAASB). (2009). *International Standard on Auditing (ISA) 500 - Audit Evidence.* https://www.ibr-ire.be/docs/default-source/fr/documents/reglementation-et-publications/normes-et-recommandations/isa/isa-english-version/isa-500_en.pdf

International Auditing and Assurance Standards Board (IAASB). (2015). *International Standard on Auditing (ISA) 701 (New), Communicating Key Audit Matters in the Independent Auditor's Report.* https://www.iaasb.org/publications/international-standard-auditing-isa-701-new-communicating-key-audit-matters-independent-auditor-s-3

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen and Wen-tau Yih. (2020). *Dense Passage Retrieval for Open-Domain Question Answering.* Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Online, November 16-20, pp. 6769-6781, Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

Khatri, Vijay, and Carol V. Brown. (2010). "Designing data governance". *Communications of the ACM,* 53(1), pp. 148-152. https://doi.org/10.1145/1629175.1629210

Kokina, Julia, and Thomas H. Davenport. (2017). "The Emergence of Artificial Intelligence: How Automation Is Changing Auditing". *Journal of Emerging Technologies in Accounting,* 14(1), pp. 115-122. https://doi.org/10.2308/jeta-51730

Krishna, Satyapriya, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay and Manaal Faruqui. (2025). "Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation". *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers),* Albuquerque, New Mexico, April 29 - May 4, pp. 4745-4759, Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.243

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel and Douwe Kiela. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". *Advances in Neural Information Processing Systems (NeurIPS 2020),* 33, pp. 9459-9474. https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Liu, Yang, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu and Chenguang Zhu. (2023). "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment". *Computer Science - Computation and Language,* 16634. https://doi.org/10.48550/arXiv.2303.16634

McNemar, Quinn. (1947). "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages". *Psychometrika,* 12(2), pp. 153-157. https://doi.org/10.1007/BF02295996

Nier, Erlend, and Ursel Baumann. (2006). "Market Discipline, Disclosure and Moral Hazard in Banking". *Journal of Financial Intermediation,* 15(3), pp. 332 -361. https://doi.org/10.1016/j.jfi.2006.03.001

Organisation for Economic Co-operation and Development. (2019). *OECD AI Principles Overview.* https://oecd.ai/en/ai-principles

Packard, Helen, and Jermy Prenio. (2023). *External Evaluation of the Use of Technological Innovation in the Prudential Supervisory Function.* En Banco de España, *Evaluation Programme.* Banco de España. https://www.bde.es/f/webbe/INF/MenuHorizontal/SobreElBanco/programa_evaluaciones/Report.pdf

Puig, Pilar, and Javier Tarancón. (2026). *Core Principles for Designing a SupTech Roadmap - Strengthening banking regulation and supervision in the Americas.* Association of Supervisors of Banks of the Americas (ASBA). https://asbasupervision.org/wp-content/uploads/2026/01/ASBA_Principles_Suptech_Roadmap_Oct_2025.pdf

Reimers, Nils, and Iryna Gurevych. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, November 3-7, pp. 3982-3992. https://doi.org/10.18653/v1/D19-1410

Robertson, Stephen, and Hugo Zaragoza. (2009). "The Probabilistic Relevance Framework: BM25 and Beyond". *Foundations and Trends in Information Retrieval,* 4(1-2), pp. 1-174. https://doi.org/10.1561/1500000019

Ru, Dongyu, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang and Zheng Zhang. (2024). "RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation". *Advances in Neural Information Processing Systems,* 37, pp. 21999-22027. https://doi.org/10.52202/079017-0692

Street, Daniel A., and Joseph H. Wilck. (2023). "Let's Have a Chat: Principles for the Effective Application of ChatGPT and Large Language Models in the Practice of Forensic Accounting". *Journal of Forensic and Investigative Accounting,* July to December, forthcoming. https://doi.org/10.2139/ssrn.4351817

Vasarhelyi, Miklos A., Kevin C. Moffitt, Trevor Stewart and Dan Sunderland. (2023). "Large Language Models: An Emerging Technology in Accounting". *Journal of Emerging Technologies in Accounting,* 20(2), pp. 1-10. https://doi.org/10.2308/JETA-2023-047

