



POLITECNICO DI TORINO

Master degree course in Digital Skills for Sustainable Societal
Transitions

Master Degree Thesis

**Compliance without compromise: a private,
high-precision AI chatbot with RAG for
answering questions about Dutch financial
regulation**

Relatore

Prof. Gianvito Urgese

Candidato

Hessel Frank Müllemeister

July 2026

Abstract

A Dutch trust office offering compliance as a service to companies under dense, layered anti-money-laundering and integrity obligations (Wwft, Wtt 2018, Sanctiewet 1977) needs deep, specialised knowledge of legislation, supervisory guidance and case law. Identifying compliance gaps means searching through statutes, implementing decrees, guidance and case law: slow, expert work in which a mistake can carry regulatory and reputational liability. Popular large language models offer faster research, but hallucination and limited auditability are key drawbacks. Strict confidentiality standards also prevent uploading sensitive client ownership structures and source-of-funds material to a public cloud model. A market scan confirms that no existing product combines Dutch trust-office coverage, private deployment and client-document handling in one tool.

This thesis asks whether it is possible to build a private, bilingual compliance assistant, designed to perform best in Dutch, that is useful enough to help and transparent enough to be auditable. To answer this, a fully on-premise, open-source retrieval-augmented generation (RAG) system was built and evaluated as a proof of concept. The pipeline ingests a four-layer corpus ranked by authority (binding regulation, supervisory guidance, case law and uploaded client files), embeds it with bge-m3 (a multilingual model that is strong on Dutch) in a local ChromaDB index, and retrieves it in two stages: a broad parallel search with a guaranteed binding-law quota, followed by a precise bge-reranker-v2-m3 pass. A locally hosted qwen3:14b model then generates a grounded, source-cited answer at temperature zero. The system runs entirely on a laptop and every answer is auditable.

Output quality was evaluated on ten representative Dutch compliance questions, each scored from 0 to 5 by an experienced trust-office specialist against the obligations a complete answer must contain. Seven of the ten reached a "Go" threshold for use as a supervised research aid, with strong performance on customer due diligence, Politically Exposed Persons (PEP), Financial Intelligence Unit (FIU) reporting and the risk-based approach. The three "No-Go" cases involved Ultimate Beneficial Owner (UBO) determination in complex structures, the Sanctions Act and source-of-wealth assessment. They exposed recurring weaknesses: norms that sit in implementing decrees rather than the parent act,

fragmentation from fixed-length chunking, residual hallucination of article numbers and a clear drop in performance on Dutch legal text compared with English. On the positive side, the tool points to internal cost savings of roughly 25% to 33%, an output increase of around 33% and the potential to grow the team with a junior rather than a senior hire.

The research question can be answered with positivity and enthusiasm. Yes, a private, locally deployed compliance assistant proves itself feasible and economically valuable as a supervised research aid. However, the language model's core reliability limits have not completely been solved. Those limits become even greater in this specific Dutch setting. Human oversight therefore remains structurally necessary, not just temporary. This thesis aims to contribute by adding the first studied characterisation of RAG behaviour on Dutch trust-office regulatory text to the field, qualitatively evaluated on whether answers capture every applicable obligation rather than just matching wording. This thesis shows how far a confidential, low-cost compliance assistant can responsibly be taken.

Acknowledgments

My gratitude goes out to professor Gianvito Urgese, for overseeing this thesis project, providing valuable feedback and helping me focus on the parts that matter.

I would also like to sincerely thank my family and friends for patiently supporting me through a lengthy period of limited thesis progression. Your unwavering trust in my belief that I will graduate someday, has finally paid off.

A special thanks goes out to my brother. Without his support, expertise, knowledge, time, patience, love and care, this thesis would not have been possible.

Contents

List of Figures	9
List of Tables	10
List of Abbreviations	11
1 Introduction	13
1.1 Background and Problem Statement	13
1.2 Research Motivation	14
1.3 Research Question	16
1.4 Scope	17
1.5 Thesis Structure Overview	18
2 Literature review	21
2.1 Conceptual Foundations of Retrieval-Augmented Generation . .	21
2.2 Compliance as a regulatory and knowledge problem	23
2.3 Why RAG Addresses Compliance Challenges - Knowledge fresh- ness and regulatory temporality	23
2.4 Auditability and provenance	24
2.5 Hallucination mitigation	25
2.6 RAG as a knowledge equaliser	25
2.7 LLM capability for regulatory-text interpretation	26
2.8 Hybrid retrieval for regulatory corpora	26
2.9 End-to-end compliance automation	26
Downsides and limitations of RAG	27
2.10 RAG does not eliminate hallucination	27
2.11 Retrieval quality does not guarantee generation quality	27
2.12 Context distraction and catastrophic regression	28
2.13 Performance degradation in non-English and multi-regulation con- texts	29
2.14 Standard evaluation metrics are ill-suited to regulatory text	30

Zooming in on the financial sector: regulatory positions and evidence gaps	30
2.15 Authoritative regulatory positions	30
2.16 Domain-specific empirical studies and the lack thereof	31
2.17 Evidence gaps	32
2.18 Note on the evidence base	32
3 Market research	33
Practical Relevance: Why no existing product closes the gap	33
3.1 Purpose	33
3.2 Stakeholder requirements	34
3.3 How potential companies were identified	35
3.4 What the market offers	35
3.5 The gap	37
3.6 From market gap to research justification	38
3.7 Conclusion	39
4 Methodology	41
4.1 Purpose of this chapter	41
4.2 Research design	42
4.3 Use case	43
4.4 The corpus	44
4.5 System design	45
4.5.1 Ingestion, cleaning and chunking	45
4.5.2 Embedding and storing	47
4.5.3 Retrieval, reranking and the regulation quota	48
4.5.4 Context assembly, prompting and generation	50
4.5.5 Interfaces	52
4.6 Evaluation design	53
4.7 Key Performance Indicators	54
5 Results	55
5.1 Overview of the evaluation	55
5.2 Scores and Go/No-Go outcomes	56
5.3 Comparison of the two configurations	57
5.4 Completeness of the answers	58
5.5 Observed behaviours	59
5.5.1 Uneven contribution of the binding-regulation layer	59
5.5.2 Residual fabrication of regulatory specifics	60
5.5.3 Relevant retrieval, incomplete answer	60
5.5.4 Behaviours in the Dutch-language setting	61
5.5.5 Comparison with cloud tools	61

5.5.6	Slower thinking	61
5.6	Key Performance Indicators	62
6	Discussion	65
6.1	Purpose of this chapter	65
6.2	The pattern behind the scores	65
6.3	The results in light of the literature	66
6.4	What the results mean for practical adoption	67
6.5	Limitations of this study	68
7	Conclusion and Future Work	71
7.1	Answers to the sub-questions	71
7.2	Answer to the main research question	72
7.3	Contribution	73
7.4	Future work	73
7.5	Closing remark	74
A	Requirements interview with the trust office representative	75
B	Documents used in the corpus of the RAG4Compliance system	77
C	Record of telephone conversation with the trust office representative	85
D	Testing results with the trust office representative	87
E	Hessel's AI Compliance Bot - Additional feedback by the trust office representative	93
	Bibliography	95

List of Figures

4.1	Pipeline overview	46
5.1	Key Performance Indicators	63

List of Tables

5.1	Expert scores and Go/No-Go outcomes across the ten test questions.	56
D.1	Full evaluation record: expert review of the ten test questions. . .	88
D.2	Scoring rubric used in the evaluation.	91

List of Abbreviations

AFM	Autoriteit Financiële Markten (Dutch Authority for the Financial Markets)
AI	Artificial Intelligence
AML	Anti-Money Laundering
AML/CFT	Anti-Money Laundering / Countering the Financing of Terrorism
AMLA	Authority for Anti-Money Laundering and Countering the Financing of Terrorism
AMLD	Anti-Money Laundering Directive
CBb	College van Beroep voor het bedrijfsleven (Dutch Trade and Industry Appeals Tribunal)
CDD	Customer Due Diligence
DNB	De Nederlandsche Bank (Dutch Central Bank)
DORA	Digital Operational Resilience Act
EBA	European Banking Authority
ECB	European Central Bank
EDD	Enhanced Due Diligence
EU	European Union
FATF	Financial Action Task Force
FIU	Financial Intelligence Unit (FIU-Nederland)
FSB	Financial Stability Board

GDPR	General Data Protection Regulation
KPI	Key Performance Indicator
KYC	Know Your Customer
LLM	Large Language Model
MiFID II	Markets in Financial Instruments Directive II
NLP	Natural Language Processing
NVB	Nederlandse Vereniging van Banken (Dutch Banking Association)
PEP	Politically Exposed Person
RAG	Retrieval-Augmented Generation
RegTech	Regulatory Technology
SIRA	Systematische integriteitsrisicoanalyse (systematic integrity risk analysis)
UBO	Ultimate Beneficial Owner
Wft	Wet op het financieel toezicht (Financial Supervision Act)
Wtt 2018	Wet toezicht trustkantoren 2018 (Trust Offices Supervision Act 2018)
Wwft	Wet ter voorkoming van witwassen en financieren van terrorisme (Money Laundering and Terrorist Financing (Prevention) Act)
XML	Extensible Markup Language

Chapter 1

Introduction

1.1 Background and Problem Statement

No pain, no gain. After the 2008 financial crisis, the Greek debt crisis, Icesave going bankrupt, and many other major financial crises, European financial regulation has never been stricter. Companies that manage other people's money, such as banks and investors, are heavily regulated. Trust offices create another layer of complexity, and have therefore received specific regulation of their own in the Netherlands. A trust office, often called a Corporate Service Provider, manages companies for clients, frequently acting as their statutory director and providing a registered address. It typically handles accounting, annual accounts, compliance, and administrative support so that (often foreign) companies can operate and remain compliant with local laws and regulations. One of the services that such a firm provides is compliance research for their customers. When a foreign company in the financial sector wants to register itself in the Netherlands, they have the responsibility to comply with Dutch and EU legislation. Furthermore, the trust office has the responsibility to do business with law-abiding customers, and to cut connections with 'bad' customers. Currently, the work required to jump through these hoops is largely specialised human work. Accuracy is extremely important, so automating this work is challenging.

The Dutch regulatory landscape governing financial institutions has grown significantly in complexity in recent years. The Financial Supervision Act (*Wet op het financieel toezicht*, Wft) forms the overarching legal framework, imposing governance, licensing, conduct, and integrity requirements on all licensed financial institutions. Layered on top of this is the Money Laundering and Terrorist Financing (Prevention) Act (*Wet ter voorkoming van witwassen en financieren van terrorisme*, Wwft). The Wwft now requires institutions to maintain a written risk policy, perform extensive customer due diligence (CDD) including Ultimate Beneficial Owner (UBO) identification and Politically Exposed Person (PEP) screening, monitor transactions on an ongoing basis, report unusual transactions

to the Financial Intelligence Unit Nederland (FIU-Nederland) without delay, and demonstrate to supervisors that every control measure taken is proportionate and defensible. For trust offices specifically, the Trust Offices Supervision Act 2018 (*Wet toezicht trustkantoren 2018*, Wtt 2018) adds a further layer of obligations, including strict integrity requirements, mandatory internal compliance functions that cannot be outsourced, and an expanded concept of integrity risk that requires judgment-based client acceptance decisions rather than mechanical rule-following. Enforcement by the Dutch Central Bank (DNB) and the Dutch Authority for the Financial Markets (AFM) has intensified accordingly, with administrative fines and licence revocations becoming an increasingly visible consequence of non-compliance.

In order to provide trustworthy and watertight compliance services, one has to know everything about the applicable financial regulation across all of these frameworks simultaneously. The challenge here is not only the vast amount of information one must know, understanding and applying the rules correctly requires a lot of extra knowledge and experience. The Wft, Wwft, and Wtt 2018 are principles-based instruments, meaning that open norms such as "adequate measures" and "risk-based approach" require contextual legal reasoning rather than simple rule application. Compliance teams must combine these laws with guidelines, best practices, legal verdicts, and Q&As that are updated continuously and at different times, translating every change into internal policies, onboarding criteria, monitoring logic, and staff procedures. Not only is this tedious and time-consuming work, but there is also a significant margin for error when specific details are overlooked. Missed compliance gaps can cause major liability, substantial fines, or worse. As the pace of regulatory change accelerates and the depth of required documentation grows, manual regulatory interpretation has very limited scalability potential for institutions that need to deliver consistent, accurate, and auditable compliance outcomes. With recent developments in artificial intelligence, however, there may be a potential solution.

1.2 Research Motivation

Before AI, finding something in a large corpus of documents would require you to search for specific keywords. This still requires humans to work with these rather crude search results. These days, a computer can turn pieces of text into numbers that capture its meaning and context. A Large Language Model (LLM) can subsequently interpret these numbers, together with your question, to give you a much more usable answer than raw search results. Such answers can be surprisingly accurate [1, 2].

These technological developments are not going unnoticed by companies who are aiming to automate compliance. A growing market of legal-tech and RegTech

vendors offers solutions that promise to automate KYC workflows, screen clients against sanctions lists, and flag unusual transactions. Likewise, general-purpose AI tools such as ChatGPT and Claude seemingly became practically omnipotent by either being able to ingest documents and answer questions contextually, or by responding surprisingly accurately based solely on training data. Alternatively, off-the-shelf RAG platforms such as Ragie.ai handle the complex mechanics of chunking, embedding, retrieving, and generating responses with minimal technical setup. However, these solutions share a fundamental limitation: they are either too generic to be trusted to reliably respond on Wft, Wwft, and Wtt 2018 applications, or they operate as black boxes that send sensitive client data to external servers, creating potentially unacceptable confidentiality risks when dealing with sensitive company data. Furthermore, most commercial tools are designed to support compliance officers in navigating pre-defined workflows, rather than answering novel, context-specific legal questions with auditable, document-grounded reasoning.

This leaves a meaningful gap: there is currently no readily available, privacy-preserving, and domain-specific tool that allows a trust office to query a specific body of applicable financial regulation, and receive trustworthy and auditable answers. Manual review processes fill this gap today, but as established in the previous section, they do not scale. What is needed is a specialised system that knows every single detail of the applicable regulatory framework, and that can respond to complex compliance questions with the reliability and consistency that manual interpretation cannot guarantee. No detail can be overlooked.

Retrieval-Augmented Generation (RAG) is a novel technique that addresses exactly this problem. RAG uses external data sources to enable a large language model (LLM) to retrieve and incorporate information at query time, rather than relying solely on what was learned during training [1]. Since an LLM with RAG first refers to a specified body of documents before responding to user queries, it is possible to create a highly specialised system grounded in a curated and controlled knowledge base [3]. With RAG, hallucination can be significantly reduced, accuracy can be increased, and the system's responses can be traced back directly to the source documents, a critical requirement in a regulatory context [4, 5]. Moreover, by running the model locally using open-source tooling, data confidentiality can be managed more closely without relying too much on third-party infrastructure, which aligns with current recommendations for on-premise LLM deployment in sensitive domains under frameworks such as GDPR [6].

The idea of this thesis is therefore to explore the potential of RAG for a trust office in the Netherlands, specifically for the service of compliance research in the financial sector. A proof of concept should demonstrate that a locally-running, RAG-powered system can:

- Showcase sufficient LLM output quality, usefulness, specificity, accuracy, and completeness
- Show potential to reduce human resources spent on compliance research
- Show potential to reduce the error margin in provided services
- Aid in achieving consistency in provided services
- Decrease the trust office's liability for overlooked compliance gaps
- Increase output scalability

1.3 Research Question

The central objective of this thesis is to determine whether a locally-running, RAG-based system can meaningfully improve the quality, consistency, and efficiency of compliance research at a Dutch trust office. This objective is captured in the following primary research question:

How can the use of Retrieval-Augmented Generation increase service quality, consistency, and resource efficiency for a trust office when providing compliance services in the Dutch financial sector?

To answer this question in a structured and measurable way, the following sub-questions guide the research:

1. **Technical feasibility:** To what extent can a locally-running RAG system be built that reliably ingests, indexes, and queries Dutch financial regulatory documents?
2. **Retrieval accuracy:** How accurately does the RAG system retrieve the most relevant regulatory passages in response to compliance-related queries, and what chunking, embedding, and retrieval strategies produce the best results?
3. **Regulatory coverage:** To what degree can the system provide complete and accurate answers across the relevant regulatory frameworks without producing hallucinated or incomplete responses?
4. **Output quality:** To what extent does the use of RAG improve the quality, specificity, and usefulness of LLM-generated compliance answers compared to a baseline model without retrieval augmentation?

5. **Confidentiality and usability:** Can the system be designed and operated in a way that ensures sensitive client and company data never leaves the local environment, making it suitable for deployment in a professional trust office context?

Together, these questions define the scope of the proof of concept and provide the criteria against which the system’s performance will be evaluated in later chapters.

1.4 Scope

In order to keep the research focused and achievable within the constraints of a thesis project, it is important to clearly define what falls within and outside the boundaries of this study.

In scope:

- **Regulatory framework:** The system is designed and evaluated against key primary Dutch financial regulatory instruments applicable to a trust office: the Wwft, and the Wtt 2018, and the Sanctiewet 1977. Where directly relevant, associated delegated regulations, supervisory guidance issued by DNB and AFM, and applicable EU directives that have been transposed into Dutch law are also included in the knowledge base.
- **Proof of concept:** The primary deliverable of this thesis is a functioning proof of concept. The proof of concept is a locally-running, RAG-powered system built in Python, using open-source tooling, capable of answering compliance-related queries grounded in the documents described above.
- **Qualitative and functional evaluation:** The system’s performance is evaluated on the basis of retrieval accuracy, answer quality, regulatory coverage, and consistency, assessed through structured test queries and qualitative analysis of the system’s outputs against the source documents.
- **Confidentiality by design:** The local deployment architecture is treated as a core design requirement and is evaluated as part of the proof of concept.

Out of scope:

- **Full production deployment:** This thesis does not aim to deliver a production-ready system. Security hardening, scalability infrastructure, user interface design, and integration with existing trust office workflows are not addressed and are left for potential future development.

- **Non-Dutch jurisdictions:** The system is scoped exclusively to Dutch and directly applicable EU financial regulation. Compliance requirements in other jurisdictions are not covered.
- **Real client data:** No actual client data, company files, or live compliance cases from the trust office are used at any stage of this research. All testing is performed using publicly available regulatory documents and synthetic or anonymised queries.
- **Quantitative benchmarking at scale:** Given the proof-of-concept nature of the project, large-scale automated benchmarking against standardised datasets is not performed. The evaluation is primarily qualitative and exploratory in nature.
- **Comparison with commercial RegTech vendors:** Although commercial alternatives are discussed in the research motivation, a systematic comparative evaluation against paid third-party platforms falls outside the scope of this thesis.

These delimitations ensure that the research remains realistic in ambition while still producing a meaningful and demonstrable result that can serve as a foundation for further development.

1.5 Thesis Structure Overview

The remainder of this thesis is structured as follows.

Chapter 2: Literature review establishes the conceptual and technical foundation for the research. It explains the basics of Retrieval-Augmented Generation (RAG). From there, it progresses to the specific mechanics of RAG: how documents are chunked, embedded, stored, and retrieved, and how retrieved context is passed to a language model to generate grounded responses. It also elaborates on drawbacks, hallucination and performance measurement. This chapter provides the theoretical basis needed to understand, design, and evaluate the proof-of-concept system developed later in the thesis.

