HSLU Lucerne University of Applied Sciences and Arts

Language Models in Finance

A Prototype for Compliance Applications

Institute of Financial Services Zug IFZ
www.hslu.ch/ifz

e.foresight

finnova.

inventxLAB
INNOVATION BY DESIGN

Kanton Zug

SFTI

SIX

SWISS BANKERS

Zürcher Kantonalbank

FH Zentralschweiz

Contents

1. Introduction

The rise of artificial intelligence (AI) and natural language processing (NLP) is opening up new opportunities for companies to process information and make decisions. At the forefront of this development are Language Models (LMs), which mimic human behaviour with increasing sophistication. The landscape of LMs spans from Small Language Models (SLMs) to Large Language Models (LLMs). However, their use brings significant challenges, particularly in regulated environments where accuracy, explainability, and data privacy are critical (Bank of England, 2024).

CBinsights (2025) projects that the enterprise AI agents and copilots market will grow over 150 percent year-over-year, reaching USD 13 billion in annual revenue by the end of 2025. In Switzerland, about half of the banks use AI or have initial applications in development according to a survey by FINMA (2025). The "Future of Finance" study by SIX Group (2024) also shows that AI remains a top technology priority with a focus on increasing efficiency and reducing costs.

Among AI applications, compliance automation is emerging as especially relevant. As regulatory demands increase, financial institutions face mounting pressure to ensure effective oversight, for example in monitoring Politically Exposed Persons (PEPs). Due to their influence and roles, PEPs pose an elevated risks for financial crime, corruption, and money laundering, and are therefore subject to enhanced due diligence requirements (CompFidus, 2024). Traditional monitoring methods are time-consuming and prone to human error, requiring financial institutions to explore AI solutions that improve efficiency and accuracy while adhering to strict regulatory standards (Castellum AI, 2024).

AI solutions have the potential to automate screening, reduce manual workloads, and enable real-time detection of PEP-related risks. By cross-referencing internal data with external sources, they improve efficiency, facilitate anomaly detection, and enhance fraud prevention (The Alan Turing Institute, 2024). For instance, if a PEP-linked director is implicated in a corruption case, an LM-based system could instantly flag the event for proactive risk management, helping to reduce regulatory penalties and reputational damage.

However, selecting the right LM for such tasks involves trade-offs. LLMs provide advanced reasoning capabilities but demand significant computational resources and raise data privacy concerns (Priyanshu, Vijay, Kumar, Naidu, & Mireshghallah, 2023). SLMs, by contrast, are more efficient, easier to deploy on private infrastructure, and provide stronger safeguards for sensitive data (IBM, 2024). Also, the decision between public and privately hosted models affects confidentiality, with private models affording greater control at the expense of access to broader general knowledge (Wang et al., 2023). Further complicating these decisions is the probabilistic nature of LMs, which often is in tension with regulatory demands for deterministic and auditable output (Risk Management Association, 2024).

Beyond technical factors, regulatory compliance for AI is essential. In Switzerland, financial institutions are subject to the requirements of the Swiss Financial Market Supervisory Authority (FINMA), which emphasizes the responsible integration of AI and mandates that its use reflects a risk-based approach to compliance management (Swiss FinTech Innovations, 2025). At the EU level, the AI Act classifies financial AI applications as high-risk, requiring strict transparency, governance, and risk management measures (European Union, 2024).

Given this context, the present study takes an initial step towards the integration of AI in banking by exploring how LMs can support critical compliance processes, with a particular focus on PEP monitoring. Specifically, it concentrates on developing and testing a prototype that leverages different LMs to automate the detection of individuals in company leadership roles, a key step in the process of PEP monitoring and risk assessment. The study compares the accuracy, efficiency, and reliability of various LMs in this sensitive compliance task, thereby contributing to the evolving discussion on AI governance in regulated environments. It should be emphasized that all models were evaluated within a standardised pipeline without any task-specific optimisation. Consequently, the reported results reflect performance under uniform conditions and may differ if the individual models were specifically adapted or fine-tuned for the use case.

2. Language Models

The release of models such as ChatGPT[1] has propelled LMs into mainstream use, demonstrating their ability to perform complex language tasks ranging from simple conversations to sophisticated reasoning. Today, a variety of LMs are available, including ChatGPT (OpenAI), Gemini (Google DeepMind), Claude (Anthropic), LLaMA (Meta), Phi (Microsoft), and DeepSeek R1 (DeepSeek AI), each featuring different technical specifications and approaches.

The field of LMs is diverse and models vary significantly in size, capabilities, resource requirements, and deployment modes. This chapter explores these distinctions and introduces the concept of Retrieval-Augmented Generation (RAG), which enhances the capabilities of LMs for data-sensitive and knowledge-intensive tasks.

2.1. Large vs. Small Language Models

LLMs have recently come to dominate the AI landscape owing to their performance across a broad range of tasks.[2] Trained on large datasets encompassing much of the publicly available internet (RedHat, 2024), they are characterised by billions or even trillions of parameters. This capacity enables LLMs to handle complex, multi-step questions, and generate coherent and contextually rich responses. However, LLMs come with substantial trade-offs. Their training and operation demand considerable computational resources, making them expensive and energy-intensive.