Wang, Shuai, Shengyao Zhuang and Guido Zuccon. (2021). *BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval.* ICTIR '21: Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval, Virtual Event, Canada, July 11, pp. 317-324. https://doi.org/10.1145/3471158.3472233

Yang, Yi, Mark Christopher Siy UY and Allen Huang. (2020). "FinBERT: A Pretrained Language Model for Financial Communications". *Computer Science - Computation and Language,* 08097. https://doi.org/10.48550/arXiv.2006.08097

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez and Ion Stoica. (2023). "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena". *Computer Science - Computation and Language,* 05685. https://doi.org/10.48550/ARXIV.2306.05685

Annex 1 Detailed robustness analysis

This appendix reports the detailed paired comparisons behind the condensed robustness discussion in Section 5. For each comparison, the underlying question–document pair is held fixed, so the test isolates the effect of changing the retrieval strategy, the generation model, or the prompt template.

A.1.1 Paired comparisons and exact tests on discordant outcomes

Let Y_{iA} and Y_{iB} denote the binary correctness outcomes for question i under configurations A and B. Many comparisons yield ties, either because both configurations answer correctly or both fail. Ties can dominate accuracy differences and can make descriptive non-inferiority shares appear high even when discordant cases systematically favour one side. We therefore complement descriptive shares with exact paired tests that focus on discordant outcomes only.

Define the discordant counts as

$$n_{10} = \sum_i \mathbb{1}(Y_{iA} = 1, Y_{iB} = 0), \qquad n_{01} = \sum_i \mathbb{1}(Y_{iA} = 0, Y_{iB} = 1).$$

Here, $n_d = n_{10} + n_{01}$ denotes the number of discordant pairs.

Under the null hypothesis of symmetric performance,

$$H_0: \Pr(Y_{iA} = 1, Y_{iB} = 0) = \Pr(Y_{iA} = 0, Y_{iB} = 1).$$

The exact McNemar test is equivalent to a two-sided exact binomial test:

$$n10 \sim \text{Binomial}(n_d, 0.5).$$

We report this exact test throughout (McNemar, 1947). Unless stated otherwise, all robustness checks use k = 5 and are evaluated over the same set of paired question-document observations.

A.1.2 Retrieval uplift holding the model fixed

We first quantify the retrieval contribution by holding the generation model fixed and comparing Semantic retrieval to lexical BM25. Table A.1 reports paired outcomes for two strong models that anchor the main deployment trade-off. For both Kimi and Llama 70B, Semantic retrieval yields an accuracy uplift of about 6.2–6.3 percentage points relative to BM25 and wins substantially more often than it loses in discordant pairs.

Table A.1

Retrieval uplift holding the generation model fixed (k = 5)

Accuracy and counts

Accuracy and discordant pairs				
Model	Acc. BM25	Acc. Semantic	Uplift	Discordant
Kimi	0.812	0.873	0.062	89
Llama 70B	0.807	0.870	0.063	92

Discordant outcomes and exact McNemar tests				
Model	Wins n10	Losses n01	Exact p	n10/(n10+n01)
Kimi	63	26	1.1e-04	0.708
Llama 70B	65	27	9.3e-05	0.706

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

a Discordant pairs are question-report observations on which BM25 and Semantic retrieval yield different correctness outcomes.
b Exact p-values follow the McNemar test on discordant pairs.

A.1.3 Model differences holding retrieval fixed

We next isolate the model contribution by holding retrieval fixed and comparing models under symmetric Semantic retrieval. Table A.2 compares three on-premise open-weight models to Kimi under the same Semantic configuration. Kimi substantially outperforms smaller on-premise models under aligned retrieval, while Llama 70B becomes statistically indistinguishable from Kimi.

To illustrate where the symmetric gap arises, Table A.3 reports the paired test by complexity for Mistral 7B vs. Kimi under Semantic retrieval.