Chapter 3: Market research examines whether a trust office needs to build a custom system at all, or whether a suitable product could simply be bought instead. It derives the requirements that follow from the firm's own situation and legal obligations, surveys the commercial tools currently available against those requirements, and shows where the offering falls short. This chapter concludes that no single product combines the regulatory depth, private deployment, client-document handling, and Dutch-specific coverage that a Dutch trust office needs, which is what justifies building and studying the proof of concept developed in the chapters that follow.

Chapter 4: Methodology describes the architecture and design choices of the proof-of-concept RAG system, including the selection of the language model, embedding strategy, chunking approach, retrieval mechanism, and local deployment setup. It also explains the evaluation methodology used to assess the system's performance.

Chapter 5: Results presents the outcomes of the proof of concept. The system's performance is assessed against the sub-questions defined in Chapter 1, covering retrieval accuracy, output quality, regulatory coverage, and suitability for confidential deployment.

Chapter 6: Discussion interprets the results in the context of the broader research question, reflects on the limitations of the current implementation, and considers what the findings mean for the practical adoption of RAG in a trust office setting.

Chapter 7: Conclusion and Future Work answers the primary research question, summarises the key findings, and outlines the most promising directions for further development. This includes potential steps toward a production-ready system and integration with real trust office workflows.

Together, these chapters move from theoretical grounding through system construction to critical evaluation, providing a complete account of whether and how RAG can serve as a viable compliance research tool for a Dutch trust office.

Chapter 2

Literature review

This chapter aims to give a comprehensive overview of what the current state of research is regarding the application of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) in conjunction with regulatory compliance, with particular attention to the financial sector and to the specific operating context of this study: a Dutch trust office (trustkantoor, or corporate service provider) that performs compliance research for its clients. A deep dive is made into conceptual foundations, documented capabilities, empirically established limitations, performance metrics, and identified research gaps. Based on the findings, RAG can be an effective solution to some key compliance challenges. Also, this review reveals that there is not much literature on specifically financial compliance applications. In general, the performance of both LLMs and RAG systems on non-English regulatory text in the context of the financial sector, is systematically understudied, even though RAG is structurally well-suited to several well-established compliance challenges, due to knowledge freshness, auditability, and hallucination reduction. However, the literature applying it to financial-compliance tasks specifically remains immature: dedicated benchmarks are largely absent and end-to-end empirical evaluations in regulated financial contexts are rare.

2.1 Conceptual Foundations of Retrieval-Augmented Generation

Retrieval-Augmented Generation was proposed already six years ago by Lewis et al. [1], when this research combined a pre-trained parametric sequence-to-sequence generator with a non-parametric dense vector retriever over an external document corpus. In simple terms, a standard LLM that was trained by its developers on many topics, such as ChatGPT, was combined with a ‘research assistant’, a retriever, that tries to fetch the most relevant parts of a collection of

documents that was not initially part of the general training dataset of the LLM. The core insight here is that the trained parameters of a neural model do not have to contain all factual knowledge, since it can also be externalised into an updatable index and retrieved at inference time. Lewis et al. introduced two versions of RAG: RAG-Sequence and RAG-Token. In the former, the model fetches a set of relevant documents only once and then uses that fixed set to base its entire answer upon. In RAG-Token, the model can be more flexible: for different parts of the answer, it can draw on different retrieved documents, picking the best source word by word. In 2020, both of these RAG versions outperformed all other existing methods on standard tests for answering open-domain questions [1].

Building on this original foundation, Gupta et al. [3] trace how RAG has developed and changed forms from its original architecture, to more sophisticated architectures. These new forms include modular RAG, graph-RAG, and adaptive RAG. In modular RAG, each part of the pipeline can be swapped out independently. In graph-RAG structured knowledge graphs are used alongside or instead of vector-based search indices. An adaptive RAG system on the other hand, automatically chooses a retrieval strategy depending on the nature of the question. These developments have created flexibility, but also opportunities for the system to behave in unintended ways [3]. Zhao et al. [7] further investigate how retrieval can augment generation. Three strategies are defined: pre-generation augmentation (retrieved passages prepended to the prompt), iterative augmentation (retrieval interleaved with generation), and post-generation augmentation (retrieved evidence used to verify or revise outputs). Each strategy carries distinct implications for compliance, where the traceability of the relationship between retrieved evidence and generated output is a governance requirement rather than a desirable feature [7].

Challenges in retrieval are also identified. The so-called “lost in the middle” problem infers that LLMs systematically give lower weight to information located in the middle of long retrieved contexts, attending preferentially to the beginning and end of input sequences [3]. This could mean that when a regulatory passage appears in the middle of a large context window, the generator may effectively ignore it, with no visible indication to the user. This failure mainly resides in the generation part of the system, rather than in the retrieval part, and it is an example of why retrieval quality is a necessary but insufficient condition for output quality in compliance applications.

2.2 Compliance as a regulatory and knowledge problem

The nature of financial compliance presents a distinctive combination of challenges that suit a RAG approach well. For example, financial regulations are usually written in voluminous, technically dense, frequently amended texts, that are expressed in specific legal language. Relevant regulatory frameworks such as the Dutch interpretation of the European Anti-Money Laundering Directive (AMLD), span hundreds of pages of primary legislation, which are supported by supervisory guidelines and best practices, producing a layered, overlapping body of obligations that must be tracked and applied in a consistent, auditable, and temporally current manner.

Three structural limitations of conventional compliance automation that are directly relevant to this challenge, are defined by Hassani et al. [8]. First, prior NLP approaches look at individual sentences as the unit of regulatory analysis, but regulatory obligations are usually explained in multiple clauses or articles. Therefore, analysing regulation sentence by sentence, systematically undercaptures the obligation scope. Second, the explanation of why a provision is relevant, is very hard to capture with classification and information-extraction pipelines. Third, existing models lack generalisability across regulatory domains. Training a model for one regulatory framework, does not mean it performs well on another without retraining. These limitations require a generation-capable architecture that integrates explicit reasoning, provenance, and domain-adaptable knowledge retrieval.

The governance dimension is also really important. A three-layered framework for LLM auditing comprising governance audits of technology providers, model audits evaluating performance and risk prior to deployment, and application audits of downstream use in specific institutional contexts is proposed by Mökander et al. [9]. Financial institutions that are supervised by authorities such as De Nederlandsche Bank, the central bank of the Netherlands (DNB), or the European Banking Authority (EBA), are already expected to maintain documentation supporting all three layers of oversight, so this framework maps directly onto the expectations of those financial supervisors.

2.3 Why RAG Addresses Compliance Challenges - Knowledge freshness and regulatory temporality

Since financial regulations are constantly subject to amendments and changes, the ability to easily update a RAG knowledge base without retraining the entire model, is a compelling argument for RAG over purely parametric LLMs

in compliance contexts. A parametric LLM trained on a static corpus has a knowledge cutoff that may render regulatory guidance obsolete within months of deployment. RAG resolves this issue by separating the knowledge base from the model weights. It is possible to add, update, or delete documents in the external index without retraining the generator. According to Lee et al. [10], this is RAG's primary practical advantage in financial applications. Hallucination risk is controlled, while regulatory knowledge bases are kept current. Nie et al. [11] also stress that regulatory temporality is an unaddressed problem in most existing FinLLM research, also noting that when RAG knowledge bases are not systematically updated, they themselves can become stale and outdated. This shifts the problem from model training to index management.

2.4 Auditability and provenance

Whoever thinks that auditability in financial-compliance systems is just a nice-to-have, is wrong. It is a hard, regulatory requirement. Pérez-Cruz et al. [12] report that AI explainability is the single most frequently cited concern when financial institutions engage with supervisors, and that a 2024 Bank of England and FCA survey found half of respondents had only partial understanding of the AI technologies they deployed due to third-party model not being transparent. When general-purpose LLMs are used in regulatory compliance, probabilistic output distributions and proprietary pretraining data make auditability very difficult.

In RAG the generator is conditioned to base the output only on retrieved passages, which can be logged and inspected. Therefore, a partial but meaningful step forward can be achieved on auditability. RAG systems can provide grounded provenance: a record of which document, passage, and version of the regulatory text informed a given output. This is categorically different from parametric generation, where the source of any claim is distributed across billions of weights with no tractable means of attribution. Brown et al. [13] recommend pipelines with provenance-awareness as the appropriate way forward for compliance applications, arguing that systems should expose uncertainty estimates alongside traceable evidence rather than generating confident outputs without attribution. The above mentioned "partial" matters: while retrieval provenance can be logged, the relationship between retrieved evidence and the generator's reasoning remains difficult to interpret, and the auditability of the generation step itself remains an open problem.

2.5 Hallucination mitigation

Text generated by an LLM is a prediction of the most probable continuation of a prompt given a specific training distribution. Unfortunately, this process has no internal mechanism for distinguishing known facts from plausible-sounding ‘facts’. “Hallucination” describes outputs that are fluent and confident but factually unsupported or incorrect [14]. By limiting generation to retrieved passages, RAG constrains the generator to produce output that is consistent with documents from the corpus, thereby reducing the chance of unsupported claims. Kang and Liu [15] demonstrate that hallucination in financial contexts is qualitatively distinct from general-domain hallucination: financial LLMs tend to fabricate plausible numerical values, cook up regulatory thresholds, and produce incorrect interpretations of financial concepts, with confidence. Due to this confidence, these errors can be particularly dangerous because they may not be immediately recognisable to a non-expert reviewer. RAG’s grounding mechanism can be especially valuable in these cases, anchoring numerical and definitional claims to specific retrieved passages and, more importantly, giving the user a way to check where these claims originate from. Roychowdhury et al. [16] demonstrate in a production financial QA system that combining retrieval with query-intention classification and multi-metric post-hoc hallucination scoring achieves confidence scores exceeding 90%, substantially above what retrieval alone could produce.

2.6 RAG as a knowledge equaliser

When tasks are knowledge intensive, such as giving professional regulatory compliance advice in the financial sector, RAG’s most consistently replicated positive finding is that it substantially improves LLM performance, compared to parametric only generation. Beauchemin and Khoury [17] evaluated 51 LLMs on a regulatory benchmark of 807 multiple-choice questions derived from Quebec insurance regulatory certification materials, comparing closed-book and RAG-augmented conditions. It turned out that RAG boosts performance in weaker parametric models by more than 35%. The authors describe this as a “knowledge-equalisation” effect, with retrieval partially compensating for gaps in a model’s parametric regulatory knowledge. This effect is less noticeable in stronger parametric models [17]. These positive findings from this study show strong RAG potential. The same study also identified significant limitations, which will be elaborated in 2.12 of this literature review.

Further advantages identified by Gokhan et al. [18] demonstrate that RAG-based retrieval meaningfully improves answer completeness relative to parametric generation, particularly for questions requiring information assembled from

multiple regulatory provisions. This capability seems compliance-specific: regulations frequently explain a single obligation in multiple articles, and a system must retrieve and synthesise all relevant passages to produce a complete and legally accurate response.

2.7 LLM capability for regulatory-text interpretation

Even without the use of RAG, the literature demonstrates that LLMs can be rather capable in interpreting regulatory text. Kumar and Roussinov [19] evaluate GPT-4's ability to detect inconsistencies and contradictions within regulatory documents using a curated corpus in which ambiguities were deliberately injected. Compliance engineers validated the results which show meaningful precision and recall in identifying regulatory contradictions. In one of the very few studies applying LLMs to regulation in the financial sector, Cao and Feinstein [20] demonstrate that GPT-4 can extract information from Basel III capital-requirement texts and produce mathematically executable representations of regulatory formulas with meaningful accuracy. This means that even in specific domains with high complexity LLMs show useful analytical capabilities that make them viable candidates for compliance-assistance functions.

2.8 Hybrid retrieval for regulatory corpora

A consistent positive finding in the retrieval component is the superiority of hybrid retrieval. The hybrid form combines dense semantic search with sparse keyword-based methods. This approach is better than either approach alone. Rayo et al. [21] demonstrate this explicitly in a system designed for regulatory text, showing that hybrid retrieval significantly outperforms pure dense or pure sparse retrieval on regulatory question answering. The intuition is straightforward: regulatory text contains both conceptually complex passages that benefit from semantic understanding and precisely defined technical terms (article numbers, defined terms, threshold values) that must be matched exactly. Neither dense nor sparse retrieval alone captures both requirements. This has direct architectural implications for any compliance system operating over corpora that combine conceptual interpretation with precise definitional language.

2.9 End-to-end compliance automation

AI reshapes how compliance checks can be executed. Modern models can be given the full regulatory picture, and they will reason accurately over the entire

obligation context at once. This approach consistently beats the older pipeline-based classification and information-extraction methods, that process obligations in independently classified fragments. Furthermore, the capacity of LLMs to explain why a provision is or is not compliant, is identified as a large leap forward in qualitative capabilities of compliance automation in which documented decision rationales are required [8, 22].

Downsides and limitations of RAG

2.10 RAG does not eliminate hallucination

Even though the use of RAG can significantly reduce hallucination, it does not fully eliminate it.

RAG-based legal-research tools were found to produce unsupported or incorrect claims between 17% and 33% of the time, depending on the product. Even though RAG performed better than baseline GPT-4 without retrieval, the significant amount of hallucination that remained proves to be too high for unsupervised deployment in applications where the stakes are high [23].

That LLMs are designed to hallucinate, is explained by Xu et al. [14]. Results from computational learning theory were used to argue that no computable LLM can correctly respond to *all* possible inputs. There will always exist queries for which the model produces inconsistent outputs regardless of the quality of retrieved context. This is a structural problem of LLMs that cannot fully be solved by changing training-data or improving retrieval. Therefore it is important to note that RAG is a strategy that can mitigate hallucination, but cannot reduce it to zero. Human oversight remains necessary.

2.11 Retrieval quality does not guarantee generation quality

Apart from issues with retrieval and LLM hallucination, the generation part of a RAG system can fail by producing incorrect answers. Agrawal et al. [24] empirically demonstrate that this can happen even when retrieved passages contain the correct facts. The generator fails in applying the legal logic contained in the retrieved text. This is especially relevant for compliance purposes, where an LLM can for example misinterpret a conditional exemption as a general permission. The resulting answer will then be dangerously misleading.

Further failure points in RAG and their respective solutions are identified by Barnett et al. [2], by drawing on three applied case studies. The following seven

parts can make a RAG system fail:

[2]

- FP1: The system answers questions it cannot support with evidence. Solution: Use semantic caching for common queries and make the system explicitly answer “unknown” when documents do not contain an answer.
- FP2: Relevant documents are ranked too low, so answers are missed. Solution: Carefully tune chunking, embeddings, retrieval and consolidation, add metadata like file and chunk IDs, consider open source embeddings for short text, and rely on continuous, runtime calibration.
- FP3: Retrieved documents with the answer are dropped during consolidation. Solution: Adjust consolidation and context-size strategies so key evidence is kept, and monitor the system to catch recurring losses of important documents.
- FP4: The answer is in context but not extracted. Solution: Use larger or better-structured contexts, add metadata to highlight key passages, and prefer domain-adapted, end-to-end trained RAG pipelines with suitable embedding models.
- FP5: The model ignores requested output formats. Solution: Improve prompts and testing, and harden the system against jailbreak prompts that override format and safety instructions.
- FP6: The answer is at the wrong level of detail. Solution: Refine prompts and domain configuration, provide clear examples of the desired specificity, and adjust based on real user behaviour.
- FP7: The answer is incomplete despite relevant context. Solution: Enhance retrieval and context construction, and validate with realistic question-answer tests that all key points are included.

2.12 Context distraction and catastrophic regression

As mentioned before in this literature review, Beauchemin and Khoury [17] tested 51 LLMs on 807 multiple-choice questions taken from official Quebec insurance exam handbooks, comparing a closed-book setup against a RAG setup that used a curated corpus of regulatory documents. Their main finding for a compliance RAG system is that retrieval does not always help. For weak models it works as a “knowledge equalizer”, but for some strong models it causes what they call “context distraction”, where adding retrieved legal text makes accuracy collapse.

The clearest case is Gemini-2.5-pro, which scored 74.65 percent closed-book but dropped to 14.06 percent with RAG, worse than random guessing, with similar collapses in Llama-3.3-70b and Grok-2. Importantly, these failures are not normal hallucinations. The first cause was verbosity: models found the right answer but wrote long explanations that did not fit the required answer format, so they scored zero. The second cause was safety refusals, where heavily safety-aligned models read the legal content as a sensitive query and replied with things like "I cannot provide legal advice". On the other hand, RAG strongly helped weaker models, for example DeepSeek-reasoner rose from 36.30 to 71.77 percent, which shows a good reasoning model does not need to memorise the law if it is given the right context. The practical lesson for a Dutch compliance RAG is that you cannot assume retrieval is always an improvement, so you should compare RAG with a closed-book baseline to investigate if retrieval is a net gain for your specific model, corpus, and question type.

2.13 Performance degradation in non-English and multi-regulation contexts

To test how well models handle legal text outside of English, Niklaus et al. [25] built LEXTREME, a legal NLP benchmark made of 11 datasets across 24 languages. They report that even the best multilingual model, XLM-R large, only reached an aggregate score of 61.3, which shows the task is still hard and leaves a lot of room to improve. For a Dutch RAG system this matters because the benchmark shows models perform worse on non-English legal text, so weaker results on Dutch regulation should be expected rather than seen as a surprise.

In "Large Legal Fictions", Dahl et al. [26] measured how often large language models hallucinate when asked specific, checkable questions about real court cases. They found hallucinations are very common, ranging from about 58% with GPT-4 to 88% with Llama 2, and the rate is not the same everywhere. For lower court jurisprudence hallucination went up. Hallucination also varied by jurisdiction and how well known a case was, so older and less central legal areas were more prone to error. The lesson for a compliance RAG is that the legal knowledge in the model's trained parameters is likely to be unreliable and uneven, which is a strong reason to ground answers in retrieved source text.

2.14 Standard evaluation metrics are ill-suited to regulatory text

Relevant literature commonly presents metrics such as Exact Match, F1 over token overlap, ROUGE, and BERTScore to assess RAG performance. However, these metrics seem to be unsuitable for compliance applications. Gokhan et al. [18] demonstrate this explicitly, introducing the RePASS metric as an alternative that evaluates whether a generated answer captures all relevant regulatory obligations without introducing contradictions. Previously mentioned metrics can score a response highly even if it gives a partial answer that is legally incomplete and therefore non-compliant. In a PRISMA-compliant systematic review of 128 RAG studies, Brown et al. [13] confirm that most evaluations in the field remain dominated by Exact Match/F1 over token overlap. Brown et al. call for evaluation frameworks that are explicitly domain-appropriate in compliance contexts. Lawrence et al. [27], on the other hand, propose a risk-aware evaluation methodology distinguishing wrong advice delivered with high confidence from safe errors such as refusals or low-confidence responses. This distinction is highly relevant in regulated contexts, where an overconfident incorrect answer is much more dangerous and harmful than an appropriate refusal.