In contrast, SLMs are designed to contain fewer parameters. This makes them faster, more efficient, and better suited for specialised applications within defined domains. While SLMs may not achieve the same level of general reasoning as LLMs, they offer significant advantages in terms of cost, speed, and ease of deployment, especially when deployed in environments with limited resources or where specific domain expertise is required (RedHat, 2024).

[1] ChatGPT stands for "Chat Generative Pre-Trained Transformer".

[2] For a discussion of LMs in finance, see Ankenbrand, Bieri, and Gattlen (2025).

2.2. Public vs. Private Language Models

Language models can also be classified based on their deployment approach as public or private models (Clairo AI, 2023).

Public LMs, accessed via external APIs, offer convenience and state-of-the-art performance without the need for local infrastructure. These models are maintained and updated by providers, ensuring they stay aligned with the latest advancements. However, they raise important concerns about data privacy and control. Sensitive information shared with public models may be stored externally, potentially exposing organisations to confidentiality risks.

Private LMs, in contrast, are deployed and managed within secure, controlled environments. Organisations retain full control over data usage, storage, and compliance, rendering private models particularly suitable for sensitive applications such as compliance monitoring. The downside is that private models require significant internal resources for deployment, maintenance, and periodic updates to remain competitive with rapidly evolving public alternatives.

2.3. Retrieval-Augmented Generation (RAG)

Although LMs can generate coherent responses, they are inherently limited to the knowledge encoded in their training data, which may not always reflect the most up-to-date or domain-specific information.

RAG addresses this challenge by integrating LMs with external retrieval mechanisms (Lewis et al., 2020). In a typical RAG workflow, the user's inquiry initially triggers a search through external knowledge sources, such as databases, private document collections, or the web, to retrieve relevant information. These retrieved documents are then combined with the original query and passed to the LM, which generates a comprehensive response grounded in both its internal knowledge and the retrieved external evidence.

3. Compliance

Compliance plays a critical role in maintaining trust and stability in the financial sector. As regulations grow more complex, institutions face increasing operational and financial challenges in meeting legal, regulatory, and internal requirements. Globally, costs related to financial crime compliance exceed USD 200 billion annually (LexisNexis Risk Solutions, 2023). In Switzerland, the increasing regulatory burden is reflected in the latest "FINMA Risk Monitor", which highlights heightened risks in areas such as sanctions, money laundering, and cyber threats, many of which place significant demands on compliance functions (FINMA, 2024). In this context, AI technologies, including LMs, offer new opportunities to support compliance functions in a more efficient and scalable manner.

3.1. Definition of Compliance

Compliance, as a distinct organisational function, plays a central role in safeguarding financial institutions against regulatory, legal, and reputational risks. Regulatory definitions provide an essential foundation for understanding this role and for designing technical solutions that align with established principles. According to the Basel Committee on Banking Supervision (2005), compliance is defined as an independent function within banks that ensures adherence to legal, regulatory, and internal requirements, thereby contributing to risk mitigation and stability. In Switzerland, FINMA (2017) describes compliance in similar terms, emphasising its role as a core control function responsible for ensuring conformity with regulatory obligations and internal policies.

These definitions reflect the preventive and risk-focused nature of compliance, which remains central as institutions seek to enhance monitoring capabilities through technological means. In the context of this study, they provide a clear reference point for identifying the objectives and boundaries of automated compliance solutions.

3.2. Evolving Compliance Monitoring

Compliance monitoring in financial institutions has historically relied on manual processes, rule-based systems, and standardised reporting (Saaradeey, Ghosh, Ray, Ganesan, & Rajagopalan, 2019). Banks use compliance teams to conduct audits, transaction reviews, and customer due diligence, often cross-checking customer data against regulatory watchlists such as those from FATF, OFAC, and Interpol (FFIEC, 2023). Know your customer and enhanced due diligence procedures help verify identities and assess risk profiles (Sanctions.io, 2025).

Banks have improved their monitoring systems to bolster compliance efforts, yet significant challenges persist, particularly regarding data quality, as risk-related databases often contain inconsistent or incomplete information (Collibra, 2024). These databases may include individuals with criminal ties or political exposure, often listed without consent, creating transparency issues and data reliability problems for compliance officers (Lindemann Law, 2023). As a result, up to 95 percent of alerts in compliance monitoring processes may be false positives, wasting resources, while false negatives allow financial crimes to go undetected (Saaradeey et al., 2019).[1]

A survey by SIX Group (2024) shows a shift in priorities, with AI becoming key to addressing these challenges. Banks now focus more on efficiency and cost reduction, especially through AI-driven solutions for compliance automation, which help improve risk detection.

3.3. Automated Compliance and Regulatory Reporting

One key compliance task is the monitoring of individuals in positions of power, who present heightened risks of financial crime and corruption. Accurately tracking such relationships is resource-intensive and error-prone, often requiring manual checks and cross-referencing data.