Table A.2

Model contribution under symmetric Semantic retrieval (k = 5)

Accuracy and counts

Accuracy gaps and discordant pairs				
Comparison	Accuracy (on-premise)	Accuracy (Kimi)	Gap	Discordants
70B+Sem. vs. Kimi+Sem.	0.870	0.873	-0.003	46
7B+Sem. vs. Kimi+Sem.	0.797	0.873	-0.077	76
3B+Sem. vs. Kimi+Sem.	0.753	0.873	-0.120	106

Discordant outcomes and exact McNemar tests			
Comparison	n10	n01	Exact p
70B+Sem. vs. Kimi+Sem.	22	24	0.883
7B+Sem. vs. Kimi+Sem.	15	61	9.8e-08
3B+Sem. vs. Kimi+Sem.	17	89	5.9e-13

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

Table A.3

Exact paired tests by question complexity. Mistral 7B + Semantic vs. Kimi + Semantic (k = 5)

Counts and exact p-values

Complexity	Total pairs	Ties	Discordant	n10	n01	Exact p
1	120	101	19	6	13	0.167
2	120	112	8	4	4	1.000
3	240	202	38	2	36	5.4e-09
4	120	109	11	3	8	0.227

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

A.1.4 Substitution effect: on-premise strong retrieval vs. cloud basic retrieval

Building on the within-model retrieval uplift, we examine mixed comparisons to assess a practical substitution channel: pairing an on-premise model with strong retrieval (Semantic) versus a cloud model with basic retrieval (BM25). Table A.4 reports the results of these mixed comparisons.

Given the central role of the 70B comparison for deployment decisions, Table A.5 stratifies 70B+Semantic vs. Kimi+BM25 by question complexity.

Table A.4

Mixed substitution margin (k = 5)

Each row compares an on-premise model with Semantic retrieval to Kimi with BM25 retrieval

Comparison	Total	Discordant	n10	n01	Exact p
70B+Semantic vs. Kimi+BM25	600	95	65	30	4.2e-04
7B+Semantic vs. Kimi+BM25	600	119	55	64	0.464
3B+Semantic vs. Kimi+BM25	600	145	55	90	0.005

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

Table A.5

Exact paired tests by complexity. Llama 70B + Semantic vs. Kimi + BM25 (k = 5)

Counts and exact p-values

Complexity	Total pairs	Ties	Discordant	n10	n01	Exact p
1	120	83	37	30	7	1.9e-04
2	120	106	14	8	6	0.791
3	240	202	38	26	12	0.034
4	120	114	6	1	5	0.219

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

A.1.5 Prompt shift robustness

To assess sensitivity to prompt specification, we run a prompt-shift exercise that keeps retrieval fixed (Semantic, k = 5) and varies only the instruction template provided to the generator. We compare an Old prompt that reflects the task-specific, supervisor-aligned prompting used in the main experiments to a New prompt that reduces tailoring while preserving the same input fields and output constraints. Annex 2 reproduces the full New prompt. Table A.4 reports the results of these mixed comparisons.

The prompt-shift exercise, as shown in Table A.6, shows substantial heterogeneity in prompt sensitivity. The New prompt sharply degrades performance for all on-premise models, while Kimi exhibits a small and statistically weak decline and Claude Sonnet 4.6 improves under the New prompt in this benchmark. A paired McNemar test on worsen events rejects equality in favour of Claude losing less than Kimi (p = 0.0288).

Table A.6
Prompt-shift robustness under Semantic retrieval (k = 5)

Old denotes the task-specific prompt used in the main experiments; New denotes a reduced-tailoring prompt

Model	n (a)	Acc. OLD (b)	Acc. NEW (c)	Delta	Discordant	(n01, n10)
Mistral 7B	580	0.790	0.378	-0.412	283	(22.261)
Llama 3B	580	0.745	0.514	-0.231	208	(37.171)
Llama 70B	580	0.866	0.707	-0.159	118	(13.105)
Kimi	580	0.871	0.845	-0.026	57	(21.36)
Claude Sonnet 4.6	580	0.788	0.857	0.069	84	(62.22)

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

a Results are reported on the 580 question-report pairs common after filtering for Claude Sonnet 4.6 availability. McNemar exact p-values test the null of no systematic change between prompts.
b OLD denotes the task-specific prompt used in the main experiments.
c NEW denotes a reduced-tailoring prompt.

Annex 2 Prompts

This appendix reproduces (see Figure A.2.1) the New reduced-tailoring prompt used in the prompt-shift robustness exercise reported in Annex 1.