Zooming in on the financial sector: regulatory positions and evidence gaps

2.15 Authoritative regulatory positions

The most authoritative views on using LLMs in financial compliance come from international regulators, and they consistently point to both the promises and the risks, a blanket statement used by authorities left and right. Even Dick Schoof, former prime minister of the Netherlands, said in a speech at the AI Impact Summit in New Delhi that “AI offers huge opportunities, but also poses great risks” [28].

The Financial Stability Board (FSB) offers a more useful insight. The FSB reports that generative AI and LLMs are mostly used today for internal operations and for summarising compliance documents and retrieving information, but that the rules governing these uses are still limited. The FSB names model risk, data quality, hallucination, and explainability as the main AI-related threats to financial stability, and it asks national authorities to check whether existing frameworks, including those for anti-money laundering, market conduct, and prudential requirements, are strong enough to handle LLM-specific risks [29]. The European Central Bank (ECB) notes that euro-area banks adopt generative

AI more slowly than US banks, especially for compliance work, citing concerns about explainability and data privacy [30].

Beyond these advisory positions, the European Union has now put binding rules in place through the AI Act (Regulation (EU) 2024/1689), which moves the discussion from supervisory opinion to hard law [31]. The AI Act takes a risk-based approach and sets duties around transparency, record-keeping, and human oversight. All of which are relevant for the auditability and oversight standards a DNB-supervised trust office must meet. A compliance-research tool of the kind built here however, would not automatically count as high-risk, since that label is reserved for listed uses such as credit scoring and recruitment, but the system still falls under the AI Act’s transparency rules and its obligations for general-purpose AI models, because it is built on an LLM. The timeline matters for any deployment claim: the bans on prohibited practices have applied since February 2025 and the general-purpose AI model obligations since August 2025, while the main high-risk deadlines were pushed back under the AI Act Omnibus agreement to December 2027 for Annex III systems [32]. The AI Act also states that it works alongside existing Union financial-services law rather than replacing it, so a trust office would have to comply with both regimes at once. For this thesis, the AI Act confirms that a working system cannot be judged on accuracy alone, since it must also be explainable and keep a human in the loop to be lawful, which supports the human-supervised design that is applied in this thesis.

2.16 Domain-specific empirical studies and the lack thereof

Even though (peer-reviewed) work applying RAG specifically to named financial-compliance regulations is rather sparse, the RIRAG study by Gokhan et al. [18] mentioned earlier in this literature review can be considered somewhat directly relevant for the application of RAG in financial-compliance regulations. The ObliQA dataset mentioned in the research, with 27,869 questions drawn from financial-regulation documents, is the largest public financial regulatory QA dataset so far, and its RePASs metric is built specifically to test whether an answer covers all obligations completely. Another domain-specific empirical study by Wang et al. [33] shows a further maturing field. The COLING 2025 Financial Regulations Challenge explained in the research is seen as the first major shared task that benchmarks language models on financial-regulatory understanding across nine task types. The authors conclude however, that LLM performance on regulatory interpretation has been badly understudied. One of the few studies that does exist on a named prudential regulation, by Cao and Feinstein [20], elements of which are mentioned earlier, sets a performance baseline for LLMs

on the Basel III financial regulatory framework, while noting that the lack of a retrieval component limits how useful static prompt-based approaches are in practice.

2.17 Evidence gaps

A conclusion that can be drawn from all literature mentioned earlier in this review, is that it seems that no peer-reviewed study has tested a RAG system on AML tasks using real or representative regulatory corpora. No published benchmark exists for the Wwft, the Wft, or DNB supervisory guidelines, and by extension none exists for the Wtt 2018 that governs trust offices. The only widely available financial regulatory QA dataset, ObliQA, is built from Abu Dhabi regulatory documents in English, so the closest available resource is far from the EU and Dutch context in both jurisdiction and language. Furthermore, Nie et al. [11] confirm that there is currently no standard benchmark for financial regulatory-compliance tasks and no established framework that tests to what extent the obligations have been met. The annotation bottleneck described by Chen et al. [34] explain why: building labelled compliance QA datasets needs expensive legal-expert time (to keep a human in the loop), and the economics of academic research make large-scale, jurisdiction-specific annotation unlikely without an institutional partner. This has a direct effect on the present study, because small, expert-annotated evaluations may be the only realistic option for a Dutch system, and the evaluation must be designed to get the most signal from a limited annotation budget.

2.18 Note on the evidence base

The gaps mentioned above point to a clear limitation with large effects. Peer-reviewed work applying RAG specifically to named financial-compliance regulations is rare, and dedicated research on RAG combined with AML obligations is effectively absent from the peer-reviewed literature as of 2026. For this reason, the discussion above draws on the best available mix of arXiv preprints, grey literature from international regulators, and related studies in nearby regulatory fields. This thin evidence base is not a side issue but a core constraint on what this thesis can build on, and it is part of what makes a Dutch proof of concept worth pursuing.

Chapter 3

Market research

Practical Relevance: Why no existing product closes the gap

3.1 Purpose

The literature review identified that research on RAG in Dutch financial-regulatory compliance remains thin. This academic gap motivates to close this gap by attempting to build a system that applies RAG to financial-regulatory compliance for trust offices in the Netherlands. Trust offices however, do not exist in a vacuum of technological innovations, and have likely attempted to work with AI. Furthermore, new or existing companies may have worked on solutions that are available in the market. Therefore, can a trust office simply buy an off-the-shelf product that helps them to use RAG for their regulatory-compliance questions? Is there any product readily available that can be purchased that does exactly what this thesis aims to do? Or does nothing of the sort exist? In the following market research the requirements that follow from the trust office's own situation are set out, and an attempt is made to find out whether or not the market meets these requirements. The identified products are explained, the fit with the requirements is assessed, and their shortcomings are highlighted. A conclusion is drawn in order to state if the purpose of this thesis, to build and deploy a proof-of-concept of such a RAG system, is justified.

NB: The vendor information below was gathered from company websites, press releases, and analyst summaries, supplemented by a small number of industry market reports. Most of the capability descriptions therefore come from the vendors themselves and should be read as claims rather than as independently verified facts. No weight should be placed on any single vendor's self-assessment.

3.2 Stakeholder requirements

Together with a representative of Peak Compliance B.V., a subsidiary of Pan-Invest B.V., a trust office based in The Hague, the Netherlands, requirements for purchasing a useful AI-based tool were identified, if the tool would be a product that is not custom built for the firm:

1. Current off-the-shelf LLMs such as Chat-GPT or Claude lack consistency in output quality, specificity, accuracy, auditability, and therefore usefulness. A custom made AI-based tool should solve those shortcomings.
2. Off-the-shelf LLMs operate as a 'black box' in the cloud that require a large amount of trust. One can never be completely sure that sensitive data is fully protected. One can also not be sure that the LLM does not use the trust office's data to train the model. Therefore a custom made AI-based tool should be as confidential, local, transparent and therefore trustworthy as possible.
3. It takes significant efforts from an expert to carry out compliance research. A custom made AI-based tool should reduce the amount of human efforts needed to complete quality compliance research.
4. Regulation and application thereof frequently changes. An AI-based tool should rely on a highly up-to-date knowledge base of regulations, supervisory guidance, and jurisprudence. It should therefore be easy to maintain.
5. Human based work inevitably introduces room for errors. An AI-based tool should help to reduce the error margin of compliance research.
6. No two specialised experts are likely to have the exact way of working, writing style, or output quality. An AI-based tool should increase consistency in provided services.
7. The human efforts needed, the lack of consistency and the margin for error make it hard for the compliance services to be scalable. The final overarching requirement is that an AI-based tool should enable more scalability of compliance services.

It is important to note here that the most important and least flexible requirement of the trust office is confidentiality of data. The trust office works with sensitive client data that contains information about the client or the client's company. If this information is made public or shared with third parties, both the client and the trust office are at risk. Therefore, the trust office representative is very hesitant in uploading sensitive data anywhere but in a private, controlled and safe environment. In comparison, the other requirements are not as rigid.

3.3 How potential companies were identified

A trust office looking for an existing product that meets their requirements, would inevitably rely on web searches, recommendations by peers and advertisements aimed at them. In order for products to be identified in this market research, they need to have some kind of visibility. Thus, web searches were carried out with relevant keywords such as ‘AI’, ‘compliance’, ‘Wwft’, ‘Wtt’, ‘AML’, ‘RAG’, and more, or a combination of those keywords. Furthermore, multiple LLMs were given the task to search far and wide for relevant products, after the requirements were explained. The resulting products were each assessed to compare their specifications with the requirements. No products were purchased, nor tested, due to the limited expert knowledge of the researcher to carry out useful tests, as well as the limited amount of time and resources the trust office’s stakeholder made available.

3.4 What the market offers

The platforms that come closest divide into three groups, and the trade-off between them is consistent enough to be the organising fact of the whole market.

Group 1: Compliance-specific platforms (products built specifically for regulatory work)

- <http://4crisk.ai/> offers a conversational assistant running on private specialised language models with human review built in, and advertises built-in coverage of DORA, MiFID II, GDPR, and AMLD; it seems to be, based on its own description, the closest single match to the use case, though Dutch-specific regulation would have to be ingested manually, which can dramatically alter the performance of the system. Also, it is not entirely clear how uploaded (sensitive) client data is handled, stored and protected.
- [FinregE](#) advertises the deepest regulatory functionality of the group. A chatbot built over a large body of regulatory sources, with explicit policy-mapping and gap analysis, which could be addressing a large part of the requirements. Unfortunately it’s entirely cloud-based with no offline capabilities, so confidentiality concerns are not solved.
- <https://squirro.com/> sits slightly apart: it is a general enterprise platform rather than a highly specialised compliance product. It is however European, it deploys fully on-premise, and it publishes a case study with a national financial supervisor in which all processing stayed inside the client’s environment. This solution is therefore strong on privacy, but weaker on ready-made regulatory content.

- <https://cube.global/solutions/cube-regplatform> and <https://www.corlytics.com/> are powerful at monitoring regulatory changes and mapping subsequent obligation updates, but are oriented toward analysts and dashboards rather than ad-hoc conversational queries via a chatbot.
- **ACE/RegAI** is particularly interesting. It's an Amsterdam based firm who offers services that are explicitly attuned to the Dutch regulatory landscape. It appears to be delivered as a consulting engagement rather than as a standalone product a firm can simply run, therefore it does not fit the requirements.

Group 2: Information-services companies with a track record

- **Wolters Kluwer**, a Dutch-headquartered global provider of professional information, software, and services for experts in healthcare, law, tax, and finance. It offers a genuine RAG-based product for legal compliance, but it's largely oriented toward the US market or toward legal professionals, rather than compliance advisors in the Dutch market. Therefore, their EU financial-regulation depth is secondary, and their deployment models are predominantly cloud-based.
- <https://www.lexisnexis.com/supportandtraining/au/lexisnexis-regulatory-compliance> RegCompliance+ platform and Protégé AI assistant demonstrate a mature, commercial approach to navigating complex regulatory obligations through AI-driven search and conversational querying. Its integration of GraphRAG further validates the product's appearance here. However, the platform is built around international and common-law jurisdictions and does not cover the Dutch-specific legislation. Moreover, its cloud-hosted SaaS model is fundamentally incompatible with the local-deployment and data-confidentiality requirement.
- **Thomson Reuters' CoCounsel** platform demonstrates a comparable commercial approach to AI-assisted compliance research, combining agentic workflows, document analysis, and regulatory monitoring through its ONE-SOURCE+ and Westlaw-grounded research capabilities. Its use of RAG-like retrieval grounded in proprietary legal content further confirms the architectural relevance of this thesis. However, like LexisNexis, Thomson Reuters' compliance content is centered on U.S., UK, and international jurisdictions, with no coverage of the Dutch-specific legislation. Additionally, CoCounsel operates as a cloud-hosted SaaS solution with processing handled by third-party models such as OpenAI, which conflicts with the on-premise, data-confidentiality requirement central to this proof-of-concept.

Each of the products in group 2 have genuine RAG-based offerings and deep legislative content, but their compliance products are largely oriented toward US markets and legal professionals rather than compliance advisors in the financial sector. Combined with non-local deployment, these are serious tools aimed at a different buyer.

Group 3: General-purpose RAG frameworks

- <https://www.vectara.com/> is a fully managed RAG-as-a-service platform with solid security, data privacy, and flexible deployment options, which makes it a relevant reference point for what this thesis is trying to build. That said, it does not come with any regulatory content included, so the entire normative layer covering the relevant regulation would still need to be sourced and fed into the system independently. Since that content pipeline has to be built from scratch regardless, it makes more sense to build a purpose-built RAG system from the ground up, where the chunking strategy, retrieval logic, and query handling can all be tailored to the specific demands of Dutch financial regulation, especially since performance can vary greatly based on language.
- <https://qdrant.tech/> is an open-source vector database built in Europe, with full support for self-hosted and air-gapped deployments, which fits well with the requirement that sensitive client data never leaves the local environment. Just like Vectara, it does not provide any compliance content of its own, so it is better understood as a building block for a compliance system rather than a compliance system in itself. Nevertheless, Qdrant can at most be recommended as a potentially useful part in a RAG system that still needs building from the ground up.

The third group provides exactly the deployment flexibility and document-handling the compliance products lack, and several are European with strong on-premise or air-gapped options. What they do not provide is any regulatory content at all: they are the means to build a compliance system, not a compliance system in itself.

3.5 The gap

Laying the three groups side by side reveals a pattern that is more informative than any individual product:

- The compliance-specific platforms (group 1) have the regulatory content but constrain deployment and rarely support combining that content with an uploaded client document.

- The information-services companies with a track record (group 2) have content and maturity but are aimed elsewhere.
- The general-purpose platforms (group 3) have the deployment flexibility and the document-handling but no regulatory content.

No single product holds the regulatory depth, the private deployment, the client-document handling, and the Dutch-specific coverage at the same time. Of these, the Dutch dimension is the most important argument. None of the investigated platforms pre-loads the law that actually governs a Dutch trust office or the clients they serve that have to follow Dutch law. The Wtt 2018, the Wwft, and DNB's supervisory material are not standard content in any of these products; at best they would have to be ingested by hand, which is precisely the part of the work the literature warns is hardest to get right and the part where errors carry the most liability. The surveyed products are built for banks, insurers, and large corporations. The trust office, with its specific obligations under the Wtt 2018, is not a market any of them is designed around. A Dutch trust office is therefore looking for a product positioned for a sector the market does not really serve, in a jurisdiction the market does not pre-cover, under a deployment constraint that the best-matched products do not meet. Furthermore, the investigated products vary greatly in price (from the low tens of thousands of euros per year into the hundreds of thousands for a custom build, indicative). Cost matters a great deal to a small office and that, if anything, strengthens the case for a lean, self-hosted approach. Also when the first step here is to prove whether or not the concept of RAG works in this application, a 'free' thesis-motivated build cannot be passed on.

3.6 From market gap to research justification

The commercial picture aligns with the academic one. The literature review concluded that no published work characterises RAG performance on Dutch financial-regulatory text, and the market review concludes that no product is built to operate over it under the specific trust office's constraints. The two gaps are the same gap seen from two sides, and together they make a strong case for building something for this specific use-case. The literature already shows cautiously that RAG can be useful in this area, but it is vital to note here that creating a RAG system with specific expertise on Dutch financial compliance aimed at trust offices, is not at all a simple task that will easily be good enough to be useful.

3.7 Conclusion

The conclusion is not that the commercial tools are weak, but that none of them line up with the particular combination of constraints a Dutch trust office works under. That misalignment, rather than any technical shortcoming in the products, is what justifies building and studying a proof of concept.

Chapter 4

Methodology

4.1 Purpose of this chapter

Since both the literature and the market establish that a gap exists for an AI-based tool that is specified for financial compliance related to trust offices in the Netherlands, this research aims to find out whether or not closing this gap could be useful for the trust office Pan-Invest B.V. The gap may exist for various reasons, such as the lack of usability, or the inability to compete with existing major tech companies. The literature review concluded however, that retrieval-augmented generation could structurally suit well to compliance work. Its behaviour on non-English, and specifically Dutch, financial-regulatory text has not been characterised though. Furthermore, the market analysis concluded that no available product combines the regulatory depth, the private deployment, the client-document handling and the Dutch-specific coverage that a trust office actually requires. Therefore, by building a dedicated AI-tool on the basis of Retrieval Augmented Generation, it can be better understood if the continuation of such work has any potential.

In this chapter a deep dive is taken into the following questions:

1. Research design: what kind of study is this, and why does this design fit the problem?
2. Use case: what requirements and considerations governed the work?
3. The corpus: what regulatory material was the system given to work with?
4. System design: How is the system constructed and why were design choices made?
5. How were the outputs evaluated?

The proof of concept as a result of this study aims to provide empirical contribution to the field, but since this contribution relies on just one use case, each of the choices made should be deliberate and defensible. All elements, parameters and design choices are therefore explained in terms of the use case, the constraints, or the evidence in the literature that motivated it.

4.2 Research design

This study uses a design-oriented, single-case approach. The output is a proof of concept RAG system that was designed and built, and given the name “RAG4Compliance”. How well the RAG system works was both qualitatively and quantitatively evaluated by a representative of the Dutch trust office Pan-Invest B.V., who is also the director of its subsidiary compliance oriented company Peak Compliance B.V. This research can therefore be classified as ‘design science’, since design science research has been defined as “research that invents a new purposeful artefact to address a generalised type of problem and evaluates its utility for solving problems of that type” [35]. Since the behaviour of retrieval-augmented generation on compliance services of a Dutch trust office is found to be uncharacterised, the natural next step is to build an instance of it and describe, in detail and against fit-for-purpose criteria, how it actually performs.

It is important to state that this study analyses the system’s behaviour on one defined set of compliance questions for one defined use case, and not on a population of users or institutions. The findings can therefore not be statistically generalised to all Dutch trust offices. Instead, this study characterises how the application of a technique with large potential like RAG, can be implemented in one carefully built system to potentially aid in the core obligations of a Dutch trust office. Secondly, the evaluation is purposely qualitative-dominant. No automated scoring metric can simply capture the perceived usefulness, completeness, and correctness in one number, just like no machine can score how delicious a cake is (or isn’t). The most important question for a compliance tool is not whether an answer is deemed correct by calculating formulas, but whether a qualified professional judges it complete and correct enough to act on. The expert-completeness orientation of the evaluation is also consistent with the recommendation in the regulatory-RAG literature that obligation coverage is the property worth measuring [18].

Furthermore, the literature review establishes that automated metrics commonly used to evaluate retrieval-augmented systems are of limited and unvalidated reliability for non-English text [25, 26]

4.3 Use case

The starting point for the design is rooted in the strong desire of a specific firm, Pan-Invest B.V., to extract the potential of AI in an effective and safe way for daily tasks. The firm was already familiar with the potential usefulness of AI for compliance research, but was unable to ensure responsible usage by its employees. Workers would for example translate sensitive documents using free online tools, risking confidentiality. Furthermore, online chatbots such as ChatGPT and Claude would answer compliance questions with a good amount of accuracy, but lacking auditability. During a semi-structured interview with a representative of the trust office, these pain points were identified by the representative. During the same interview, the requirements for a custom system to solve these pain points were also discussed. The resulting requirements summarise the outcome of this interview (Annex A).