LMs offer a transformative solution. By processing large volumes of unstructured data like news articles, financial disclosures, and legal documents, LMs have the potential to dynamically identify and assess individuals in leadership roles, improving accuracy and reducing reliance on potentially low-quality sources. This study's next step is to develop and evaluate a prototype that tests how well LMs can identify individuals in leadership positions. This capability is crucial for various compliance tasks, and the prototype aims to assess the feasibility, accuracy, and limitations of AI-assisted monitoring.

[1] A "false positive" occurs when a compliance system incorrectly flags a harmless individual or transaction as suspicious, leading to wasted resources. A "false negative" occurs when a genuine threat is overlooked, allowing potential financial crimes to go undetected.

4. Prototype Architecture and Operation Modes

The developed prototype for automated leadership identification operates in two core modes: Mode A (LM-only mode; Section 4.1) and Mode B (RAG mode; Section 4.2). The pipelines of these modes are illustrated in Figure 4.1 and described in more detail in the following sections. The implementation code for both modes is available on GitHub.[1]

4.1. Mode A: LM-Only

Mode A omits any external information retrieval. The model relies solely on its pre-trained internal knowledge. The corresponding pipeline consists of the following steps:

1. **Database**
   A name (e.g., "Jane Doe") is retrieved from a predefined dataset (see Section 5.3.1) for assessment.

2. **Prompt construction**
   A predefined instruction is combined with the name to form a minimal input prompt. The format of the prompt is as follows:

Prompt format

Question:
Is {name} a member of the management team (e.g. CEO, CFO, CTO, COO, CAO, CIO, EVP, head, director) or board of directors of a publicly listed company?

Answer with exactly one word: Yes or No.

3. **LM**
   The model processes the prompt and responds with a binary decision ("Yes" or "No") based solely on its internal knowledge, without access to current web sources or external context. To accommodate potential model-specific response structures (e.g., introductory markers such as "<thinking>"), responses are allowed up to a maximum length of 20 tokens, where one token typically corresponds to around three to four characters in English and German text.

4. **Output logging**
   The model's response, together with the input name, is recorded to ensure traceability. This logging supports subsequent evaluation, analysis, and auditing of inference results.

[1] See our GitHub repository here.

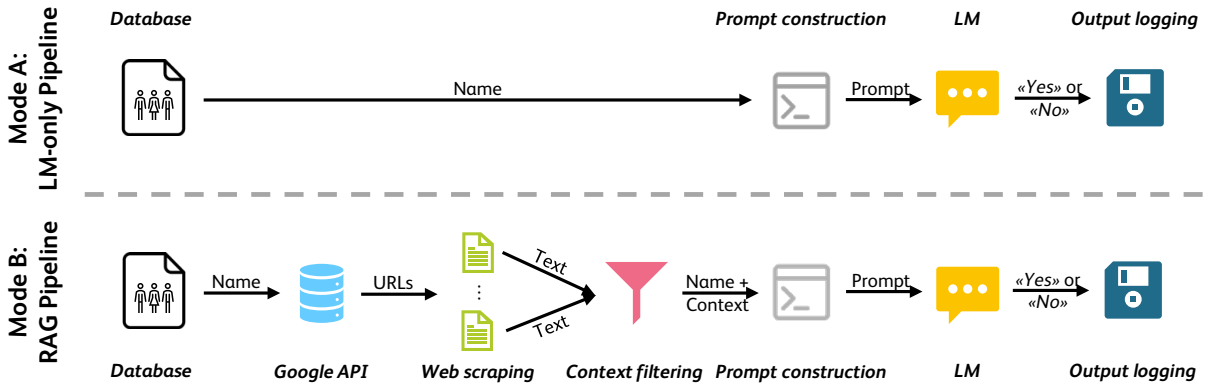

Figure 4.1: General architecture of the prototype

While computationally efficient, this mode is limited by its reliance on internal knowledge, which prevents it from capturing recent or highly specific information.

4.2. Mode B: Retrieval-Augmented Generation (RAG)

In Mode B, a complete RAG pipeline is employed, combining an LM with external information retrieval. This allows the model to incorporate current web-based evidence into its reasoning process. The corresponding pipeline consists of the following steps:

1. **Database**
   A name (e.g., "Jane Doe") is retrieved from a predefined dataset (see Section 5.3.1) for assessment.

2. **Google API**
   A query restricted to information from the past year is issued using the input name as an exact match. The top ten most relevant web links are then selected for further processing.

3. **Web scraping**
   The selected websites are parsed using `BeautifulSoup`, focusing strictly on HTML elements (e.g., `<p>`, `<h1>`, etc.). Non-HTML content is excluded.

4. **Context filtering**
   Each document corresponds to the content retrieved from a single URL fetched via Google. Documents are divided into paragraphs, which are considered relevant only if they contain at least part of the target name (i.e., first and/or last name) and at least one leadership-related keyword (in German or English).[2] Relevant paragraphs are scored based on keyword occurrences and ranked accordingly.

   For each document, the highest-ranked paragraphs are concatenated until a per-document character limit is reached. This limit is calculated based on the target model's context window size: the maximum number of tokens the model can process is first multiplied by three, reflecting the approximate number of characters per token in English and German text, and then divided by the number of documents processed (typically ten) to ensure that each document contributes equally and the aggregated context does not exceed the model's input size. Paragraphs that exceed this per-document limit are discarded.