Figure A2.1

Full text of the reduced-tailoring prompt (New) used in the prompt-shift robustness exercise. Placeholders in curly braces are populated from the input JSON

New prompt used in the prompt-shift robustness exercise

You are an auditor at the Banco de España, expert in report drafting and audit analysis.

The question to be answered is: {question}, as provided in the input JSON.

Answers to the audit questions must be clear and concise.
If no answer is available, respond with "NO INFORMATION".

Respond ONLY using the information contained in the provided context.

The valid values for the answer are: {valid_values}.

The context is: {content}.

Format the output as valid JSON, escaping double quotation marks when necessary, with two fields: the answer under the key "answer" and the traceability under the key "trace".

The field "trace" must include the original text that justifies the answer, together with the page number and the block identifier from which the text is extracted.

Return ONLY parseable JSON using json.loads() in Python, on a single line.
Example of a valid response:
{"answer": "Yes", "trace": "The answer is yes because the document states that a key audit matter exists. Page 12, block 3."}

Within the field "trace", never use double quotation marks or newline characters.
When adding metadata in "trace", never use a dictionary-like format; instead, write the information in natural text to avoid parsing errors.

SOURCE: Banco de España.

Annex 3 Additional robustness results

This appendix reports additional robustness results that complement Section 5. These results are omitted from the main text for conciseness. They do not alter the qualitative conclusions, but provide transparency on counts, discordant structure, and heterogeneity by question complexity.

A.3.1 Paired retrieval comparison: semantic vs. bm25 (all models)

The table A.3.1 reports paired comparisons of Semantic retrieval versus BM25 for each generation model (k = 5). For each model, we report the number of discordant pairs, the counts of Semantic wins (n10) and BM25 wins (n01), descriptive dominance shares, and the exact McNemar p-value. The global row aggregates across all models.

Table A3.1
Paired comparison of Semantic vs. BM25 by model (k = 5)

Model	Pairs	Discordant	n10	n01	Semantic >= BM25	Semantic > BM25	Semantic < BM25	Exact p
Claude Sonnet	600	114	70	44	0.927	0.117	0.073	0.019
Kimi	600	89	63	26	0.957	0.105	0.043	1.1e-04
Llama 3B	600	148	80	68	0.,887	0.133	0.113	0.366
Llama 70B	600	92	65	27	0.955	0.108	0.045	9.3e-05
Mistral 7B	600	117	75	42	0.930	0.125	0.070	0.003
All models	3,000	560	353	207	0.931	0.118	0.069	7.2e-10

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.
NOTE: Semantic wins are (1,0) outcomes; BM25 wins are (0,1) outcomes; exact p-values from McNemar tests on discordant pairs.

A.3.2 Additional stratifications for mixed comparisons (semantic vs. bm25 across deployment)

Section 5 reports stratification by complexity for the central mixed comparison 70B+Semantic vs. Kimi+BM25. For completeness, Tables A.3.2 and A.3.3 report the same stratification for the mid-sized and small open models (Mistral 7B and Llama 3B, respectively). These tables clarify where the aggregate patterns arise and highlight cases with limited discordant mass.

Table A3.2
Exact paired tests by complexity. Mistral 7B + Semantic vs. Kimi + BM25 (k = 5)

Counts and exact p-values

Complexity	Total pairs	Ties	Discordant	n10	n01	Exact p
1	120	85	35	22	13	0.175
2	120	105	15	8	7	1.000
3	240	182	58	24	34	0.237
4	120	109	11	1	10	0.012

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

Table A3.3

Exact paired tests by complexity. Llama 3B + Semantic vs. Kimi + BM25 (k = 5)

Counts and exact p-values

Complexity	Total pairs	Ties	Discordant	n10	n01	Exact p
1	120	82	38	22	16	0.418
2	120	105	15	7	8	1.000
3	240	165	75	25	50	0.005
4	120	103	17	1	16	2.8e-04

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

A.3.3 Additional stratification for symmetric retrieval (70b+semantic vs. kimi+semantic)

Section 5 reports that 70B+Semantic is statistically indistinguishable from Kimi+Semantic in the aggregate. Table A3.4 reports the corresponding stratification by complexity. The aggregate null is consistent with balanced discordant outcomes overall, while some tiers contain relatively few discordant pairs and therefore limited power.