From that exercise, seven requirements were understood, taking into account that these requirements specifically apply to a custom built system as opposed to an off-the-shelf system. They are best interpreted as a single coherent set rather than a checklist, because together they define the particular combination that the market review found no existing product to satisfy.

1. The most important requirement, and the least negotiable, is private and controlled deployment: a trust office handles client ownership structures, source-of-funds documentation, its own integrity-risk assessments, and works with internal policy documents that are valuable for competitive firms. All of this is precisely the material that may not be sent to a third-party model provider without compromising client confidentiality and the firm's own position.
2. The second requirement states the necessity of a well maintained, current knowledge base of regulation, supervisory guidance and case material. The trust office operates under dense, complex regulation and it offers its deep expertise of this regulation by providing compliance services. The firm is therefore under high pressure to deliver watertight compliance advice without inaccuracies or outdated information. Any assisting tool should be highly up to date.
3. Third is that this knowledge base can be accessed via a conversational interface, so that an adviser can ask a question in natural language rather than navigate a document database.
4. Fourth is that the system should accept an uploaded client document during a query, so that a specific structure or file can be assessed against the regulatory material in one place.

5. Fifth is that the system should be fully auditable, so that information that is given can be traced back to actual documents, rather than it being potentially made up.
6. The sixth requirement is that the system should be able to perform a gap analysis, comparing a client's situation against applicable obligations.
7. Finally, the seventh requirement is that the system keeps a human in the loop, since, as the literature establishes beyond reasonable doubt, no retrieval-augmented system can be trusted to give unsupervised compliance advice [23, 14]. The design described in the remainder of this chapter is the attempt to satisfy as many of these requirements as a simple, self-hosted system can, and the evaluation is the attempt to establish how far that attempt succeeds.

4.4 The corpus

Since a retrieval-augmented system relies on retrieving information from a corpus, the value of the system is limited by the quality of what it can retrieve. How the corpus is constructed is therefore a methodological decision in its own right rather than a rigid, given set. Due to the high familiarity of the trust office representative with relevant documents, a list was created to serve as a guideline. In this guideline three different groups of documentation were listed, differentiated by legal authority:

1. Binding regulation: primary legislation and the implementing decrees that carry the force of law.
2. Supervisory guidance: the normative but non-binding material, such as Q&As and Good Practices, through which the Dutch central bank (DNB) explains how the open norms of the legislation are to be interpreted and applied.
3. Jurisprudence: enforcement decisions and court rulings that show how the norms have been interpreted and sanctioned in practice.

A fourth layer is part of the corpus the system retrieves from, but documents in this layer are not permanently stored, therefore it's listed separately. This is a transient, per-session layer holding any document the user uploads for the duration of a single conversation, which allows a concrete client file to be assessed against the regulatory material without that file ever becoming part of the permanent store.

All documents used in the system, the organisations they belong to, the dates and a general idea of their contents can be found in Annex B.

One aspect of how the corpus was prepared has direct influence on the results and is recorded here as a deliberate, known constraint. This aspect is the handling of statutory source files. Dutch legislation is published in a structured XML format in which the legal text is interleaved with a substantial volume of structural markup. Ingesting that markup directly and without stripping would pollute the vector representations and it would make real statutory passages retrieve less effectively. The system is therefore designed to strip all tags from the XML and retain only the human-readable text nodes before the text enters the rest of the pipeline. Unfortunately, this cleaning step is imperfect: if a source file is malformed, the system falls back to reading the raw file, thus the occasional passage consisting largely of publication metadata rather than article text can still enter the index, a behaviour observed and discussed in the results.

4.5 System design

The pipeline of the RAG4Compliance system is in line with the design of Lewis et al. [1]: documents are ingested, divided into chunks, embedded and indexed. A query then retrieves the most relevant chunks, and a generator produces an answer grounded in those chunks. The entire pipeline runs on local infrastructure. Several specific choices in this pipeline are shaped by two important constraints: (1) the system should work especially well with Dutch documents, queries and responses, and (2) the system should run privately. For reference, the system runs entirely on a 2021 Apple MacBook Pro, with an 8-core M1 Pro chip, and a 14-core integrated GPU.

4.5.1 Ingestion, cleaning and chunking

In order to keep the ingestion manageable, all documents enter a single pipeline. The system is designed to handle documents in the four most common formats for regulatory, supervisory and judicial documents: XML, PDF, plain text and Word. After the tag-stripping for XML files described in the step above, each document is segmented into overlapping chunks using a recursive character-based splitter configured with a target size of 1,200 characters and an overlap of 300 characters. Chunks are split preferentially at paragraph, then line, then sentence boundaries. The fixed target size is a bit of a compromise: a chunk must be large enough to contain a coherent provision together with enough surrounding context to be interpretable, but small enough that retrieval can be precise and that several chunks fit within the generator’s context window. The 300-character overlap guards against the most damaging failure of fixed-size segmentation: a single

RAG4Compliance

Hierarchy-aware retrieval for Dutch financial-compliance research

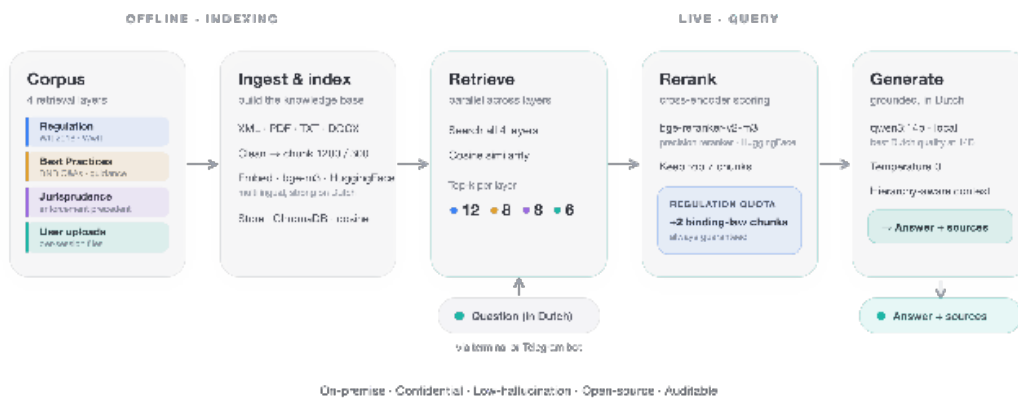


Figure 4.1: Pipeline overview

obligation may be split across a boundary after which neither resulting fragment makes sense by itself. The overlap then ensures such an obligation appears intact in at least one chunk. This configuration is not aware of the article structures of underlying legislation, and that limitation is acknowledged here as a known cost of the approach, consistent with the literature’s finding that retrieval quality is materially sensitive to indexing and chunking configuration [36].

```
1 CHUNK_SIZE      = 1200
2 CHUNK_OVERLAP   = 300
3
4 def build_splitter() -> Any:
5     # ...
6     return RecursiveCharacterTextSplitter(
7         chunk_size=CHUNK_SIZE,
8         chunk_overlap=CHUNK_OVERLAP,
9         length_function=len,
10        separators=["\n\n", "\n", ". ", " ", ""],
11    )
```

Listing 4.1: The chunking configuration (config.py, ingestion.py)

4.5.2 Embedding and storing

After the texts are segmented, each chunk is converted into a dense vector representation using the BGE-M3 embedding model, with embeddings normalised so that cosine similarity is the operative distance measure. The BGE-M3 model was picked because of the high importance of the Dutch language in this system. For this language fewer digital text, annotated datasets, and language technologies are available, making it harder to train accurate and robust models compared to high-resource languages such as English. It therefore remains structurally under-represented in multilingual NLP resources and benchmarks, which typically leads to lower performance than for English. Consequently, the embedding and retrieval components of a pipeline degrade on lower-resource languages and multilingual embedding models retain a residual gap on Dutch even where they outperform English-only alternatives, which the literature establishes [37, 38]. Against this background, the retrieval component in this system adopts BGE-M3, a recently proposed multilingual text-embedding model that is explicitly designed for information retrieval in more than 100 working languages and input lengths up to 8,192 tokens. Chen, J., Xiao, S., et al. [39] show that BGE-M3 (M3-Embedding) achieves (near) state-of-the-art results on a range of multilingual, cross-lingual, and long-document retrieval benchmarks by unifying dense, sparse and multi-vector retrieval within a single architecture. This makes it a strong candidate for Dutch-centric retrieval, without limiting it to just solid performance in the Dutch language. Although Dutch remains less resourced than English,

recent Dutch embedding work (MTEB-NL / E5-NL) also uses BGE-M3 as a competitive multilingual baseline, indicating solid zero-shot performance on Dutch despite the general resource gap [38]. In light of this evidence, BGE-M3 offers a better-justified choice than English-only or clearly English-first embedders, which are optimised primarily on English data and can therefore be expected to show less reliable behaviour on Dutch documents.

The resulting vectors are then stored in ChromaDB, a lightweight vector store that persists to local disk and requires no separate database server. This satisfies some important deployment constraints identified earlier: it runs in-process and on-premise, it is simple to operate without dedicated infrastructure, and it does not create a dependency on an external service. Crucially, the four corpus layers are held in four separate ChromaDB collections rather than in a single index, which is what makes the authority-aware retrieval described next possible.

4.5.3 Retrieval, reranking and the regulation quota

The retrieval stage is divided into two parts: a fast first stage, and a precise second stage. Together they can achieve what neither of those stages can achieve alone. In the first stage, the system queries all four layers in parallel and retrieves a fixed budget of candidate chunks from each: twelve from the regulation layer, eight each from the best-practices and jurisprudence layers, and six from the uploaded-document layer, if any document is present. The asymmetry of these budgets is deliberately implemented to mimic how the compliance expert would work. Binding regulation carries the most weight, because it has the highest authority. The difficult statutory language retrieves less easily than the text in the guidance documents, because the latter is written more like a conversation. Therefore, the chunks from the binding regulation layer are the hardest to surface. As a result, this layer is given the largest first-stage budget, in order to boost the chances of chunks appearing in the final selection. In the second stage, the first-stage candidates are re-scored by a cross-encoder reranker, BGE-reranker-v2-m3. A cross-encoder reads each candidate together with the query instead of just comparing independent embeddings, which yields a significantly more accurate relevance estimate. This retrieve-then-rerank arrangement is the configuration the regulatory-retrieval literature recommends [21]. The model was chosen because it shares the multilingual design of the BGE-M3 embedding model described earlier, which matters given the documented degradation of retrieval in Dutch [38]. The system then keeps the seven highest-scoring chunks whose score clears a rerank floor, which is a minimum relevance score below which a passage is treated as too weakly related to keep. This floor is set at 0.10 on the reranker's scale, which has a range of 0 to 1. The purpose here is to keep weakly related passages out of the context, where they would dilute the evidence and introduce claims without grounds. Because a strict floor could in principle reject every candidate,

the system applies a keep-at-least rule, meaning a guaranteed minimum number of chunks (1) that are retained regardless of score. Retaining at least one chunk prevents an empty context, which would otherwise force a blind refusal or invite unsupported generation. The weak retrieval is recorded as a warning in the system’s logs; no signal about retrieval confidence is passed on to the generator itself.

```

1 RETRIEVE_K = {                                # Per-layer retrieval budgets
2     "regulation": 12,
3     "best_practice": 8,
4     "jurisprudence": 8,
5     "session": 6,                             # Uploaded documents -- per-chat temporary layer
6 }
7 TOP_K      = 7                                # Final chunks after reranking

```

Listing 4.2: The first-stage retrieval budgets per layer and the final selection size (config.py)

```

1 MIN_RERANK = 0.10                            # Lower bound on the rerank score (bge-reranker-v2-m3, scale 0-1)
2 MIN_KEEP   = 1                               # Always keep at least this number of best-scoring chunks
3
4 # Filter on rerank score (more reliable than cosine for Dutch text).
5 kept = [item for item in scored if item[3] >= MIN_RERANK]
6 if len(kept) < MIN_KEEP:
7     kept = scored[:MIN_KEEP]
8 kept = kept[:TOP_K]

```

Listing 4.3: The rerank floor and the keep-at-least rule (config.py, retrieval.py)

This second stage also incorporates an important design choice in the retrieval pipeline, which is that a guaranteed quota for binding regulation is set. After reranking, the system always adds up to two extra regulation chunks that pass a lowered relevance threshold (0.05) and were not already retained, even when the selection already contains regulation-layer chunks. As a result, the final context can hold up to nine chunks rather than seven. The motivation is that explanatory, conversation-like guidance documents tend to retrieve easier than complex statutory text. If left uncorrected, this tendency produces answers grounded in guidance about the law rather than in the law itself. Rayo et al. [21] describe this countermeasure and it is therefore implemented.

```

1 REG_QUOTA      = 2                            # Number of regulation chunks (Layer 1) that may always be added
2 MIN_RERANK_REG = 0.05                        # Lowered rerank threshold for the regulation quota
3
4 kept_ids = {id(item[0]) for item in kept}
5 reg_extra = [
6     item for item in scored
7     if item[2] == "regulation"
8     and item[3] >= MIN_RERANK_REG
9     and id(item[0]) not in kept_ids
10 ]

```

```
11 for item in reg_extra[:REG_QUOTA]:
12     kept.append(item)
```

Listing 4.4: The regulation quota (config.py, retrieval.py)

4.5.4 Context assembly, prompting and generation

After the two-stage retrieval, retained chunks are assembled into a single context in descending order of authority, regulation first, and each is labelled with its source and layer so that the generator can attribute its claims and auditability is ensured. The assembled context is passed to the generator together with a system prompt engineered specifically against the failure modes the literature documents for this class of system. The system prompt exists both in Dutch and in English, and the applicable prompt is selected based on the query language. Although the system is intended primarily for Dutch use, it must remain usable for English queries. Therefore, a lightweight language-detection step decides which language the answer is produced in. Each incoming question is passed to the langdetect library, and the result selects between a Dutch and an English variant of the system prompt, so that the model is instructed to answer in the language of the question. The logic is deliberately conservative and biased toward Dutch: a question shorter than ten characters, where detection is unreliable, is treated as Dutch by default, any result other than English is also treated as Dutch, and a detection failure falls back to Dutch as well. This binary, Dutch-default design reflects the system's primary use case and avoids the risk of answering a Dutch compliance question in English on the basis of an uncertain detection.

The English system prompt for reference:

```
1 SYSTEM_PROMPT_EN = (
2     "You are a precise financial regulatory assistant. "
3     "Answer the user's question using ONLY the context provided below. "
4     "Do NOT use any factual knowledge from your training data; the context is your only source.\n\n"
5     # ... (layer definitions and rules for using the four authority layers)
6     "Referencing and citation:\n"
7     "- Support every factual claim with a reference to the source document name and the layer "
8     "(e.g., 'According to [dnb-qas-en-good-practices-wwft.pdf, Layer 2]...').\n"
9     "- State a specific article, paragraph or sub-paragraph number ONLY if it literally appears "
10    "in the context. NEVER invent or reconstruct article numbers from your own knowledge.\n"
11    # ... (contradiction flagging and the refusal rule)
12    "Be concise and factual. Avoid padding or speculation. "
13    "Always respond in English, regardless of the language of the source documents.\n"
14    "Use the source name EXACTLY as it appears in the [Source: ...] header. "
15    "Never invent file names or extensions (such as '.pdf')."
16 )
```

Listing 4.5: The English system prompt, abridged (prompts.py)

This prompt instructs the model to:

- Answer using only the supplied context and not its parametric knowledge
- Never state an article or paragraph number unless that number appears verbatim in the context, which targets the fabricated-citation failure mode known to be characteristic of legal and financial generation [23, 15]
- Enumerate all applicable obligations rather than a subset, which targets the obligation-incompleteness failure mode [18]
- Attribute every factual claim to a named source; and to flag explicitly any apparent contradiction between layers.

Next, generation is performed by the Qwen3:14B model run locally through Ollama at a temperature of zero. Temperature is a setting that controls how much randomness the model uses when it chooses each next word. At higher values the model varies its wording and can be more creative, but it also becomes less predictable. At zero, the model always picks the most likely option, so the same question produces the same answer every time. For a compliance tool this is the safer choice: it makes the system consistent and reproducible, and it avoids the creative variation that could introduce unsupported or invented content.

The choice of this model matters most in the generation stage, and for the present use case Qwen3:14B is the obvious option, starting from the task. The system must read and reason over dense Dutch regulatory text and produce faithful, grounded answers in Dutch, so the most important quality is multilingual competence on a lower-resource language, not English reasoning. The literature establishes that performance on Dutch rises with model scale and that the capability ceiling of compact open models on Dutch sits well below that of larger ones [40, 41]. This sets a clear direction: within the limits of the available local hardware, the model that brings the most capacity to Dutch is preferred. Qwen3:14B was chosen as the largest model that remains practical to run locally on the available hardware while offering significantly better Dutch quality than smaller alternatives, and its training is explicitly multilingual [42]. Another choice could have been gpt-oss 20b, the free local model from OpenAI. Since OpenAI is a company well-renowned with university students, one may confidently pick this model due to positive experiences with the online ChatGPT interface. This model also carries a much larger headline parameter count, but still it is the far weaker fit for two reasons. First, its larger number is misleading: gpt-oss 20b is a mixture-of-experts model that activates only about 3.6 billion of its parameters per token, whereas Qwen3:14B is a dense model that engages all of its parameters, so on the measure that matters (the capacity actually applied to a Dutch token) Qwen3:14B is the larger model despite the smaller name. Second, gpt-oss was trained with a strong English and reasoning focus and shows clear weakness outside English, while Qwen3:14B is built for broad multilingual coverage. For a Dutch task this difference points strongly toward Qwen3:14B. Taken

together, the use case and the evidence make Qwen3:14B the clear choice. Finally, running the generator locally through Ollama, rather than performing an API call to an online model, is the technical expression of the confidentiality requirement: this is what allows the system to be described as fully on-premise. It should still be stated as a constraint that no Dutch-language regulatory benchmark exists for this specific model, so its suitability is inferred from general Dutch-capability evidence rather than demonstrated on the target task.

4.5.5 Interfaces

The system can be used via two interfaces. The first is a terminal interface in PyCharm or another IDE in which a question is typed and the answer (together with the retrieved source chunks and a summary of which layers contributed) is printed locally. The second is a Telegram bot that offers the exact same functionality, but then through a familiar chat interface, including the ability to upload a client document into the fourth layer and a command to display the sources behind the most recent answer. Telegram delivers each incoming message to a program through a webhook, which is simply a web address that Telegram sends every new message to. This means the program must have a public address that is reachable over the internet and secured with HTTPS. The system, however, runs on a local machine that has no public address of its own. To bridge this gap, the tool ngrok is used. ngrok creates a temporary public HTTPS address and forwards any traffic sent to it back to the local machine, which gives the local program a public front door without placing it on a server. Behind that address, a small local web server built with Flask receives each message, passes the question to the RAG pipeline, and sends the generated answer back through Telegram. The bot also supports simple commands, for example to display the sources behind the last answer or to upload a document for the current session. One point should be made clear. Because messages pass through Telegram's servers and the ngrok tunnel, the Telegram interface does not keep data on the local machine. It was used as a convenient way to test and demonstrate the system, while the terminal remains the fully private, on-premise route. It is therefore included for the primary purpose of testing the output quality of the system without the trust office representative having to be in the same location. Also, the trust office representative has little programming knowledge and is therefore very unfamiliar with the interface of an IDE or a terminal. To accommodate, a more familiar interface is chosen in order to make sure the output quality is judged purely based on its contents, and not on the way it is presented or the interaction that preceded. The Telegram interface, however, differs greatly in terms of confidentiality: the terminal interface keeps all processing on local infrastructure, whereas the Telegram interface routes message text through Telegram's servers and a public tunnel. The latter should therefore be understood solely as a

demonstration interface.