   The final context is formed by aggregating the truncated contents from all documents into a single input string.

5. **Prompt construction**
   A predefined instruction is combined with the name and the retrieved context to form the input prompt. The format of the prompt is as follows:

   Prompt format

   Context:
   {context}

   Question:
   Is {name} a member of the management team (e.g. CEO, CFO, CTO, COO, CAO, CIO, EVP, head, director) or board of directors of a publicly listed company?

   Answer with exactly one word: Yes or No.

6. **LM**
   The model processes the prompt and responds with a binary decision ("Yes" or "No") based on its internal knowledge and the context provided. To accommodate potential model-specific response structures (e.g., introductory markers such as "<thinking>"), responses are allowed up to a maximum length of 20 tokens.

7. **Output logging**
   The model's response, together with the input name, is recorded to ensure traceability. This logging supports subsequent evaluation, analysis, and auditing of inference results.

Although computationally more demanding, Mode B improves factual accuracy by grounding responses in up-to-date and online evidence.

[2] Keywords include German terms such as the stem "Verwaltungsra" (to capture variations like "Verwaltungsrat" and "Verwaltungsrätin") as well as full forms like "Geschäftsleitung", "Präsident", and "Direktor", along with English terms such as "director", "board", "CEO", "executive", "and "founder".

5. Evaluation of Language Models

Selecting the appropriate LM is a critical design decision when developing AI-driven solutions for sensitive compliance tasks. In this study, the objective is to automate the identification of individuals in company leadership roles, a key requirement for effective PEP monitoring.

5.1. Evaluation Scope and Criteria

The evaluation focuses on comparing the performance of a diverse set of LMs for the given use case. To ensure a transparent and systematic process, models were chosen based on the following criteria:

- **Instruction following**: Ability to accurately interpret and respond to structured prompts.
- **Reasoning capability**: Proficiency in inferring leadership roles from limited, ambiguous, or incomplete data.
- **Context window size**: Capacity to process extended contextual information, particularly in RAG-mode.
- **Availability and ecosystem support**: Preference was given to models available on Hugging Face under open or permissive licences, which ensures easy integration, reproducibility, and suitability for deployment in controlled environments.

5.2. Models Selected for Evaluation

A diverse range of models was selected to reflect different design philosophies and capabilities, from more efficient SLMs to advanced LLMs capable of deeper contextual analysis. This selection enables a balanced and nuanced assessment of performance trade-offs in the context of compliance-related use cases. The selected models are listed in Table 5.1. This variety ensures representation across a spectrum of model sizes, architectures, and intended application scenarios.

In addition to the locally deployed models, GPT-4.1 accessed via OpenAI's API was also included in the evaluation. While this model does not fully meet the selection criteria, particularly with regard to deployment control and ecosystem openness, it represents one of the most advanced publicly accessible LLMs. Its inclusion enables a broader perspective on performance trade-offs between private and public model deployments in compliance-sensitive use cases.

5.3. Evaluation Methodology

The performance of each prototype was assessed using a benchmark framework specifically designed for this study. The evaluation methodology is illustrated in Figure 5.1 and consists of multiple components discussed in the following subsections.

Model	Parameters	Context window	Provider	Source
TinyLlama-v1.1	1.1B	2k tokens	TinyLlama Team	Link
Phi-4-mini-instruct	3.84B	128k tokens	Microsoft	Link
gemma-3-4b-it	4.3B	128k tokens	Google DeepMind	Link
zephyr-7b-beta	7.24B	32k tokens	Hugging Face H4 Team	Link
Mistral-7B-Instruct-v0.3	7.25B	32k tokens	Mistral AI	Link
Llama-3.1-8B-Instruct	8.03B	128k tokens	Meta AI	Link
DeepSeek-R1-Distill-Qwen-14Btree fonts	14.8B	128k tokens	DeepSeek AI	Link
Mixtral-8x22B Instruct v0.1	141B	64k tokens	Mistral AI	Link
GPT-4.1	Undisclosed	1m tokens	OpenAI	Link

**Note: Mixtral-8x22B is a mixture-of-experts model, activating only a subset of parameters per inference step.*

Table 5.1: Evaluated language models, number of parameters, context windows, providers, and corresponding sources

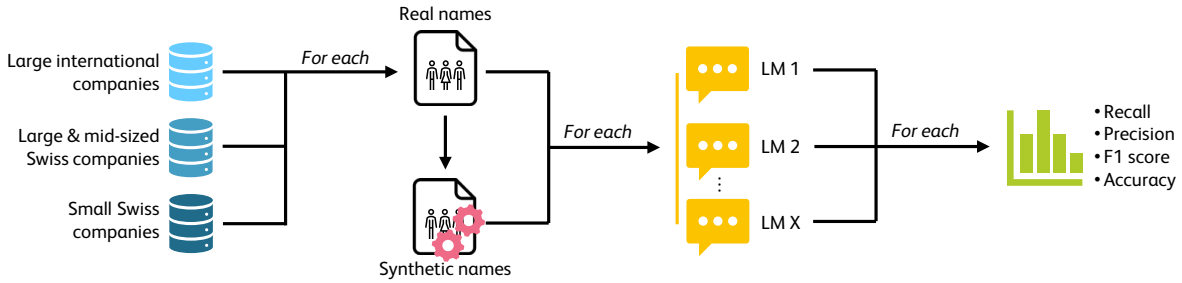

Figure 5.1: Evaluation process

5.3.1 Benchmark Datasets

To test the performance of different LMs, three real datasets and their corresponding synthetic datasets were used. All datasets consist of individuals in leadership roles at publicly listed companies:

- **International leadership dataset:** 100 individuals holding leadership positions at large international publicly listed companies.
- **Swiss large- and mid-cap leadership dataset:** 100 individuals in leadership roles at large and mid-sized publicly listed Swiss companies.
- **Swiss small-cap leadership dataset:** 100 individuals in leadership roles at small publicly listed Swiss companies.

For each of these real datasets, a corresponding synthetic name dataset was generated. These synthetic datasets contain 100 artificially constructed names each, created by randomly combining first and last names from the respective real datasets. They are used to test the prototypes' ability to avoid false positives.

5.3.2 Evaluation Metrics

The prototype pipeline described in Chapter 4 was applied to each of the LMs selected for this study. The following classification metrics were used to assess performance based on each LM's output:

- **True positives (TP):** Real leaders correctly identified.
- **False positives (FP):** Synthetic names incorrectly flagged as leaders.
- **True negatives (TN):** Synthetic names correctly identified as non-leaders.
- **False negatives (FN):** Real leaders not detected.

From these, the following derived metrics were calculated:

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP}$$

  Proportion of predicted leaders that are actually correct.

- **Recall:**

$$\text{Recall} = \frac{TP}{TP + FN}$$

  Proportion of real leaders that were successfully identified.

- **F1 score:**

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Harmonic mean of precision and recall, balancing both.

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Overall proportion of correct classifications.

5.4. Infrastructure

All private LMs were performed using cloud resources provided by Lambda[1]. The models ran on NVIDIA GH200 480GB GPUs, selected for their large memory capacity and computational throughput, which make them particularly suitable for large-scale inference tasks and processing long contexts in RAG-mode. Their high memory bandwidth allows for efficient handling of LMs without the need for aggressive quantisation or complex parallelisation strategies. In addition, running all models on identical hardware allowed for a standardised comparison of runtime performance. This enabled fair assessment of inference speed differences between models of varying sizes and architectures (i.e., modes).

Inference for GPT-4.1 followed the same methodological principles as for the locally deployed models. However, instead of running on Lambda infrastructure, GPT-4.1 was accessed via the OpenAI API. The main deviations concern the interaction modality and output control. Whereas privately deployed models operated through a direct text generation pipeline, GPT-4.1 required chat-based prompting with role-defined instruction, notably the use of a system role to enforce task behaviour, along with explicit stopping criteria to ensure consistent classification results.

[1] https://lambda.ai

6. Empirical Results of Prototype Evaluation

This chapter presents the empirical results of the prototype evaluation. We compare the performance of RAG and non-RAG modes (Section 6.1), evaluate performance metrics across all LMs in RAG mode (Section 6.2), assess false negative rates (Section 6.3), and analyze execution times (Section 6.4). All results are derived from a standardised pipeline without task-specific optimisation, which should be considered when interpreting the findings.

6.1. RAG vs. Non-RAG Mode

Empirical results demonstrate a clear advantage of RAG over Non-RAG approaches across nearly all tested models, as highlighted in Figure 6.1, which shows the F1 scores for the nine in-scope LMs on the dataset covering large internationally listed companies. Non-RAG models, relying solely on pre-trained knowledge, exhibited notable limitations in the given tasks where up-to-date and context-rich information is essential. In contrast, RAG integration substantially improved classification performance. With the exception of *TinyLlama-v1.1*, all models recorded significant increases in F1 scores when augmented with retrieved context. This underscores the importance of supplementing internal model knowledge, particularly in compliance-related tasks where factual accuracy is critical.

The findings are further supported by similar patterns observed in the other two datasets, which cover large and mid-sized listed Swiss companies as well as small listed Swiss companies.[1] These results confirm the general applicability of the observed performance advantages associated with RAG. Therefore, all subsequent evaluations are conducted in RAG-mode only.

6.2. Comparison of LMs in RAG-Mode

The empirical results, visualized in Figure 6.2, reveal clear patterns in model performance across the three datasets.

[1] See our GitHub repository here.