Table A3.4

Exact paired tests by complexity. Llama 70B + Semantic vs. Kimi + Semantic (k = 5)

Counts and exact p-values

Complexity	Total pairs	Ties	Discordant	n10	n01	Exact p
1	120	101	19	13	6	0.167
2	120	113	7	4	3	1.000
3	240	226	14	2	12	0.013
4	120	114	6	3	3	1.000

SOURCE: Authors' calculations on the external audit supervision dataset of the Banco de España.

OCCASIONAL PAPERS

2420 MARIO ALLOZA, JORGE MARTÍNEZ, JUAN ROJAS and IACOPO VAROTTO: Public debt dynamics: a stochastic approach applied to Spain. (There is a Spanish version of this edition with the same number).

2421 NOEMÍ LÓPEZ CHAMORRO: El camino hacia la supremacía cuántica: oportunidades y desafíos en el ámbito financiero, la nueva generación de criptografía resiliente.

2422 SOFÍA BALLADARES and ESTEBAN GARCÍA-MIRALLES: Fiscal drag: the heterogeneous impact of inflation on personal income tax revenue. (There is a Spanish version of this edition with the same number).

2423 JULIO ORTEGA CARRILLO and ROBERTO RAMOS: Parametric estimates of the Spanish personal income tax in 2019. (There is a Spanish version of this edition with the same number).

2424 PILAR L'HOTELLERIE-FALLOIS, MARTA MANRIQUE and DANILO BIANCO: EU policies for the green transition, 2019-2024. (There is a Spanish version of this edition with the same number).

2425 CATERINA CARVALHO-MACHADO, SABINA DE LA CAL, LAURA HOSPIDO, SARA IZQUIERDO, MARGARITA MACHELETT, MYROSLAV PIDKUYKO and ERNESTO VILLANUEVA: The Survey of Financial Competences: description and methods of the 2021 wave.

2426 MARINA DIAKONOVA, CORINNA GHIRELLI and JUAN QUIÑÓNEZ: Economic Policy Uncertainty in Central America and the Dominican Republic.

2427 CONCEPCIÓN FERNÁNDEZ ZAMANILLO and CAROLINA TOLOBA GÓMEZ: Sandbox regulatorio español: impacto en los promotores de los proyectos monitorizados por el Banco de España.

2428 ANDRES ALONSO-ROBISCO, JOSE MANUEL CARBO, EMILY KORMANYOS and ELENA TRIEBSKORN: Houston, we have a problem: can satellite information bridge the climate-related data gap?

2429 ALEJANDRO FERNÁNDEZ CEREZO, BORJA FERNÁNDEZ-ROSILLO SAN ISIDRO and NATIVIDAD PÉREZ MARTÍN: The Banco de España's Central Balance sheet data office database: a regional perspective. (There is a Spanish version of this edition with the same number).

2430 JOSE GONZÁLEZ MÍNGUEZ: The Letta report: a set of proposals for revitalising the European economy. (There is a Spanish version of this edition with the same number).

2431 MARIYA MELNYCHUK and JAVIER MENCÍA: A taxonomy of macro-financial risks and policies to address them.

2432 DMITRY KHAMETSHIN, DAVID LÓPEZ RODRÍGUEZ and LUIS PÉREZ GARCÍA: El mercado del alquiler de vivienda residencial en España: evolución reciente, determinantes e indicadores de esfuerzo.

2433 ANDRÉS LAJER BARON, DAVID LÓPEZ RODRÍGUEZ and LUCIO SAN JUAN: El mercado de la vivienda residencial en España: evolución reciente y comparación internacional.

2434 CARLOS GONZÁLEZ PEDRAZ, ADRIAN VAN RIXTEL and ROBERTO PASCUAL GONZÁLEZ: Navigating the boom and bust of global SPACs.

2435 PATROCINIO TELLO-CASAS: El papel de China como acreedor financiero internacional.

2436 JOSÉ RAMÓN MARTÍNEZ RESANO: CBDCs, banknotes and bank deposits: the financial stability nexus.

2501 PEDRO DEL RÍO, PAULA SÁNCHEZ, MARÍA MÉNDEZ, ANTONIO MILLARUELO, SUSANA MORENO, MANUEL ROJO, JACOPO TIMINI and FRANCESCA VIANI: La ampliación de la Unión Europea hacia el este: situación e implicaciones para la economía española y la Unión Europea.