4.6 Evaluation design

The literature explains various ways in which performance of a RAG-system can be evaluated. Multiple performance indicators exist that aim to assess whether or not the output of RAG-systems is likely to be accurate. For this use case, none of those metrics apply. This is mostly caused by an asymmetry of field knowledge. The trust office representative is an expert in the field of financial compliance and this knowledge requires years of studying and in-depth experience. Therefore, the output of the RAG4Compliance system can not be self-assessed. For example, an architect can design a building for a client, but the architect can not decide for the client whether or not the building is to the client's taste. Instead, the architect should ask the client for their opinion. Similarly, for this use case, the trust office representative is required to give feedback on the system's performance, and is therefore asked to provide a list of system inputs that are representative for the intended use. The representative provided ten highly representative test questions that can be used to assess if the system's answers are good enough to be useful and honest enough about their limits to be safe. Each of the questions targets a core obligation of a Dutch trust office and together they span the principal areas of the firm's regulatory work: customer due diligence, the determination of ultimate beneficial ownership in complex structures, the identification of and enhanced measures for politically exposed persons, the indicators of unusual transactions, the obligations triggered by a sanctions-list match, the integrity-risk analysis required under the Wtt, the reporting obligation to the Financial Intelligence Unit, the verification of source of wealth in complex international structures, the application of the risk-based approach to a high-risk client, and the signals that should prompt a review of ongoing due diligence.

Then, for each question, a set of expected elements was defined in advance. For example, the specific obligations, conditions and steps that a complete and correct answer would have to contain. These expected elements are included to provide structure to the evaluation, so that it is done against a fixed, pre-registered standard rather than a loose impression formed after the fact. In addition, a reference answer was generated for each question by a closed-book frontier model (Claude Opus 4.8) without the use of RAG, used to compare the RAG4Compliance system's output against a highly performing model that the representative was already familiar with. The system's answers themselves were produced under two successive configurations, allowing the evaluation to function as a comparison between iterations rather than a single snapshot: an initial configuration and a refined configuration incorporating the cross-encoder reranker, the regulation quota and revised prompts described above. The changes were made on the basis

of the feedback that the trust office representative provided on the first iteration.

The assessment then was done by a domain expert: a compliance professional at the participating trust office, who reviewed each answer against its expected elements and recorded a structured judgement. This is the same person that the requirements were created with. The judgement had three components. The first was a score on a six-point scale running from 0 to 5, anchored with explicit descriptors: 0 for an answer that is incorrect or irrelevant, 1 for "very incomplete" (only one element correct), 2 for "incomplete" (several elements missing), 3 for "good" (most elements present, slight gaps), 4 for "complete" (all elements correct and logical), and 5 for an answer that is complete, legally accurate and accompanied by sound reasoning. The second was a binary Go or No-Go judgement, expressing whether the answer would be safe to rely on in practice and translating the graded score into the operational decision that matters to the firm. This binary decision is intertwined with the 0-5 score, since a score lower than 3 would be a No-Go. The third element of evaluation is a written piece of text for each question recording what the answer covered well, what it potentially should have included, and what risk it would carry if the answer was acted upon.

4.7 Key Performance Indicators

The key performance indicators used to judge the system's practical value were not drawn from a running deployment but from the trust office representative directly. In a telephone conversation on 19 June, the same representative who defined the evaluation questions provided estimates, based on operational experience of the firm, of the internal-cost, output and hiring effects that a supervised tool could deliver. These estimates were written up as a short summary, which the representative reviewed and confirmed, and which is included as Annex C. They are treated throughout as informed estimates from a single source rather than as measured results, and are reported with that limitation made explicit.

Chapter 5

Results

5.1 Overview of the evaluation

This chapter reports how the RAG4Compliance proof of concept performed on the evaluation as described in the previous chapter. It presents what output the system produced and how the trust office representative interpreted and assessed it. The interpretation of these outcomes, their meaning for the research question, and the comparison with the literature is left to the discussion and the conclusion.

To shortly recap the information laid out in the methodology chapter: the evaluation used ten test questions from different areas of the real day-to-day compliance work the representative selected as representative. Each question targets a core obligation of a Dutch trust office, and for each question a set of expected elements was fixed in advance. For every question the system produced an answer, which the representative reviewed against those expected elements and then recorded in three interconnected ways: a score from 0 to 5, a binary Go or No-Go label as a result of the score, and a short written explanation on what the answer covered, what it missed, and what risk the gap would carry. A reference answer generated by a closed-book model (Claude Opus 4.8) was available alongside each system answer as a familiar comparator. The system's answers were produced under two configurations, an initial one and a refined one based on the evaluation of the initial output, so the results also show how the system changed between iterations.

No automated retrieval or generation metrics were computed, in line with the evaluation design, so this chapter reports no Recall@K, NDCG@K, faithfulness or answer-relevance scores. The results are the expert scores, the Go or No-Go labels, the completeness of the answers, and the behaviours that were observed during the review. The chapter presents the scores first, then the difference between the two configurations, then how complete the answers were, and finally the qualitative behaviours that stood out.

5.2 Scores and Go/No-Go outcomes

The ten questions and their outcomes are shown in Table 5.1.

Table 5.1: Expert scores and Go/No-Go outcomes across the ten test questions.

Question	Topic	Score (0-5)	Go / No-Go?
1	Customer Due Diligence (CDD)	4	Go
2	UBO determination in a complex structure	2	No Go
3	PEP and enhanced due diligence	4	Go
4	Indicators of unusual transactions	3	Go
5	Sanctions Act obligations on a match	2	No Go
6	Wtt integrity risk analysis (SIRA)	3	Go
7	FIU reporting obligation	4	Go
8	Source of wealth in complex structures	2	No Go
9	Risk-based approach, high-risk country	4	Go
10	Ongoing monitoring signals	3	Go

Across the ten questions the system scored an average of 3.1 out of 5. The scores fell in a narrow range, from 2 to 4. No answer scored 0, 1 or 5. Seven questions received a Go and three a No-Go, which implies a 70% Go-rate. For clarification: every question that scored 4 or 3 was a Go, and every question that scored 2 or lower was a No-Go, which follows the rule that every score below 3 is treated as unsafe to rely on and introduces risks for the firm if acted upon.

It is important to point out that four questions scored a 4. These were *Customer Due Diligence (CDD)*, *Politically Exposed Person (PEP) and enhanced due diligence*, *the FIU reporting obligation*, and *the risk-based approach*. On these topics the answers were correct, logically built up, and supported by source references that the representative could follow and check. A score of 5 would be desirable but difficult to reach. Such a score would imply an answer so accurate, all-encompassing and good that the trust office representative would have nothing to add or change, which could indicate an almost complete transfer of compliance knowledge to a computer system. Three questions scored a lower but still adequate 3: *the indicators of unusual transactions*, *the Wtt integrity risk analysis*, and *ongoing monitoring signals*. Here the answers were substantively good but each missed a single element from the expected elements, such as the distinction between objective and subjective indicators, or one of the four risk categories in the integrity risk analysis.

The three weakest answers each scored a 2 and were all marked No-Go: *UBO determination in a complex structure*, *the Sanctions Act*, and *the verification of source of wealth*. In these cases the answers were not incorrect in what they said, but they left out elements that are not allowed to be missed. For the question

about Ultimate Beneficial Owners (UBOs), the mandatory fallback to senior management as a pseudo-UBO was absent, and trust-specific roles such as trustee, protector and beneficiary were not named. For the question about the Sanctions Act, the freezing of assets was not presented as the first and primary measure, and the tipping-off prohibition (the duty of confidentiality regarding any unusual transaction that you report to FIU) was missing altogether. Regarding the question about the source of wealth, the answer stayed a bit too general and did not name the concrete verification documents, such as annual accounts, legal opinions and bank statements, that would need to be requested.

Two questions illustrate the two ends of this range. At the strong end, the question on the topic of Risk-based approach resulted in a strong answer that covered all three expected elements, with only a minor point of attention regarding not explicitly mentioning approval by senior management. This score and evaluation read together indicate that the output was scoring closer toward a 5, but that this score is very hard to reach. The question Customer Due Diligence (CDD) also indicates the existence of a glass ceiling at the score of 4. The representative judged that even though the output fully covered all six expected components and that the second refined answer was the stronger of the two, the output was not given a score of 5 because the ownership and control structure was addressed implicitly rather than as a separate step. At the weak end however, the Sanctions Act answer shows a clear No-Go. The expected core elements were the freezing of assets, the report to the supervisor, and the prohibition on informing the client. The trust office representative wrote that both attempts missed two of the three core elements. The process of assessing possible sanctions matches (true positive versus false positive) and the duty to report to the competent authority were correctly described, but the freezing of assets and economic resources was only mentioned to a limited extent and not as an immediate primary measure, including the prohibition on making funds, goods or services available. Also, the tipping-off prohibition was entirely absent. The answer was therefore scored a 2 and marked No-Go, with the recorded risk that failing to freeze or tipping off the relationship would be a serious breach of the Sanctions Act that carries criminal liability and high fines. Acting solely on this system output would therefore be very dangerous.

5.3 Comparison of the two configurations

As described above, the RAG4Compliance system was run in two configurations. The initial configuration used the basic retrieval and filtering setup. The outputs were handed to the trust office representative and were subsequently evaluated. Based on this initial review, a refined configuration of the Python script was created. The refined configuration added the cross-encoder reranker, because

apparently useful chunks were not taken into consideration. Also the regulation quota was implemented, because this layer showed up too little in the resulting sources, even though it certainly contained useful information. Furthermore, clearer source titles were implemented to avoid potential pollution, and the prompts were rewritten to tighten the generation part. It's important to note here that the scores in Table 5.1 reflect the refined configuration.

The refined configuration proved itself to be useful in all questions, since none of the outputs in the second attempt decreased the quality of the answers. On the contrary, most answers were slightly to significantly more complete. The clearest positive change appeared on the PEP question. Under the initial configuration the system did not produce a substantive answer whatsoever and instead returned the fallback message that it could not find an answer in the available documents, stating that the context contained no binding regulation on the topic. This response is for a good part responsible for the changes made in the refined configuration, because the trust office representative explained that the corpus certainly contained the answer, or at least major parts of it. Fortunately, under the refined configuration, the system then produced a full answer that covered four of the five expected elements. The Customer Due Diligence answer also significantly improved: the refined configuration referred explicitly to the duty to carry out the check before entering into the relationship and to the retention obligation, and its source references were easier to follow. On several questions the two configurations shared the same gap rather than differing, for example on the UBO question, where both versions missed the pseudo-UBO fallback. Taken across the ten questions, the refined configuration produced answers that were always equal to or more complete than the initial configuration, and none of the refined answers was weaker than its earlier version, proving that the changes made were sensible and effective.

5.4 Completeness of the answers

Because the evaluation scored answers against a fixed list of expected elements, it is possible to report how completely each answer covered those elements. On three of the ten questions the answer captured every expected element: Customer Due Diligence, the FIU reporting obligation, and the risk-based approach. On the remaining seven questions the answer captured some but not all of the expected elements.

The size of the gap varied and maps directly onto the Go and No-Go split. On four of the seven partial answers only a single expected element was missing, therefore these answers were still judged safe enough to use as a starting point and were marked Go. These were the PEP question, where the twelve-month run-off period after a person leaves office was not mentioned; the unusual-transactions

question, where the split between objective and subjective indicators was not made explicit; the integrity risk analysis, where the service or product risk was not named as its own category; and ongoing monitoring, where the addition of new entities was folded into general structural change rather than treated on its own. On the other three partial answers several elements were missing at once, and these were the three No-Go answers on UBO determination, the Sanctions Act, and source of wealth. In short, the system reached full coverage on three questions, near-full coverage with one missing element on four, and partial coverage with several missing elements on three.

5.5 Observed behaviours

Above, the output was judged as an element in isolation: none of the other elements relevant for performance or experience were taken into account. Therefore, this paragraph will go over other displayed elements of the system's performance, such as the difference of retrieval between layers, the fabrication of source details, the difference in performance with English queries, and the comparison with the closed-book online LLM. Examples are also given.

5.5.1 Uneven contribution of the binding-regulation layer

The layer holding the primary legislation did not always contribute similarly to all questions, even though this layer has the highest authority. On the Customer Due Diligence question, where the relevant terms appear in the text of the Wwft, the statutory layer surfaced and the refined answer cited specific articles of the act. On some other questions, however, the statutory layer added little. On the unusual-transactions question the system stated in the first configuration that "the specific indicators are not stated in [BWBR0024282_2026-01-01_0.xml, Layer 1]", after which it based its answer on supervisory guidance and jurisprudence instead. On the PEP question the initial configuration retrieved nothing usable from the binding-regulation layer and returned its fallback message, replying that it "could not find an answer to this question in the available documents" and that "the context contains no binding regulation (Layer 1) or uploaded document (Layer 4) that explicitly defines when someone qualifies as a PEP". Due to the nature of the questions and the vast amount of information available in layer 1, a retrieval of zero useful chunks from this layer was highly unlikely. The contribution of the regulation layer was therefore significantly improved in the second attempt.

5.5.2 Residual fabrication of regulatory specifics

As indicated in the literature, the use of RAG can significantly reduce hallucination, but it does not fully eliminate it [23]. This makes sense, because the use of RAG changes what the model can see and what it bases its output on. The generator remains a next-word predictor, predicting the most likely outcome. Like any prediction, it can be wrong. Therefore, as expected, basing the answers on retrieved passages significantly reduced made up content but did not fully remove it. In a small number of answers the system stated specific article or paragraph numbers that were, in fact, not present in the retrieved passages, and on one occasion it referred to a source filename that did not match any retrieved document. This happened in question 4 (indicators of unusual transactions), in which the answer cites “jurisprudentie-bunq-2022.pdf” in the refined output. This exact document title does not exist in the corpus, but another document, titled “beslissing-op-bezwaar-bestuurlijke-boete-wwft-bunq-b-v.pdf”, contains the information both answers refer to. This means that the model invented a tidier, plausible-sounding filename instead of using the real source label. The system fortunately displays the sources behind each answer, so the representative was able to see which passages were actually used and could therefore see the invented title.

5.5.3 Relevant retrieval, incomplete answer

On some questions the retrieval brought back relevant material that the generated answer then failed to use in full. The retriever can provide the right passages, but the generator still decides what to put in the answer. It will optimise for a fluent, coherent, response that feels complete, not for a checklist. When the relevant rule is a small part of a longer passage, or when the answer already reads as finished, the model can simply skip it. So a provision that was retrieved can still be missing from the final text. This occurred in question 3 (PEP and enhanced due diligence). In this question the correct chunk (15/66) from “nvb-standaard-politically-exposed-persons-nl-herzien.pdf” was retrieved, in which it is stated that the continued application of enhanced measures for at least twelve months after a person leaves their prominent public function, the so-called run-off period, should be applied. However, the answer did not contain this passage and the trust office representative therefore flagged this as a missed element. On the questions that required pulling together a layered ownership structure, such as UBO determination and source of wealth, the retrieval returned little of substance and the answer was subsequently thin. In one UBO case the answer was built partly on passages that were not the right ones.

5.5.4 Behaviours in the Dutch-language setting

The behaviours above appeared while the system was operating entirely in Dutch. The uneven retrieval of complex Dutch statutory text and the invented Dutch legal references are the concrete forms this took during the review.

5.5.5 Comparison with cloud tools

As a practical point of reference, the same ten questions were also given to a cloud tool, Claude Opus 4.8. Claude was faster than the local system and gave more concrete answers with clearer follow-up steps. When asked for sources, however, Claude pointed only to its training data, which cannot be checked. Also, using either tool would require uploading sensitive client data to the cloud. The local system was slower (± 3 minutes response time) and less polished, but it kept all data on the local machine and every answer could be traced back to a retrieved source. This makes perfect sense, because compared to a world-leading model like Claude Opus 4.8, the used Qwen3:14b model has much fewer parameters and can generally be considered 'less smart'. Anthropic does not disclose the amount of parameters in Opus 4.8, but it is safe to assume that it is a multitude of the 14-15 billion parameters in Qwen3:14b.

5.5.6 Slower thinking

The used hardware, a 2021 MacBook Pro 14 inch with an M1 Pro processor and onboard 14-core graphics chip, processes the outputs of an LLM relatively slow compared to the vast amount of computational power Anthropic can use. This results in a waiting time of around 3 minutes for each answer generated by the RAG4Compliance system. No analysis has been done on the amount of time each step takes exactly, but it is safe to assume that the initial retrieval from the local ChromaDB database does not contribute significantly. The second reranking step also does not take a lot of time, since the retrieved sources including cosine and rerank score appear within seconds. Therefore, the generation part takes up most of the 3 minutes thinking about an answer and outputting it word by word.

The trust office representative indicated that the long waiting time compared to Claude Opus 4.8 is a downside of the RAG4Compliance system, but not a fundamental one. The system functions as a tool that assists an expert in compliance research, but the expert remains the one that combines knowledge, expertise and the tool's output into a coherent advice. The entire research process takes a multitude of 3 minutes, and the tool can therefore work in the background while the expert works on something else at the same time.

5.6 Key Performance Indicators

Apart from the individual scores, the proof of concept RAG system was also assessed against some identified key performance indicators that matter to the trust office. Where the previous section reported how good the answers were, this section aims to translate that result into business value. The starting point is the central finding of the evaluation: on seven of the ten representative questions the system produced answers that were good enough to serve as a supervised first draft. It is this first-pass capability, an expert checking and finishing the system's work rather than starting from a blank page, that makes the gains mentioned hereafter possible. Three indicators were considered: internal cost, output, and hiring.

The first indicator is internal cost. A large share of the office's work consists of recurring tasks such as periodic client screenings and new client intakes. This work is carried out by specialists and billed at approximately 100 euro per hour (depending on the client). The internal cost in labour reflects this amount. Because a supervised tool can assist greatly in major parts of this recurring work, the representative estimates that the internal cost of these tasks can fall by roughly 25 to 33 percent. For a small office where expensive human expert time is the largest cost driver, a saving of a quarter to a third on high-volume routine work is meaningful rather than marginal.

The second indicator is potential output. The office's capacity is limited by how much its experts can handle, since demand is higher than that. At the moment one expert can provide recurring compliance services for a maximum of about three clients per week, which is a hard ceiling. With a tool that is good enough to speed up the research and intake stages, the representative expects that instead of three, four new clients per week becomes realistic, an increase of 33 percent. For a small firm this matters because the extra output is won without an extra hire, at a time when experienced compliance staff are scarce, expensive and slow to recruit. Thus, more clients can be served with the same team.