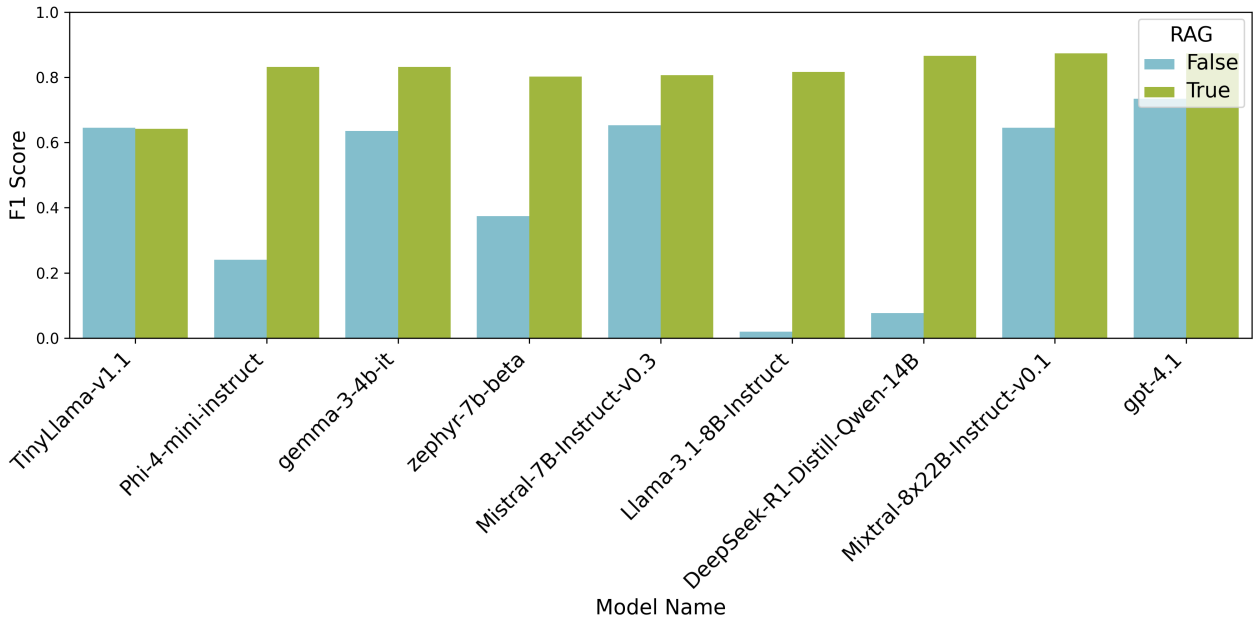

Figure 6.1: F1 Score comparison of various language models on the international dataset, with and without RAG enabled

The figure shows radar plots for each model, summarising key metrics, i.e., precision, recall, F1 score, and accuracy, across the international, Swiss large- and mid-cap ("Switzerland"), and Swiss small-cap ("Switzerland Nebenwerte") datasets.