2502 BANCO DE ESPAÑA: In-person access to banking services in Spain. 2024 monitoring report. (There is a Spanish version of this edition with the same number).

2503 ANDRÁS BORSOS, ADRIAN CARRO, ALDO GLIELMO, MARC HINTERSCHWEIGER, JAGODA KASZOWSKA-MOJSA and ARZU ULUC: Agent-based modeling at central banks: recent developments and new challenges.

2504 ANDRES ALONSO-ROBISCO, ANDRES AZQUETA-GAVALDON, JOSE MANUEL CARBO, JOSE LUIS GONZALEZ, ANA ISABEL HERNAEZ, JOSE LUIS HERRERA, JORGE QUINTANA and JAVIER TARANCON: Empowering financial supervision: a SupTech experiment using machine learning in an early warning system.

2505 JÉSSICA GUEDES, DIEGO TORRES, PAULINO SÁNCHEZ-ESCRIBANO and JOSÉ BOYANO: Incertidumbre en el mercado de bonos: una propuesta para identificar sus narrativas con GDELT.

2506 LAURA JIMENA GONZÁLEZ GÓMEZ, FERNANDO LEÓN, JAIME GUIXERES PROVINCIALE, JOSÉ M. SÁNCHEZ and MARIANO ALCAÑIZ: Evolución de la investigación neurocientífica del efectivo: revisión y perspectivas actuales.

2507 LUIS FERNÁNDEZ LAFUERZA, IRENE ROIBÁS and RAQUEL VEGAS SÁNCHEZ: Indicadores de desequilibrios de precios del mercado inmobiliario comercial.

2508 PANA ALVES and OLIVIER HUBERT: ¿Influye la eficiencia energética en el precio de la vivienda en España?.

2509 ALEJANDRO FERRER and ANA MOLINA: The interaction of liquidity risk and bank solvency via asset monetisation mechanisms. (There is a Spanish version of this edition with the same number).

2510 ISABEL ALCALDE and PATRICIA STUPARIU: Financial education at an early age. (There is a Spanish version of this edition with the same number).

2511 ALEJANDRO GONZÁLEZ FRAGA, AITOR LACUESTA GABARAIN, JOSÉ MARÍA LABEAGA AZCONA, MARÍA DE LOS LLANOS MATEA ROSA, SOLEDAD ROBLES ROMERO, MARÍA VALKOV LORENZO and SERGIO VELA ORTIZ: Estructura del mercado de electrolineras.

2512 FERNANDO ARRANZ GOZALO, CLARA I. GONZÁLEZ MARTÍNEZ and MERCEDES DE LUIS LÓPEZ: Sovereign assets and sustainable and responsible investment: the importance of climate metrics. (There is a Spanish version of this edition with the same number).

2513 IRMA ALONSO-ÁLVAREZ and DANIEL SANTABÁRBARA: Decoding Structural Shocks in the Global Oil Market.

2514 JOSÉ MANUEL CARBÓ, CLAUDIA TOLEDO and ÁNGEL IVÁN MORENO: Hacia un diccionario panhispánico de sentimiento de la estabilidad financiera.

2515 IGNACIO FÉLEZ DE TORRES, CLARA I. GONZÁLEZ MARTÍNEZ and ELENA TRIEBSKORN: The puzzle of forward-looking climate transition risk metrics.

2516 MIGUEL GARCÍA-POSADA: Un análisis de los efectos de la introducción del procedimiento especial de insolvencias para microempresas sobre la propensión a concursar.

2517 ADRIAN VAN RIXTEL: Whatever it takes? Economic policymaking in China in the context of a possible deflationary spiral.

2518 PATRICIA STUPARIU and JUAN RAFAEL RUIZ: Adding up the benefits: education, numeracy and financial literacy. (There is a Spanish version of this edition with the same number).

2519 DAVID CUBERES, AITOR LACUESTA, MARÍA DE LOS LLANOS MATEA and DANIEL OTO-PERALÍAS: El efecto de la regulación sobre el tamaño de las plantas fotovoltaicas.

2520 MARINA GARCÍA GIL and DMITRY KHAMETSHIN: Ayudas directas de la línea COVID-19: los determinantes de la asignación y el efecto sobre el crédito bancario.