The third indicator is hiring. Pan-Invest is currently looking to recruit a senior compliance specialist, at an internal cost of roughly 92,000 to 110,000 euro per year, a role that is both expensive and hard to fill. If a tool can prove itself and consequently take over part of the routine research, a less experienced colleague could become productive sooner and on a level that is acceptable for sufficient output quality. A junior hire, at around 56,000 to 72,000 euro per year, then becomes a realistic alternative, because the tool helps to reduce the experience gap. That is a difference of roughly 36,000 to 38,000 euro per year, about a third lower, and it also widens the pool of candidates the office can consider, potentially reducing costs for the hiring process. The tool does not replace the expert, but it lowers the level of experience a new colleague needs on the first day.

Two points should be stated plainly about all three indicators. First, these

are estimates from a single proof of concept and the judgement of one experienced representative, not measured outcomes from a system in daily use, so they should be read as an informed value case rather than as proven figures. The RAG4Compliance system should be tested for a prolonged period of time, measuring internal cost, output and hiring results, in order to prove these claims thoroughly.

In conclusion, the business value does not come from blindly running a fully autonomous system. The real value comes from an increase in speed and consistency on recurring work, while expert judgement is kept for the review and the final decision. From that perspective, the value case is quite realistic for a small trust office: lower cost on routine work, more clients per expert, and a cheaper, more feasible next hire.

The figures in this section were provided by the trust office representative during a telephone conversation and are documented, and confirmed by the representative, in Annex C.

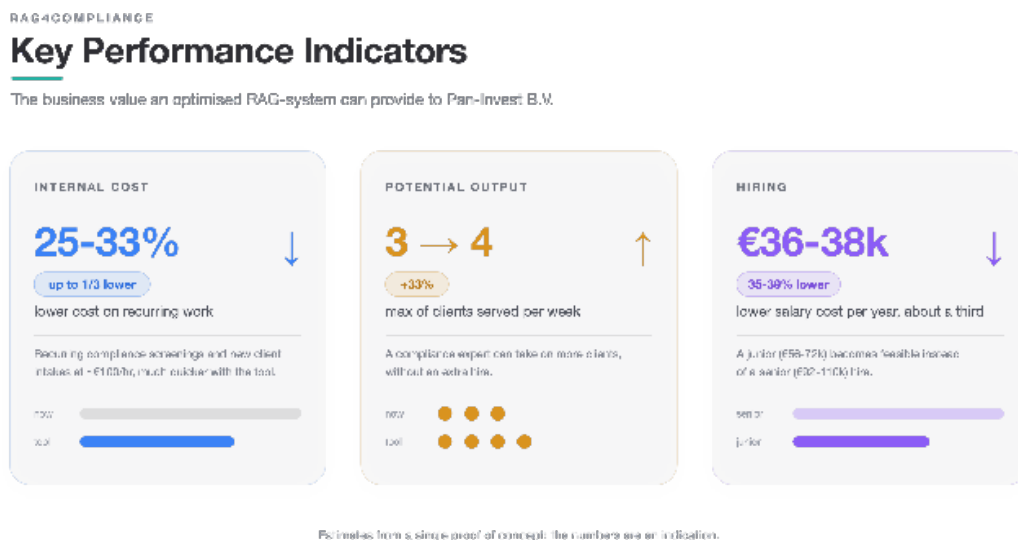


Figure 5.1: Key Performance Indicators

Chapter 6

Discussion

6.1 Purpose of this chapter

The previous chapter elaborated on what the RAG4Compliance system produced and how the trust office representative judged it, but the interpretation of these results are left for this chapter. First, the pattern behind the scores is examined: why did the system succeed on some questions and fail on others? Second, the observed behaviours are placed next to the literature from Chapter 2, in order to establish which documented strengths and weaknesses of RAG were confirmed in this specific Dutch trust-office setting, and which were not. Third, the practical meaning of the results for adoption by the trust office is discussed. Finally, the limitations of the study itself are reflected upon, so that the conclusions laid out in the next chapter can be read with the right amount of caution.

6.2 The pattern behind the scores

With an average score of 3.1 out of 5 and a Go-rate of 70%, the results look moderately positive and somewhat unremarkable. Zooming in on the numbers, a more interesting finding is identified: the split between Go and No-Go was not random. The four questions that scored a 4 (customer due diligence, PEP and enhanced due diligence, the FIU reporting obligation, and the risk-based approach) share an important property. Their obligations are described explicitly, and in retrievable language, in the texts of the corpus. The relevant terms of a CDD question appear literally in the Wwft, so the regulation layer provided useful chunks and thus the answer could cite specific articles of the act. In other words: when the norm is written down where the retriever can find it, in a way that works with the chunk size/overlap settings, the system performs quite well.

The three No-Go questions share the opposite property. For the UBO question, the mandatory fallback to senior management as pseudo-UBO can be found in

implementing decrees rather than in the parent act, and the same holds for trust-specific roles such as trustee, protector and beneficiary. For the Sanctions Act question, the freezing obligation and the tipping-off prohibition were retrieved, but not used in the answer as the primary measures they are. For source of wealth, the concrete verification documents (annual accounts, legal opinions, bank statements) were never named. What these failures have in common is that the complete obligation, i.e. the perfect answer to the question, is spread across multiple documents, or across multiple articles within a document, and that a fixed-length chunking strategy fragments exactly this kind of layered content. It is important to note that the RAG4Compliance system did not answer the test questions incorrectly. Instead, it answered them incompletely, which in a compliance setting can be just as dangerous. Nevertheless, incompleteness can be considered less dangerous than incorrectness.

Furthermore, the distinction between incorrect and incomplete deserves emphasis, because it is the single most important qualitative finding of the evaluation. In all ten questions, no answer scored a 0 or a 1. The representative repeatedly recorded that what the system said was substantively correct, but that elements were missing which are not allowed to be missed. This is precisely the failure mode that Hassani et al. [8] predicted: regulatory obligations are usually explained across multiple clauses or articles, and any approach that analyses regulation in fragments systematically captures too little of the complete obligation scope. The chunking configuration of this system, which is not aware of article structures, is a similar approach, and the results show what downsides it can have. It also confirms that it was sensible to design the evaluation around obligation coverage in the first place. A more traditional token-overlap metric would likely have scored the Sanctions Act answer rather well, since nothing in it was wrong. Only an expert reading against a fixed list of expected elements could catch what was absent, which is exactly the argument Gokhan et al. [18] make with the RePASs metric, and the distinction between dangerous confident errors and safe gaps that Lawrence et al. [27] describe.

6.3 The results in light of the literature

Several key findings from the literature review were directly confirmed by the observed behaviour of the system. The first is that RAG reduces hallucination but does not eliminate it [23, 14]. The system fabricated article numbers that did not appear in the retrieved passages, and on one occasion invented a tidier, plausible-sounding filename ("jurisprudentie-bunq-2022.pdf") for a source that actually carried a much longer name. It is worth noting that the design did its job here: because the sources behind each answer are displayed, the representative could see which passages were actually used and could clearly see the invented

title, without it being able to slip through. This supports the position of Brown et al. [13] that provenance-aware pipelines are a partial but meaningful step forward. The fabrication was not prevented, but it was detectable, and that difference matters a great deal in a supervised setting.

The second confirmed finding is that retrieval quality does not guarantee generation quality [24, 2]. The clearest example is the PEP question, where the chunk containing the twelve-month run-off period was demonstrably retrieved, but the generated answer left it out anyway. This maps directly onto failure points FP4 and FP7 of Barnett et al. [2]: the answer was in the context but not extracted, and the answer was incomplete despite relevant context. The generator optimises for a fluent response that feels complete, not for a checklist, and no amount of retrieval tuning changes that. The generator, on the other hand, could be further optimised to avoid this.

Third, the expected degraded performance on Dutch legal text was observed rather than avoided. The literature establishes that models perform worse on non-English legal text [25] and that multilingual embedding models retain a residual gap on Dutch [38]. The deliberately Dutch-oriented choices in the design, BGE-M3 for embedding and reranking and Qwen3:14B for generation, optimised for the Dutch language focus, but they could not avoid the remaining gap. The difficult language of the binding-regulation layer retrieved less easily than conversational guidance documents, exactly as anticipated in the design of the regulation quota.

Fourth, on the positive side, the improvements between the two configurations confirm the recommendations of Rayo et al. [21]. Adding the cross-encoder reranker and the guaranteed regulation quota turned a complete refusal on the PEP question into an answer covering four of the five expected elements, and none of the ten answers got worse. The retrieve-then-rerank arrangement that the regulatory-retrieval literature recommends earned its place in this system on observable results, not just on theory. In a similar way, the comparison with the closed-book cloud model reflects the trade-off the literature describes: Claude Opus 4.8 was faster and more concrete, but could point to nothing more than "training data" when asked for sources. For a firm supervised by DNB, where auditability is a hard requirement rather than a preference, that answer would not suffice, no matter how good the output is.

6.4 What the results mean for practical adoption

For the trust office, the practical question is not whether the system is perfect, but whether it is useful enough to help as a tool in the daily workload. The results suggest that it is, under one strict condition: constant supervision. Seven of the ten answers were judged safe to use as a starting point that an expert checks and finishes. The three No-Go answers show what happens without a

thorough check, and the recorded risks (criminal liability under the Sanctions Act, enforcement risk on UBO determination) are not risks a trust office wants to carry. The representative's own feedback points in the same direction: the tool works and bases its answers on the ingested material, but the difficult and layered questions still need the expert.

Within that supervised role, the business value case from Chapter 5 is more realistic than speculative, even though the figures remain estimates that need to be proven. A cost reduction of roughly 25-33% on recurring work, an output increase from 3 to 4 new clients per week, and the option of a junior hire instead of a more expensive senior one, are meaningful numbers for a small office where time spent by employees is the largest cost driver. The response time of around 3 minutes was judged as an annoying downside but not a fundamental one, since the tool can work in the background while the expert does something else. However, it should be repeated that these effects were estimated by one representative on the basis of a proof of concept, and that only a longer testing period with daily use can turn them into measured outcomes.

One further point should be made plainly. The literature establishes that hallucination is a structural property of language models that cannot be reduced to zero [14], and the results of this study confirm that even a carefully constrained, grounded, temperature-zero system still invents details occasionally. Consequently, human oversight remains a permanent part of the design. It is therefore important that the RAG4Compliance system is used as a versatile tool during compliance research that merely helps the expert work more efficiently. It should not be seen as a full-fledged adviser that replaces the expert. This is also something that the AI Act requires, with its obligations around transparency and human oversight.

6.5 Limitations of this study

The findings above should be interpreted alongside the limitations of the study itself. The first limitation was stated in the methodology and is repeated here because it bounds everything: this is a single-case design. The system was evaluated on one set of ten questions, for one firm, by one expert. The results characterise how this system behaved in this setting, and they cannot be statistically generalised to all Dutch trust offices.

Second, the evaluating expert is the same person who defined the requirements, selected the test questions, and provided the KPI estimates. The pre-registered expected elements give the evaluation structure and protect against a loose impression formed after the fact, but they do not fully remove the risk that a single involved assessor judges the system more favourably, or more harshly, than a panel of independent experts would. Pan-Invest however, the firm that

cooperated with this study, only had one expert available.

Third, the refined configuration changed several things at once: the reranker, the regulation quota, the source titles and the prompts. The evaluation shows that the combination helped, since no answer got worse and several improved, but the improvements cannot be attributed to an individual change. Similarly, no closed-book baseline was run for Qwen3:14B itself. The comparison was made against a much larger cloud model, which is the comparison that is much more realistic for real world usage, but it means the net gain of retrieval for this specific generator was not measured in isolation.

Fourth, the key performance indicators are informed estimates from a single source, documented and confirmed in Annex C, not measurements from a running deployment. They should be treated as a value case, not as proof. Finally, two known imperfections in the pipeline itself qualify the results: the XML cleaning step occasionally lets publication metadata into the index, and the Telegram interface routes message text outside the local machine and was therefore in light of the confidentiality requirements only suitable as a demonstration route. However, neither undermines the central findings. Both belong in an honest account of what was actually built.

Chapter 7

Conclusion and Future Work

7.1 Answers to the sub-questions

The five sub-questions from Chapter 1 can now be answered on the basis of the previously laid out results.

1. **Technical feasibility.** Yes, to a degree that exceeded expectations for the hardware, expertise and time involved. A fully local pipeline that ingests, indexes and queries Dutch financial regulatory documents in four common formats was built with open-source components and runs entirely on a 2021 consumer laptop. A minor cost of this feasibility is speed: roughly 3 minutes per answer.
2. **Retrieval accuracy.** This is partially achieved, and significantly improved by design choices. The two-stage arrangement of broad parallel retrieval followed by cross-encoder reranking, combined with the guaranteed quota for binding regulation, showed better answer quality between the two configurations. However, complex regulatory text remained harder to retrieve correctly than conversational guidance, and the fixed-length chunking fragmented obligations that span multiple articles.
3. **Regulatory coverage.** This is where the system fell short most clearly. Full coverage of the expected elements was reached on three of the ten questions, near-full coverage on four, and partial coverage on three. The recurring gaps involved norms located in implementing decrees rather than the parent act, and the fabrication of article numbers and one source name showed that hallucination was reduced but not eliminated.
4. **Output quality.** Compared with its own initial configuration, the refined system was never worse and usually more complete. Compared with a closed-book frontier model, the local system was slower and less polished,

but it offered what the cloud model could not: answers traceable to retrieved source passages and confidentiality. Since the closed-book comparator was a different and far larger model, this comparison is indicative rather than a clean measurement of the effect of retrieval alone.

5. **Confidentiality and usability.** Yes, with one qualification. Through the terminal interface, no client data, query or answer ever leaves the local machine, which satisfies the least negotiable requirement of the trust office. The Telegram interface used for remote testing does not meet this standard and was included solely for demonstration. Usability was judged sufficient for a background research tool, with response time as the main practical drawback.

7.2 Answer to the main research question

The main research question asked how the use of Retrieval-Augmented Generation can increase service quality, consistency, and resource efficiency for a trust office when providing compliance services in the Dutch financial sector. The answer is that it seems quite capable of providing useful benefits here.

RAG increases service quality in a specific and limited way: as a supervised first draft. On seven of ten representative questions the system produced answers good enough for an expert to check and finish rather than having to start from a blank page, including source references that could easily be verified. The increase in consistency is also included in the design: generation at temperature zero means the same question produces the same answer every time, which no team of human specialists can match. It increases resource efficiency according to the firm's own estimates, with internal costs on recurring work falling by roughly 25-33%, output per specialist rising by about 33%, and a junior hire becoming a feasible alternative to a senior one.

RAG does not remove the need for the expert. This system will always be limited by the models that it uses. Those limits become larger, not smaller, in a Dutch-language, multi-framework setting, where the models tend to perform worse. To mitigate this, models with better multilingual performance were chosen. Human oversight is therefore a permanent feature of any responsible deployment, not a temporary one. Within that boundary, however, a private, locally deployed compliance assistant seems both feasible and economically valuable, and the particular combination the market does not offer (regulatory depth, private deployment, client-document handling and Dutch coverage) can be built by a single person with open-source tooling on ordinary hardware, underlining that it can not only be done by large corporations.

7.3 Contribution

This thesis contributes the first characterisation of RAG behaviour on Dutch trust-office regulatory text, evaluated on whether answers capture every applicable obligation rather than on wording overlap. It documents, for one carefully built system, where such a system succeeds (explicit norms in retrievable regulatory text), where it fails (norms in implementing decrees, layered structures, fragmented obligations), and which failure modes from the general literature reappear in this setting. For the participating firm, it delivers a working proof of concept and a documented value case. For researchers, it offers a small but honest data point in a field where, as Chapter 2 established, peer-reviewed evidence on RAG for AML obligations is effectively absent.

7.4 Future work

The results point to a clear set of next steps, in rough order of expected impact.

The first is article-aware chunking. The most damaging failure pattern traced back to fixed-length segmentation that ignores the article structure of legislation. A splitter that takes into account the boundaries of articles and paragraphs would potentially solve the completeness problem at its root, and the literature on chunking sensitivity suggests the gain could be substantial.

The second is careful expansion and documentation of the corpus, in particular the implementing decrees where several missed norms turned out to live. The pseudo-UBO fallback and the trust-specific roles were not missed because the generator failed, but because the corpus and chunking did not serve them up. Furthermore, a maintained update process for the index is needed, since a knowledge base that is not updated simply recreates the very problem RAG is meant to solve [11].

The third is hybrid retrieval. The current system is dense-only with reranking. Rayo et al. [21] show that combining dense semantic search together with sparse keyword matching outperforms either alone on regulatory text, where exact terms such as article numbers and thresholds must be matched precisely.

The fourth is a longer measurement phase. The KPI figures in this thesis are estimates from one representative. A prolonged pilot in daily practice, measuring internal cost, output and hiring effects, together with a larger question set, multiple independent assessors, and a closed-book baseline for the same local model, would turn the value case into evidence. Changes should then be introduced one at a time, so that improvements can be attributed to individual design choices.

Finally, there is the path toward production, which was explicitly out of scope here: security hardening, access control, a private web interface to replace the Telegram demonstration route, and integration with the trust office’s existing

workflows. The representative's own suggestions deserve exploration in this phase, including a secured environment that uses a stronger cloud model under a strict process in which sensitive data is never shared, next to the fully local route for the material that may never leave the office.

7.5 Closing remark

This thesis started from the observation that European financial regulation has never been stricter, and that the specialised human work of compliance research does not scale. No pain, no gain, as Chapter 1 put it. The pain of dense, layered regulation is not going away. What this proof of concept shows is that a meaningful part of the gain no longer requires an expensive platform or a leap of faith in a cloud provider. The RAG4Compliance system can run on a consumer laptop with limited power, it can cite its sources, and keep all sensitive documents inside the office by not sending it to the cloud. That is a modest conclusion, but in a field where an overlooked obligation carries criminal liability, modest and honest is exactly what a compliance tool should be.

Appendix A

Requirements interview with the trust office representative

Interview with the trust office representative in which the requirements for the system were laid out.

Date: 17 March 2026

Format: In person meeting

Duration: 3 hours

Participants: Hessel Mullemeister (author) and the representative of the trust office Pan-Invest B.V., referred to by role to preserve the individual's anonymity. This is the same representative who later provided the ten evaluation questions and assessed the system's answers (Chapters 4 and 5).

Purpose: to identify the compliance-research challenges the firm currently experiences, to understand what an ideal solution would look like, and to establish the requirements for the proof-of-concept system developed in this thesis.

Background

A meeting took place between the trust office representative and the author to identify the challenges the firm currently experiences in its compliance-research work and to describe what an ideal situation would look like. The representative indicated that artificial intelligence could offer a solution, but that he was not yet sure how to implement it successfully to reach the firm's goals. The concept of Retrieval-Augmented Generation (RAG) was introduced and explained during the meeting, and it was received with enthusiasm as a promising direction.

Objectives identified for the proof of concept

From the discussion, the proof of concept should aim to demonstrate that Retrieval-Augmented Generation can:

- improve the quality, usefulness, specificity, accuracy and completeness of the language model's output;
- reduce the human resources spent on compliance research;

- reduce the error margin in the services provided;
- increase the consistency of the services provided;
- decrease the trust office's liability for overlooked compliance gaps;
- increase the scalability of output.