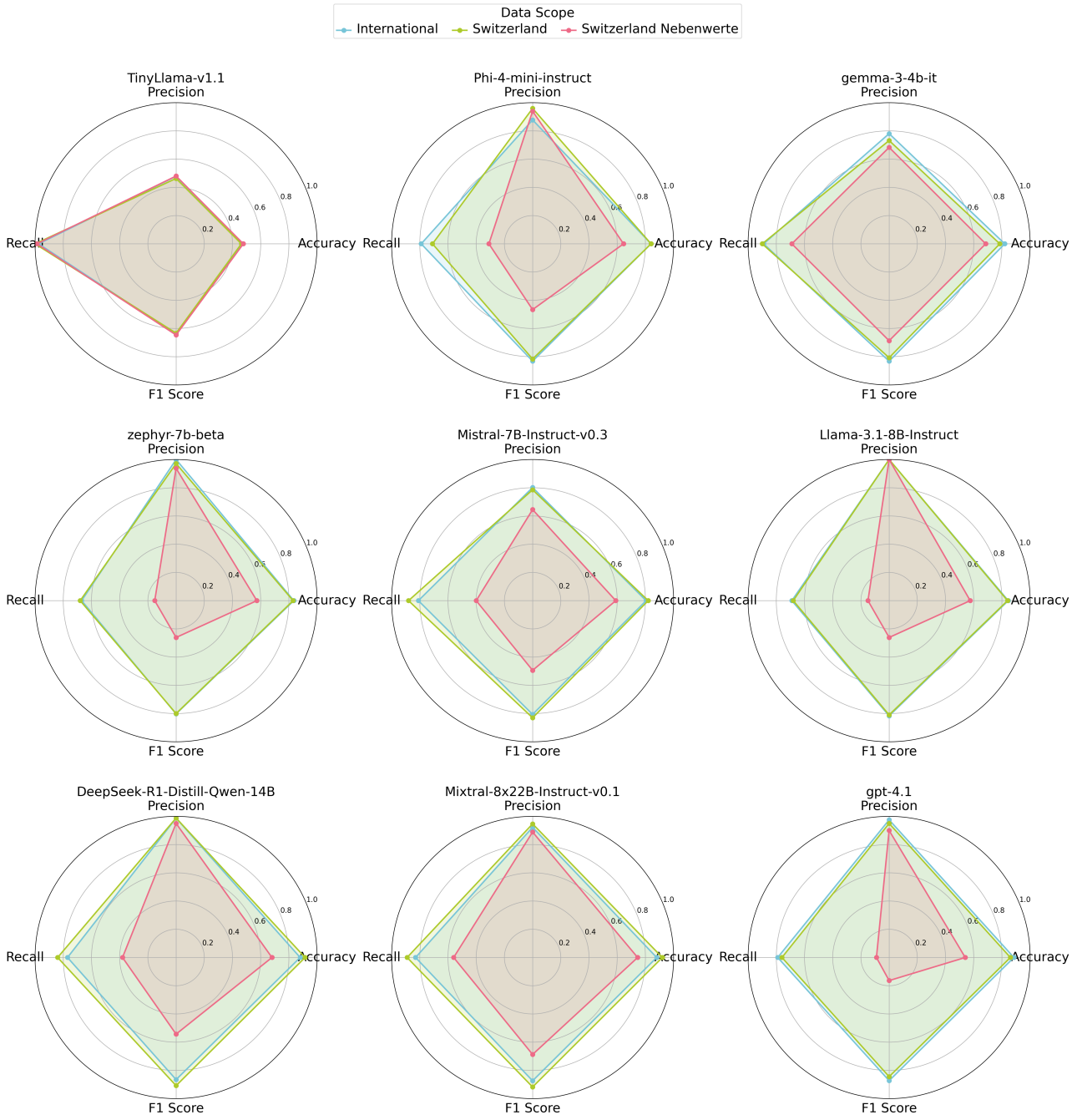

Figure 6.2: Comparison of different language models with different data scopes

The results highlight substantial performance differences between the datasets. Performance was, on average, highest on the international dataset, where the public visibility of company executives ensures rich and relevant web-based context. Swiss large- and mid-cap companies showed slightly lower performance, reflecting a reduced but still substantial online presence. By contrast, Swiss small-cap companies posed the greatest challenge. F1 scores dropped considerably due to limited online information and lower representation in pre-training corpora, making retrieval less effective.

Larger models such as *DeepSeek-R1-Distill-Qwen-14B*, *GPT-4.1*, and *Mixtral-8x22B-Instruct-v0.1* tended to deliver the most consistent results, often achieving F1 scores above 0.85 for the international and Swiss large and mid-cap datasets. In contrast, smaller models struggled to benefit from RAG, suggesting that limited capacity constrains their ability to process additional context effectively.

Moreover, the models revealed different trade-offs between precision, recall, and accuracy. *TinyLlama-v1.1*, for instance, achieved very high recall but low precision and accuracy, indicating a tendency to over-predict positive cases, likely due to limited capacity for nuanced interpretation of retrieved context. In contrast, models such as *Llama-3.1-8B-Instruct* and *zephyr-7b-beta* prioritised precision and accuracy at the expense of recall, reflecting a more conservative approach in classifying positive cases, particularly in the small-cap domain. Overall, while RAG substantially enhances performance in data-rich environments, its benefits are limited when retrieved context is sparse. Moreover, model size and architecture remain key factors in determining how effectively external information can be leveraged for accurate classification.

6.3. False Negatives Detection

In compliance applications, false negatives are critical, as failing to identify relevant individuals may entail regulatory and reputational risk. Figure 6.3 visualises the number of false negatives observed in RAG-mode on the international dataset. The results show notable variation across models. Larger models such as *GPT-4.1*, *DeepSeek-R1-Distill-Qwen-14B*, and *Mixtral-8x22B-Instruct-v0.1* produced moderate false negative counts (17–23) of a total

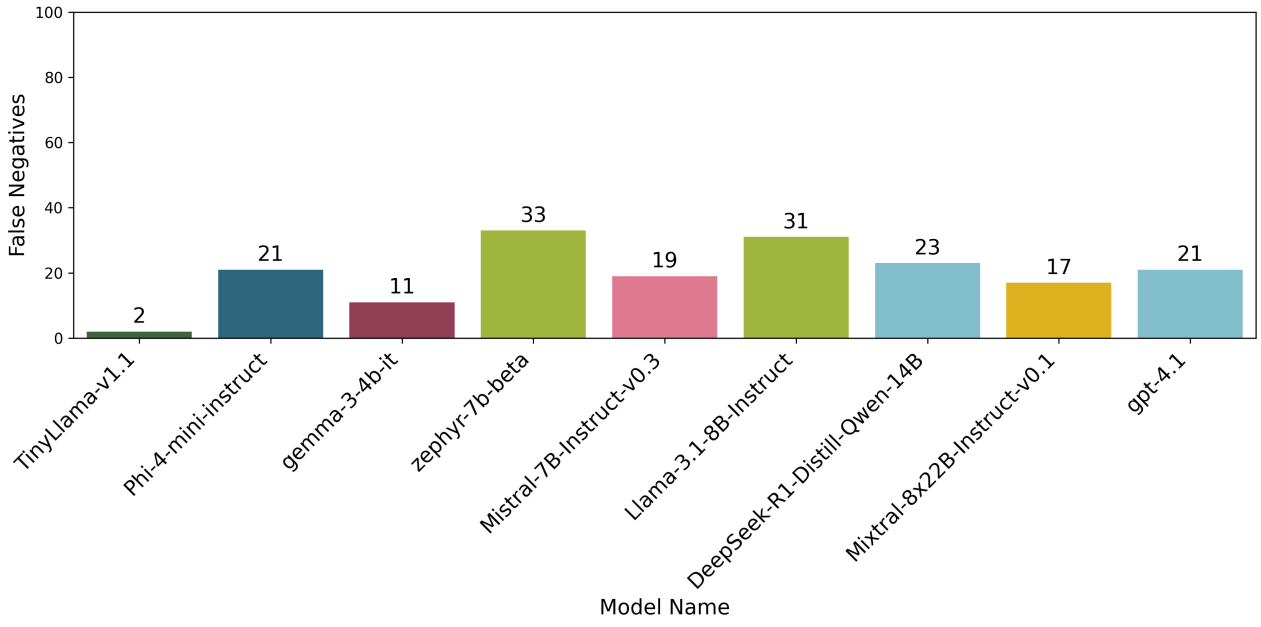

Figure 6.3: Comparison of false negatives detection by model, with RAG enabled and international scope