2521 LUCÍA CUADRO-SÁEZ, CORINNA GHIRELLI, MAXIMILIANO MORENO-LÓPEZ and JAVIER J. PÉREZ: Monitoring and forecasting food prices in the euro area.

2522 LAURA HOSPIDO, JÚLIA MARTÍ LLOBET and CARLOS SANZ: ¿Qué políticas son efectivas para reducir la exclusión social? Evaluación de cinco proyectos piloto de inclusión social a través de ensayos aleatorizados.

2523 MARIO ALLOZA, MARÍA ELENA CRISTÓBAL RODRÍGUEZ, JULIA GARCÍA-ROYO DÍAZ, BEATRIZ GONZÁLEZ, ALBERTO MARTÍN DEL CAMPO SOLA, ANE MARTÍN UGARTE, ENRIQUE MORAL-BENITO and IRENE PINILLA MELGAREJO: Diagnóstico y consecuencias económicas del grado de competencia en las licitaciones públicas.

2524 ADRIÁN CARRO, JORGE E. GALÁN, ENRIC MARTORELL and RAQUEL VEGAS: A literature review on ex-ante and ex-post analysis of the implications of borrower-based macroprudential measures.

2525 PAULA SEMPERE and RAQUEL PANTOJA: Efectivo y *neuromarketing:* Concepto, funciones y aplicación.

2601 RODOLFO G. CAMPOS, JACOPO TIMINI, FRANCESCA VIANI and ELENA VIDAL: The EU-Mercosur agreement: analysis of its characteristics from a sectoral perspective. (There is a Spanish version of this edition with the same number).

2602 PABLO AGUILAR, CORINNA GHIRELLI and SAMUEL HURTADO: MTBE v2025: new version of the Quarterly Model of the Banco de España.

2603 IRMA ALONSO-ALVAREZ, EKATERINA BUKINA, MARINA DIAKONOVA, NINO KHITARISHVILI, JAVIER J. PÉREZ and PEDRO PIQUERAS: Geopolitical risk: a database of general and bilateral indices.

2604 RUBÉN DOMÍNGUEZ-DÍAZ, MARTA GARCÍA-RODRÍGUEZ, JAVIER QUINTANA and RUBÉN VEIGA-DUARTE: Estimación del crecimiento potencial de la economía española: una revisión metodológica.

2605 CARLES MANERA, FERRAN NAVINÉS, JAVIER FRANCONETTI, MIQUEL QUETGLAS and JOSÉ A. PÉREZ-MONTIEL: Regional outlook on labor productivity in Spain, 2000–2022.

2606 RICARDO BARAHONA: Current trends in performance and flows in the Spanish pension funds industry.

2607 FRANCISCO GONZÁLEZ RODRÍGUEZ, ALBERTO ORTS TORRES, JOSÉ MARÍA SERENA GARRALDA and MIQUEL TARÍ SÁNCHEZ: Issuance yield of MREL-eligible bank debt: a benchmarking model.

2608 MÓNICA CORREA-LÓPEZ, MAR DELGADO-TÉLLEZ and MARTA SUÁREZ-VALERA: Europa en transición energética: respuestas y desafíos tras dos crisis consecutivas.

2609 AITOR LACUESTA, MARÍA VALKOV and MICOLE DE VERA: Un análisis de los mecanismos de protección de los operadores económicos establecidos por la Ley de garantía de la unidad de mercado.

2610 Encuesta Financiera de las Familias (EFF) 2024: métodos, resultados y cambios desde 2022.

2612 MICAELA ARIAS, LUCÍA CRIADO, JAVIER GARCÍA-VERDUGO, EDUARDO GUTIÉRREZ, ALMUDENA KESSLER, JOSÉ MANUEL MONTERO, SARA PEREIRA and PAU ROLDÁN-BLANCO: Evolución de la estructura de mercado y de los indicadores de competencia en España en las dos últimas décadas.

2613 ANDRÉS ALONSO-ROBISCO, JOSÉ MANUEL CARBÓ, CARLOS JOSÉ GARCÍA, JORGE QUINTANA and JAVIER TARANCÓN: Retriever or reasoner? Decomposing retrieval-augmented generation performance in external audit supervision.