Confidentiality requirement

Confidentiality was raised as a major concern, given that the work involves sensitive company data, potential compliance gaps and large sums of money. Sending such data to an external technology company that operates a chatbot was considered unacceptable. The representative therefore stressed that the RAG-based system should ideally process all data locally, so that confidentiality is not put at risk. This requirement became one of the least flexible constraints on the design.

Verification and consent

The representative consented vocally to taking part in this study and to the use of this record.

Appendix B

Documents used in the corpus of the RAG4Compliance system

The overview below lists the documents that form part of the corpus used by the RAG4Compliance system, grouped by source type.

Regulation:

1. Wet ter voorkoming van witwassen en financieren van terrorisme (Wwft) – Netherlands

Organisation: Government of the Netherlands | Date: 1 January 2026 (consolidated version; originally 2008)

Sets out the Dutch anti-money laundering and counter-terrorist-financing framework, including customer due diligence, reporting obligations and supervision of financial and designated non-financial institutions.

Source: <https://wetten.overheid.nl/BWBR0024282/2026-01-01>

2. Uitvoeringsbesluit Wwft 2018 (Implementation Decree Wwft) – Netherlands

Organisation: Government of the Netherlands | Date: 1 January 2025 (consolidated version; originally 2018)

Provides detailed implementing rules for the Wwft, including risk assessment, transaction monitoring and specific due-diligence measures for obliged entities.

Source: <https://wetten.overheid.nl/BWBR0024275/2025-01-01>

3. Wet op de economische delicten (Economic Offences Act) – Netherlands

Organisation: Government of the Netherlands | Date: 1 July 2023 (consolidated version; originally 1950)

Defines economic offences and applicable sanctions, including criminal enforcement for violations of regulatory frameworks such as financial-supervision and anti-money-laundering legislation.

Source: <https://wetten.overheid.nl/BWBR0003296/2023-07-01>

4. Besluit prudentiële regels Wft (Bpr Wft) – Netherlands

Organisation: Government of the Netherlands | Date: 22 April 2021 (consolidated version; originally 2015)

Lays down prudential requirements for Dutch financial institutions under the Financial Supervision Act (Wft), such as capital, liquidity and risk-management standards.

Source: <https://wetten.overheid.nl/BWBR0041190/2021-04-22>

5. Wet op het financieel toezicht (Wft) – Netherlands

Organisation: Government of the Netherlands | Date: 18 July 2025 (consolidated version; originally 2007)

Constitutes the core Dutch financial-supervision statute, covering licensing, conduct and prudential supervision of banks, insurers, investment firms and other financial institutions.

Source: <https://wetten.overheid.nl/BWBR0026204/2025-07-18>

6. Besluit marktmisbruik Wft (Market Abuse Decree Wft) – Netherlands

Organisation: Government of the Netherlands | Date: 4 February 2025 (consolidated version; originally 2016)

Implements EU market-abuse rules in Dutch law, setting detailed obligations on disclosure of inside information, insider lists and reporting of suspicious orders and transactions.

Source: <https://wetten.overheid.nl/BWBR0041193/2025-02-04>

7. Besluit gedragstoezicht financiële ondernemingen Wft (Bgfo Wft) – Netherlands

Organisation: Government of the Netherlands | Date: 1 March 2025 (consolidated version; originally 2007)

Specifies conduct-of-business rules for financial undertakings under the Wft, including customer information duties, product governance and suitability and appropriateness assessments.

Source: <https://wetten.overheid.nl/BWBR0041583/2025-03-01>

8. Directive (EU) 2015/849 (Fourth Anti-Money Laundering Directive, consolidated to 30 December 2024)

Organisation: European Union | Date: 30 December 2024 (consolidated version; original directive of 20 May 2015)

Sets the EU-wide framework to prevent the use of the financial system for money-laundering and terrorist-financing purposes, including risk-based customer due diligence, beneficial-ownership transparency and cooperation between financial-intelligence units.

Source: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02015L0849-2024>

Supervisory guidance / best practices:

1. NVB Risicogebaseerde Standaard – Bron van middelen (SoF)

Organisation: Dutch Banking Association (NVB) | Date: November 2023

Describes banks' risk-based approach to assessing a customer's source of funds during customer due diligence.

Source: <https://www.nvb.nl/media/oksc0lq2/nvb-risicogebaseerde-standaard-bron-van-middelen.pdf>

2. NVB Standaard – Pseudo-UBO

Organisation: Dutch Banking Association (NVB) | Date: May 2023

Describes when and how senior management may be designated as a 'pseudo-UBO' as a fallback where no UBO can be established on the basis of ownership or control.

Source: <https://www.nvb.nl/media/bbkh4zgw/nvb-standaard-pseudo-ubo-nl.pdf>

3. NVB Risicogebaseerde Standaard – EDD-maatregelen EC derde landen met een hoog risico (HRTC)

Organisation: Dutch Banking Association (NVB) | Date: May 2023 (revised)

Describes the risk-based application of enhanced due diligence (EDD) to business relationships and transactions involving high-risk third countries designated by the European Commission.

Source: <https://www.nvb.nl/media/sbzg2paa/nvb-risicogebaseerde-standaard-edd-for-nl-revised.pdf>

4. NVB Risicogebaseerde Standaard – Verwacht Transactieprofiel (ETP)

Organisation: Dutch Banking Association (NVB) | Date: May 2023

Explains how banks use an expected transaction profile to detect deviating or unusual customer transaction behaviour.

Source: <https://www.nvb.nl/media/hlllne10/nvb-standaard-verwacht-transactie-profiel.pdf>

5. NVB Risicogebaseerde Standaard – Actualisatie van klantgegevens

Organisation: Dutch Banking Association (NVB) | Date: May 2023

Describes the risk-based moments, methods and sources banks use to update customer data throughout the customer relationship.

Source: <https://www.nvb.nl/media/eemfqobt/nvb-standaard-actualisatie-van-klantgegevens.pdf>

6. NVB Risicogebaseerde Standaard – Politiek Prominente Personen (PEP's)

Organisation: Dutch Banking Association (NVB) | Date: November 2023 (revised)

Describes the risk-based approach to enhanced due diligence for politically exposed persons (PEPs).

Source: <https://www.nvb.nl/media/x0klygxb/nvb-standaard-politically-exposed-personen.pdf>

7. NVB Risicogebaseerde Standaard – Technische modeldocumentatie

Organisation: Dutch Banking Association (NVB) | Date: June 2024

Sets out principles for the technical documentation of advanced analytics, AI and machine learning models used by banks within the AML/CFT framework.

Source: <https://www.nvb.nl/media/j3fcji2n/20240625-nvb-industry-baseline-tmd-nl.pdf>

8. NVB Risicogebaseerde Standaard – Cash

Organisation: Dutch Banking Association (NVB) | Date: June 2025

Describes the risk-based approach to assessing money laundering and terrorist financing risks associated with customers' cash transactions.

Source: <https://www.nvb.nl/media/uy4l3gsx/nvb-standaard-cash-nl.pdf>

9. NVB Risicogebaseerde Standaard – UBO-identificatie en -verificatie

Organisation: Dutch Banking Association (NVB) | *Date:* May 2023

Describes the risk-based approach to identifying and verifying customers' ultimate beneficial owners (UBOs).

Source: <https://www.nvb.nl/media/hg4iry1t/nvb-standaard-ubo-nl.pdf>

10. NVB Risicogebaseerde Standaard – Modellen voor het genereren van alerts en events

Organisation: Dutch Banking Association (NVB) | *Date:* July 2023

Describes the risk-based use of (advanced) models to generate money laundering and terrorist financing alerts and events within ongoing due diligence.

Source: <https://www.nvb.nl/media/kogiaksy/nvb-standaard-modellen-voor-het-genereren-van-alerts-en-events-nl.pdf>

11. NVB Risicogebaseerde Standaard – Ongoing Due Diligence (ODD)

Organisation: Dutch Banking Association (NVB) | *Date:* July 2023

Describes the risk-based framework for ongoing customer monitoring (ongoing due diligence) by banks.

Source: <https://www.nvb.nl/media/efpmhrq4/nvb-standaard-ongoing-due-diligence-nl.pdf>

12. The FATF Recommendations – International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation

Organisation: Financial Action Task Force (FATF) | *Date:* Adopted February 2012 (periodically updated)

Contains the 40 FATF Recommendations that form the international standard for AML/CFT/CPF measures.

Source: <https://www.fatf-gafi.org/content/dam/fatf-gafi/recommendations/FATF%20Recommendations%202012.pdf.coredownload.inline.pdf>

13. Guidance for a Risk-Based Approach: Life Insurance Sector

Organisation: Financial Action Task Force (FATF) | *Date:* October 2018

Guidance for a risk-based approach to AML/CFT in the life insurance sector.

Source: <https://www.fatf-gafi.org/content/fatf-gafi/en/publications/Fatfrecommendations/RBA-Life-Insurance-Sector.html>

14. Guidance for a Risk-Based Approach: Securities Sector

Organisation: Financial Action Task Force (FATF) | *Date:* October 2018

Guidance for a risk-based approach to AML/CFT in the securities sector.

Source: <https://www.fatf-gafi.org/content/fatf-gafi/en/publications/Fatfrecommendations/Rba-securities-sector.html>

15. High-Risk and Other Monitored Jurisdictions

Organisation: Financial Action Task Force (FATF) | *Date:* Continuously updated (three times a year)

Overview ('black' and 'grey' lists) of jurisdictions with strategic deficiencies in their AML/CFT regimes.

Source: <https://www.fatf-gafi.org/en/topics/high-risk-and-other-monitored-jurisd.html>

16. Guidance on Beneficial Ownership of Legal Persons (Recommendation 24)

Organisation: Financial Action Task Force (FATF) | Date: 10 March 2023

Guidance for implementing the revised Recommendation 24 on the transparency of beneficial ownership of legal persons.

Source: <https://www.fatf-gafi.org/content/fatf-gafi/en/publications/Fatfrecommendations/Guidance-Beneficial-Ownership-Legal-Persons.html>

17. Guidelines on ML/TF Risk Factors (EBA/GL/2021/02, gewijzigd bij EBA/GL/2024/01)

Organisation: European Banking Authority (EBA) | Date: Amending guidelines January 2024, applicable as of 30 December 2024

Guidelines for assessing and managing money laundering and terrorist financing risk factors by financial institutions, including CDD/EDD/SDD measures.

Source: <https://www.eba.europa.eu/legacy/regulation-and-policy/regulatory-activities/anti-money-laundering-and-countering-financing-1>

18. Guidelines on the Role of AML/CFT Compliance Officers (EBA/GL/2022/05)

Organisation: European Banking Authority (EBA) | Date: 14 June 2022 (applicable as of 21 November 2022)

Guidelines on the role, tasks and responsibilities of the AML/CFT compliance officer and the management body of institutions.

Source: <https://www.eba.europa.eu/activities/single-rulebook/regulatory-activities/anti-money-laundering-and-countering-financing-1>

19. Van herstel naar balans – Een vooruitblik naar een meer risicogebaseerde aanpak van het voorkomen en bestrijden van witwassen en terrorismefinanciering

Organisation: Dutch Central Bank (DNB) | Date: 2022

DNB report advocating a more risk-based, proportionate gatekeeper approach by banks to preventing and combating money laundering and terrorist financing.

Source: <https://www.dnb.nl/media/2ambmvxt/van-herstel-naar-balans.pdf>

20. Response to FATF Public Consultation on AML/CFT and Financial Inclusion (R.1, 10 and 15)

Organisation: The Wolfsberg Group | Date: February 2025

Response to the FATF consultation on strengthening a risk-based approach in relation to financial inclusion through Recommendations 1, 10 and 15.

Source: <https://wolfsberg-group.org/resources/rba/199>

21. Statement on the Risk-Based Approach

Organisation: The Wolfsberg Group | Date: July 2025

Statement defining the core elements (proportionality, prioritisation, effectiveness) of a risk-based approach to financial crime risk management.

Source: <https://wolfsberg-group.org/resources/rba/203>

22. Response to the FBAs' Notice of Proposed Rulemaking for the AML Program Rule

Organisation: The Wolfsberg Group | Date: 30 September 2024

Response to the US federal banking agencies (FBAs) regarding the proposed AML Program Rule.

Source: <https://wolfsberg-group.org/resources/effectiveness/182>

23. Response to the Consultation on FinCEN's Notice of Proposed Rulemaking (NPRM) for the AML/CFT Program Rule

Organisation: The Wolfsberg Group | Date: 3 September 2024

Response to FinCEN's proposed AML/CFT Program Rule, with recommendations focused on effectiveness and risk-based implementation.

Source: <https://wolfsberg-group.org/resources/effectiveness/177>

24. Statement on Effective Monitoring for Suspicious Activity, Part I: Moving Beyond Automated Transaction Monitoring

Organisation: The Wolfsberg Group | Date: 1 July 2024

Advocates a risk-based, effectiveness-focused approach to monitoring for unusual and suspicious activity, moving beyond traditional transaction monitoring.

Source: <https://wolfsberg-group.org/resources/innovation/168>

25. Response to FATF Public Consultation on R.16/INR.16

Organisation: The Wolfsberg Group | Date: 26 April 2024

Response to the FATF consultation on the revision of Recommendation 16 and the accompanying Interpretative Note on payment information (the 'travel rule').

Source: <https://wolfsberg-group.org/resources/correspondent-banking/165>

26. Wolfsberg Group on Countering Terrorist Financing

Organisation: The Wolfsberg Group | Date: 20 March 2024

Replaces the 2002 statement and describes the role of financial institutions and risk-based control measures for combating terrorist financing.

Source: <https://wolfsberg-group.org/resources/general/156>

27. Country Risk FAQs (2024)

Organisation: The Wolfsberg Group | Date: 2024

Frequently asked questions on setting up and applying a methodology for assessing financial crime country risk.

Source: <https://wolfsberg-group.org/resources/rba/155>

28. Wolfsberg Group–IIF Joint Response to the FATF's Public Consultation on Risk-Based Guidance on Recommendation 25

Organisation: The Wolfsberg Group / Institute of International Finance (IIF) | Date: 8 December 2023

Joint response with the IIF to the FATF consultation on risk-based guidance regarding beneficial ownership transparency of legal arrangements (trusts).

Source: <https://wolfsberg-group.org/resources/rba/138>

29. Wolfsberg Group Payment Transparency Standards

Organisation: The Wolfsberg Group | Date: 24 October 2023

Revised standards on payment message transparency aimed at preventing abuse of the payments system.

Source: <https://wolfsberg-group.org/resources/correspondent-banking/>
136

30. Wolfsberg Response to EBA Consultation on Revised Guidelines on Money Laundering and Terrorist Financing Risk Factors

Organisation: The Wolfsberg Group | Date: 25 August 2023

Response to the EBA consultation (EBA/CP/2023/11) on revising the ML/TF risk factor guidelines, including for crypto-asset service providers.

Source: <https://wolfsberg-group.org/resources/rba/134>

31. Comment Letter on the EU AML/CFT Legislative Package 2022

Organisation: The Wolfsberg Group | Date: 28 October 2022

Response to the EU AML legislative package (including the establishment of AMLA), with recommendations on harmonisation, information sharing and the risk-based approach.

Source: <https://wolfsberg-group.org/resources/effectiveness/94>

32. Wolfsberg Principles for Using Artificial Intelligence and Machine Learning in Financial Crime Compliance

Organisation: The Wolfsberg Group | Date: 2022

Five principles for the responsible use of AI and machine learning within financial crime compliance programmes.

Source: <https://wolfsberg-group.org/resources/innovation/93>

33. Wolfsberg Group Transaction Monitoring Request for Information (RFI) Best Practice Guidance

Organisation: The Wolfsberg Group | Date: 23 August 2022

Best practice guidance on making and answering requests for information (RFIs) arising from transaction monitoring, particularly in correspondent banking relationships.

Source: <https://wolfsberg-group.org/resources/general/57>

34. Source of Wealth and Source of Funds FAQs

Organisation: The Wolfsberg Group | Date: August 2020

Frequently asked questions on identifying and assessing source of wealth (SoW) and source of funds (SoF), particularly for private banking and wealth management customers.

Source: <https://wolfsberg-group.org/resources/general/32>

35. Consultation on the Draft Guidelines on Business-Wide Risk Assessment (Art. 10(4) Regulation (EU) 2024/1624)

Organisation: Authority for Anti-Money Laundering and Countering the Financing of Terrorism (AMLA) | Date: Opened 16 April 2026 (deadline 15 July 2026)

Consultation on draft guidelines setting minimum requirements for the business-wide risk assessment (BWRA) by entities subject to the AML Regulation.

Source: <https://www.amla.europa.eu/policy/public-consultations/consultation-draft-en>

36. Voorkomen witwassen en terrorismefinanciering – naleving Sanctiewet (themapagina en Leidraad Wwft en Sanctiewet)

Organisation: Dutch Authority for the Financial Markets (AFM) | *Date:* Guidance revised 2024 (webpage continuously updated)

AFM theme page with supervisory information and the Wwft and Sanctions Act Guidance for the gatekeeper role of supervised institutions.

Source: <https://www.afm.nl/nl-nl/sector/themas/voorkomen-witwassen-terrorismedinanciering>
Jurisprudence

1. ECLI:NL:CBB:2026:25 – Boetes wegens overtreding Wwft, Sanctiewet en SIRA-verplichtingen bij een betaalinstelling (nemo tenetur-beginsel bij vrijwillige overlegging van stukken)

Organisation: Dutch Trade and Industry Appeals Tribunal (CBb) | *Date:* 27 January 2026

The CBb rules that voluntary submission of a report does not justify reliance on the nemo tenetur principle, and upholds fines for breaches of the Wwft, the Sanctions Act and SIRA obligations.

Source: <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:CBB:2026:25>

2. Beslissing op bezwaar – bestuurlijke boete Wwft bunq B.V.

Organisation: Dutch Central Bank (DNB) | *Date:* 14 October 2025

DNB's decision on bunq B.V.'s objection to an administrative fine for breach of Section 3(2) of the Wwft (customer due diligence); the objection is declared unfounded.

Source: <https://www.dnb.nl/media/cr2ajxmx/beslissing-op-bezwaar-bestuurlijke-boete-bunq-bv.pdf>

3. ECLI:NL:CBB:2025:372 – Overtreding van de Wwft-meldplicht (art. 16 Wwft)

Organisation: Dutch Trade and Industry Appeals Tribunal (CBb) | *Date:* 15 July 2025

The CBb rules that a first breach of the reporting obligation may be fined, and corrects the mitigation of the fine for exceeding the reasonable time limit (maximum €2,500 for delays of up to one year).

Source: <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:CBB:2025:372>

Appendix C

Record of telephone conversation with the trust office representative

Date: 19 June 2026

Mode: Telephone

Approximate duration: 0 hours, 53 minutes

Participants: Hessel Mullemeister (author) and the trust office representative of Pan-Invest B.V. This is the same representative who provided the requirements described in Chapter 3 and who assessed the system's answers.

Purpose: To obtain the representative's estimates of the business value that a supervised, RAG-based compliance tool could deliver for the firm, expressed as key performance indicators.

Summary of the figures provided:

Internal cost. A large share of the firm's work consists of recurring tasks such as periodic client screenings and new client intakes, carried out by specialists and billed at approximately 100 euro per hour. The representative estimated that a supervised tool taking over the first pass of this recurring work could reduce the internal cost of these tasks by roughly 25 to 33 percent.

Output. One specialist can currently take on a maximum of about three new clients per week. With a sufficiently capable tool to speed up research and intake, the representative expected this to rise to about four new clients per week.

Hiring. The firm is seeking to recruit a senior compliance specialist at approximately 92,000 to 110,000 euro per year. The representative indicated that, if a tool could take over part of the routine research, a junior hire at approximately 56,000 to 72,000 euro per year would become a feasible alternative.

Nature of the figures: These are the representative's informed estimates based on operational experience of the firm, not strictly measured outcomes. They are reproduced here as primary data for the evaluation of the study's key performance indicators.

Verification: The representative reviewed this summary and confirmed it as

an accurate record of the conversation on 19 June 2026.

Consent: The representative consented to taking part in this study and to the use and documentation of these figures.

Appendix D

Testing results with the trust office representative

Table D.1: Full evaluation record: expert review of the ten test questions.

Question	Topic	Test question	Expected elements	Score (0-5)	Findings	Risk if wrong	Go/No-Go
1	Customer Due Diligence (CDD)	What steps must an institution take to carry out customer due diligence in accordance with the Wwft?	Identification, verification, UBO, purpose/nature, risk, monitoring	4	Attempt 2 is the best version. Both attempts fully address all six required components of customer due diligence: identification and verification, UBO determination, purpose and nature of the relationship, risk-based due diligence, documentation and ongoing monitoring. Attempt 2 is stronger because it explicitly refers to the due diligence before entering into the relationship (Art. 4) and the retention obligation (Art. 33). The content is correct, logically structured, and the source references are correct and easy to follow. A point of attention is that insight into the ownership and control structure is not mentioned as a separate step, although it is addressed implicitly.	Incomplete CDD leads to accepting clients without an established risk profile; a breach of the Wwft, an administrative fine/instruction from the AFM/DNB, money-laundering risk and reputational damage.	Go
2	UBO determination	How do you determine the UBO in a complex structure?	Control, contractual control, alternative UBO	2	Both attempts address the main aspects of UBO determination within complex structures, including the >25% threshold, indirect control, actual or strategic control, and verification using registers and documentation. However, the answers are not complete. The mandatory fallback option, whereby senior management is designated as pseudo-UBO when no UBO can be established, is missing from both answers. In addition, specific roles within trust and fund structures, such as trustee, protector and beneficiaries, are not named, even though these are precisely relevant for complex ownership and control structures. As a result, several required elements are only partially covered.	An incorrect or missed UBO means the true beneficial owner remains out of sight; inadequate due diligence (Art. 3(2)/(4) Wwft); increased money-laundering and/or sanctions-evasion risk and enforcement risk.	No-Go
3	PEP-EDD	When does someone qualify as a PEP and which measures are mandatory?	Definition of PEP, EDD measures, source of wealth, senior management approval, monitoring	4	Attempt 1 contains no substantive answer and therefore scores 0. Attempt 2 fully addresses four of the five required elements: the PEP definition including family members and close associates, having an appropriate risk-management system, approval by senior management, investigation of the source of wealth and funds, and enhanced ongoing monitoring. The treatment is substantively strong, but not complete because the statutory 12-month run-off period after a prominent public function ends is not mentioned.	Failing to recognise a PEP or the absence of EDD (no senior approval, no source of wealth) is a breach of Art. 8 Wwft; high corruption/money-laundering risk, fines and reputational damage.	Go

Question	Topic	Test question	Expected elements	Score (0-5)	Findings	Risk if wrong	Go/No-Go
4	Unusual transactions	Name indicators that may point to unusual transactions for a client?	Objective, subjective, examples	3	Both attempts contain relevant and concrete examples of unusual transactions, such as large cash amounts, the use of high denominations, smurfing, round amounts, deviating address details, unclear origin of funds and possible front men. The examples are substantively strong and practically applicable. The main shortcoming is that the distinction between objective and subjective indicators is not explicitly explained. Attempt 1 refers only indirectly to subjective indicators and attempt 2 does not mention this distinction at all. As a result, an essential expected main element is missing, despite the qualitatively good examples.	Failure to report, or late reporting, to the FIU carries a fine and can also have criminal consequences. In the sector there is still often much confusion about when a report must be made, which indicators and factors must be assessed, and that on suspicion a report must be made without delay.	Go
5	Sanctions Act	What must an institution do when a client matches a sanctions list?	Freeze, report, do not inform	2	Both attempts correctly describe the process of assessing possible sanctions matches (true positive versus false positive) and the duty to report to the competent authority. However, the answers miss two of the three core obligations. Freezing assets and economic resources is only mentioned to a limited extent and not as an immediate primary measure, including the prohibition on making funds, goods or services available. In addition, the tipping-off prohibition is entirely absent: it is not mentioned that the person or entity concerned may not be informed about the sanctions match or the measures taken. As a result, two essential expected elements are insufficiently addressed.	Failing to freeze, or freezing too late, and/or tipping off the relationship is a (serious) breach of the Sanctions Act with criminal liability, high fines and international reputational risk. This is a high-risk theme where mistakes may be made.	No-Go
6	Wtt risk analysis	Which risks must a trust office include in the integrity risk analysis under the Wtt?	Client, country, service, transaction risks	3	Both attempts address the main risk categories largely correctly. Client risk, complexity of the ownership and control structure, geographic risk, transaction risk and the presence of a PEP are explicitly named. However, the risk arising from the nature of the specific service or product is not clearly addressed as a separate risk category. As a result, only three of the four expected risk categories are fully developed. Attempt 2 is substantively the strongest because it correctly refers to the Wtt 2018 and Annex III, but full application of the answer requires consulting Annex III.	An incomplete SIRA means there is a chance that integrity risks are not properly and timely controlled. This is a breach of the Wtt 2018 and can lead to an instruction and/or fine from DNB.	Go

Testing results with the trust office representative

Question	Topic	Test question	Expected elements	Score (0-5)	Findings	Risk if wrong	Go/No-Go
7	FIU reporting obligation	When is an institution required to file a report with FIU-NL and what must that report contain?	Indicators, deadlines, rationale	4	Both attempts address the three expected elements fully and correctly. The trigger for a report, including an unusual transaction and situations in which it is only later recognised as unusual, is clearly explained. The obligation to report without delay and the required content of the report, including the rationale and relevant circumstances, are also well covered. Attempt 1 is slightly more complete because it also names the mandatory data for the report and registration with FIU-NL. This attempt does contain a small inaccuracy, however, because a sanctions match in itself is not grounds for an FIU report but falls under the sanctions-law track. Nonetheless, all three expected elements are adequately covered.	Failure to report, or late reporting, to FIU-NL (Art. 16 Wwft) can lead to a fine from the supervisor as well as criminal consequences. In addition, failing to report or reporting late increases the risk of facilitating money laundering.	Go
8	Source of wealth	How do you verify wealth in complex international structures?	annual accounts, legal opinions, bank statements	2	Both attempts describe the context of enhanced due diligence (EDD) well and address relevant points such as the origin of wealth, the background and economic rationale of complex structures, tax aspects and risks associated with high-risk jurisdictions. However, the answers remain too general. The expected verification documents, such as annual accounts, legal opinions and bank statements, are not explicitly named as sources for the investigation. As a result, the concrete specification of which documents must be requested and assessed is missing, leaving an essential part of the answer uncovered.	An insufficiently verified source of wealth leaves the plausibility of a legal origin open; a core EDD shortcoming (Art. 8 Wwft), high money-laundering risk in international structures and enforcement risk.	No-Go
9	Risk-based approach	How do you apply the risk-based approach to a new client from a high-risk country?	Additional documentation, monitoring, EDD	4	Attempt 2 addresses all three expected elements fully and correctly. The answer explicitly describes the application of enhanced due diligence (EDD), requesting additional information about the client and UBO, investigation of the source of funds, and more intensive ongoing monitoring of the business relationship and transactions. The reference to high-risk third countries is also correct and relevant. The treatment is substantively strong, logically structured and practically applicable. As a minor point of attention, approval by senior management is not explicitly mentioned, but this does not detract from the completeness of the answer to the core requirements.	No EDD for a high-risk country means acceptance without appropriate controls; a breach of Art. 8/9 Wwft, strongly increases money-laundering/sanctions risk and enforcement risk.	Go

Testing results with the trust office representative

Question	Topic	Test question	Expected elements	Score (0-5)	Findings	Risk if wrong	Go/No-Go
10	Ongoing monitoring	Which signals during the client relationship can be grounds to review the customer due diligence?	New or changed structures, transaction anomalies, new entities, behavioural changes	3	Both attempts adequately address the main signals for reviewing customer due diligence. Changes in the ownership or UBO structure, deviating transactions, increased risk indicators, outdated client data and the need to revise the risk profile are clearly named. The addition of new entities is included implicitly under changes in the structure, but receives little separate attention. Although this is substantively correct, this aspect could be developed more extensively because the introduction of new entities can be an important signal within the UBO investigation and can give rise to additional customer due diligence.	Missed monitoring signals mean an outdated risk profile and undetected unusual transactions; a breach of the ongoing-monitoring obligation (Art. 3 Wwft) and money-laundering risk.	Go

Testing results with the trust office representative

Table D.2: Scoring rubric used in the evaluation.

Score	Description	Criteria
0	Incorrect	Answer contains errors or is irrelevant
1	Very incomplete	Only 1 element correct
2	Incomplete	Several elements missing
3	Good	Most elements present, slight gaps
4	Complete	All elements correct and logical
5	Excellent	Complete, legally accurate, with reasoning

Appendix E

Hessel's AI Compliance Bot - Additional feedback by the trust office representative

19 June 2026. Written statement by the trust office representative:

"Privacy is important to us when implementing an AI search engine for use in our office. This is partly due to the concern that employees, out of convenience, might enter sensitive information. Pan-Invest is a Dutch trust office, and as such we hold a great deal of (privacy-)sensitive information, including on clients whose structure contains a stock-exchange listing. Because we are legally required to investigate all of our clients, their (UBO) structures, sources of income and the origin of their wealth, and to screen against PEP and sanctions lists, we hold a large amount of information that is not publicly known, or that may not be made public. We need this data in order to comply with the obligations arising from the Wwft, the Sanctions Act and the Wtt 2018.

Because legislation is important to us, because we are supervised by the Dutch Central Bank (DNB), and because we take our role as gatekeeper very seriously, we want our employees to be able to ask data-sensitive questions in a safe way. This is the basis for developing our own bot. This bot takes as its input the legislation, guidelines, best practices and case law that apply to us.

Hessel's tool works: it answers the questions asked and bases its answers on the information the bot has taken in. However, we note the following:

- 1. A relatively simple question took a lot of time, but did give a correct answer.*
- 2. A slightly more difficult question took a lot of time, but gave no concrete and no correct answer. It looked like a combination of text, without any depth or context.*
- 3. A relatively difficult question took a lot of time, but gave a reasonable answer, in which it was suggested that further investigation be carried out.*
- 4. An unstructured question took a lot of time, but nevertheless gave a good answer, with well-formulated follow-up steps.*
- 5. A difficult question took a lot of time and gave an incomplete and non-concrete answer,*

without correct follow-up steps. The context was placed only on the UBO, and not on whether a UBO can even exist. The question indirectly contained two questions, which are nevertheless closely connected to each other. For determining the UBO of a listed (stock-exchange) client, I had expected the answer to be broader than just a focus on the UBO identification obligation, without concrete follow-up questions.

The same questions were also put to Claude AI. The speed of its answers cannot be matched by Hessel's bot, and in addition the answers were concrete, including clear follow-up steps or suggestions. Claude used correct sources, but was much less auditable. A major drawback is the data sensitivity that we cannot expose to such AI tools. That said, I would like to ask Hessel what we can set up. I am thinking of the following:

- 1. Setting up our own environment in which the knowledge input is secured, but which still makes use of Claude's language model. It may be necessary to establish a clear process whereby personal and/or sensitive data may never be shared.*
- 2. Setting up our own environment in which files can be uploaded, including sensitive data, but where there is no processing of data with Claude. This should also include an option to translate sensitive data and/or files."*

Bibliography

- [1] Patrick Lewis et al. *RAG for Knowledge-Intensive NLP*. NeurIPS. 2020. URL: <https://doi.org/10.48550/arXiv.2005.11401>.
- [2] C. Barnett et al. *Seven Failure Points in RAG*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2401.05856>.
- [3] Gupta et al. *Comprehensive RAG Survey*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2410.12837>.
- [4] Kernshell. *How RAG reduces AI hallucinations and improves accuracy*. Kernshell. 2025. URL: <https://www.kernshell.com/how-rag-reduces-ai-hallucinations-and-improves-accuracy/kernshell>.
- [5] GFOSS. *How to reduce hallucinations in RAG systems*. GFOSS. 2026. URL: <https://gfooss.eu/how-to-reduce-hallucinations-in-rag-systems/gfooss>.
- [6] Kairntech. *LLM on-premise: Deploy AI locally with full control*. Kairntech. 2026. URL: <https://kairntech.com/blog/articles/llm-on-premise/kairntech>.
- [7] Zhao et al. *RAG for AI-Generated Content Survey*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2402.19473>.
- [8] Hassani et al. *Rethinking Legal Compliance Automation*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2404.14356>.
- [9] Mökander et al. *Auditing LLMs: Three-Layered Approach*. AI & Ethics. 2023. URL: <https://doi.org/10.1007/s43681-023-00289-2>.
- [10] Lee et al. *Survey of FinLLMs*. Neural Computing and Applications. 2024. URL: <https://doi.org/10.1007/s00521-024-10495-6>.
- [11] Nie et al. *Survey of LLMs for Financial Apps*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2406.11903>.
- [12] Pérez-Cruz et al. *AI Explainability for Regulators*. BIS FSI Paper 24. 2025. URL: <https://www.bis.org/fsi/fsipapers24.pdf>.

- [13] Andrew Brown, Muhammad Roman, and Barry Devereux. *Systematic Literature Review of RAG*. arXiv. 2025. URL: <https://doi.org/10.48550/arXiv.2508.06401>.
- [14] Xu et al. *Hallucination is Inevitable*. arXiv. 2025. URL: <https://doi.org/10.48550/arXiv.2401.11817>.
- [15] Kang & Liu. *LLM Hallucination in Finance*. arXiv. 2023. URL: <https://doi.org/10.48550/arXiv.2311.15548>.
- [16] Roychowdhury et al. *Hallucination-Minimized Framework*. arXiv. 2023. URL: <https://doi.org/10.48550/arXiv.2311.07592>.
- [17] Beauchemin & Khoury. *LLMs for Quebec Insurance Regulation*. Semantic Scholar. 2026. URL: <https://www.semanticscholar.org/search?q=Benchmarking%20Large%20Language%20Models%20for%20Quebec%20Insurance%3A%20From%20Closed-Book%20to%20Retrieval-Augmented%20Generation>.
- [18] Gokhan et al. *RIRAG + ObliQA Dataset*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2409.05677>.
- [19] Kumar & Roussinov. *GPT-4.0 for Regulatory Compliance*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2412.20602>.
- [20] Cao & Feinstein. *LLM in Financial Regulatory Interp.* arXiv q-fin. 2024. URL: <https://doi.org/10.48550/arXiv.2405.06808>.
- [21] Rayo et al. *Hybrid IR+RAG for Regulatory Texts*. arXiv. 2025. URL: <https://doi.org/10.48550/arXiv.2502.16767>.
- [22] Hassani. *Enhancing Legal Compliance with LLMs*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2404.17522>.
- [23] Magesh et al. *Hallucination-Free? AI Legal Tools*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2405.20362>.
- [24] Agrawal et al. *Mindful-RAG: Failure Points*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2407.12216>.
- [25] Niklaus et al. *LEXTREME Benchmark*. EMNLP Findings. 2023. URL: <https://doi.org/10.18653/v1/2023.findings-emnlp.200>.
- [26] Dahl et al. *Large Legal Fictions: LLM Hallucinations*. Journal of Legal Analysis. 2024. URL: <https://doi.org/10.1093/jla/laae003>.
- [27] Lawrence et al. *LLMs in High-Risk Domains*. GenBench/EMNLP. 2023. URL: <https://doi.org/10.48550/arXiv.2311.14966>.
- [28] Schoof. *[Reference details to be added]*. REFERENCE DETAILS TO BE ADDED; speech at the AI Impact Summit, New Delhi; cited in Section 2.15. 2026.
- [29] Financial Stability Board. *Financial Stability Implications of AI*. FSB Report. 2024. URL: <https://www.fsb.org/uploads/R14112024.pdf>.

- [30] European Central Bank. *Rise of AI: Benefits and Risks*. ECB FSR May 2024. 2024. URL: https://www.ecb.europa.eu/press/financial-stability-publications/fsr/special/html/ecb.fsrart202405_02~58c3ce5246.en.html.
- [31] European Union. *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. Official Journal of the European Union. 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- [32] European Commission. *[Reference details to be added]*. REFERENCE DETAILS TO BE ADDED; cited in Section 2.15. 2024.
- [33] Wang et al. *COLING 2025 Financial Regulations Challenge*. arXiv. 2025. URL: <https://doi.org/10.48550/arXiv.2412.11159>.
- [34] Chen et al. *LLMs for Finance, Healthcare, Law*. arXiv. 2024. URL: <https://doi.org/10.48550/arXiv.2405.01769>.
- [35] Venable & Baskerville. *Eating Our Own Cooking: Toward a More Rigorous Design Science of Research Methods*. Electronic Journal of Business Research Methods, 10(2). 2012. URL: <https://academic-publishing.org/index.php/ejbrm/article/view/1296/1259>.
- [36] Bashir et al. *[Reference details to be added]*. REFERENCE DETAILS TO BE ADDED; cited in Section 4.5.1. 2025.
- [37] Banar et al. *[Reference details to be added]*. REFERENCE DETAILS TO BE ADDED; cited in Section 4.5.2. 2024.
- [38] Banar, N., Daelemans, W., et al. *MTEB-NL and E5-NL: Embedding benchmark and models for Dutch*. arXiv. 2025. URL: <https://arxiv.org/abs/2509.12340arxiv>.
- [39] Chen, J., Xiao, S., et al. *M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation*. Findings of the Association for Computational Linguistics. 2024. URL: <https://aclanthology.org/2024.findings-acl.137/aclanthology>.
- [40] Ahuja et al. *MEGA: Multilingual Evaluation of Generative AI*. EMNLP. 2023. URL: <https://doi.org/10.48550/arXiv.2303.12528>.
- [41] Builtjes et al. *[Reference details to be added]*. REFERENCE DETAILS TO BE ADDED; cited in Section 4.5.4. 2025.
- [42] Qwen Team. *Qwen3 Technical Report*. arXiv. 2025. URL: <https://doi.org/10.48550/arXiv.2505.09388>.
- [43] Beijing Academy of Artificial Intelligence. *BAAI/bge-m3*. Hugging Face. Computer software. 2024. URL: <https://huggingface.co/BAAI/bge-m3huggingface>.