100, while *Llama-3.1-8B-Instruct* and *zephyr-7b-beta* exhibited the highest counts (31 and 33, respectively). By contrast, *TinyLlama-v1.1* reported only two false negatives. This, however, reflects a tendency to over-predict positives, leading to high recall but poor precision. While this reduces false negatives, it significantly increases false positives and review effort.

Overall, the results highlight the importance of balancing recall and precision. Conservative models risk missing true positives, while aggressive models may produce excessive noise. For compliance tasks, optimising this trade-off is essential to ensure both reliable detection and operational efficiency.

6.4. Comparison of Execution Times

Figure 6.4 shows a scatterplot of F1 scores and execution times across all models and datasets in RAG-mode. The results reveal a general trade-off between predictive performance and computational cost. While higher F1 scores are often associated with longer execution times, this relationship is not strictly linear. Some models combine strong performance with relatively efficient execution, while others require more time to achieve comparable results.

This variation can be explained by differences in model architecture, optimisation, and how effectively each model leverages retrieved context. Models that efficiently integrate relevant information tend to achieve higher F1 scores without incurring excessive computational overhead. Others, particularly very large models or those less optimised for inference speed, may require more processing time to reason over extended context windows or compensate for weaker retrieval processing capabilities. Notably, *Mixtral-8x22B-Instruct-v0.1* consistently exhibited the longest execution time, reflecting its status as the largest model tested within a private deployment.

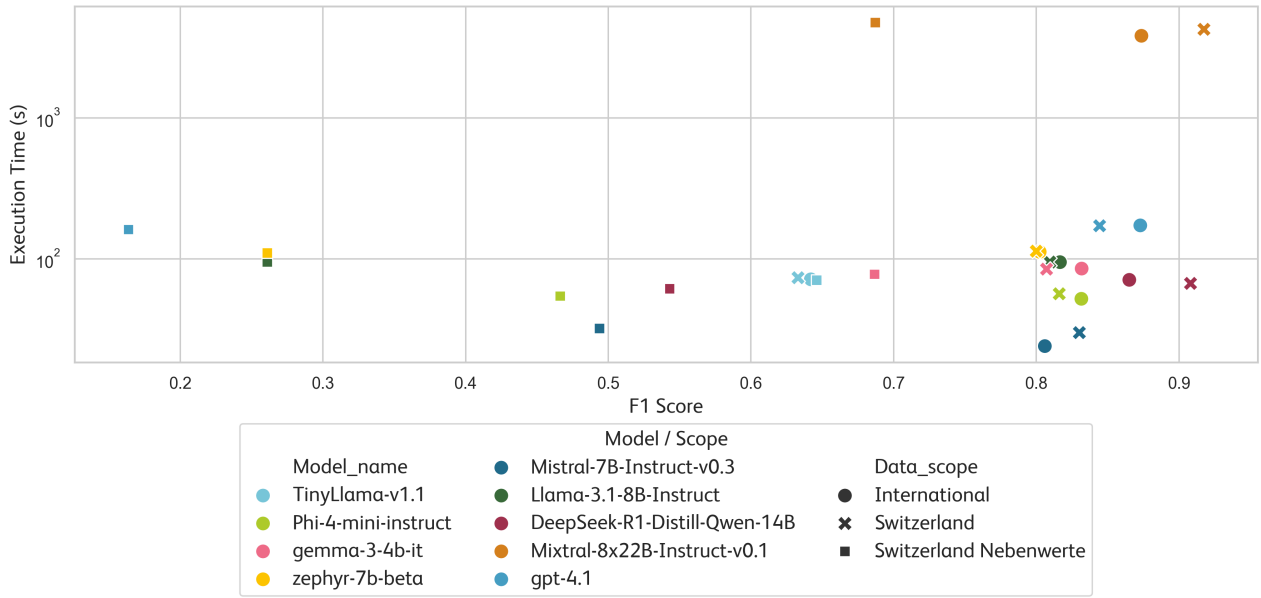

Figure 6.4: Overview of the execution time of different models and data scopes, related to the model's F1 score with RAG enabled

7. Deployment and Operational Considerations

Deploying LMs for compliance use cases, such as the prototype presented in this study, requires careful consideration of several interdependent factors. These include regulatory alignment, infrastructure requirements, cost efficiency, operational scalability, and architectural integration. Figure 7.1 provides an overview of these core dimensions and serves as a conceptual framework for the subsequent sections. Given the sensitivity of compliance-related data and processes, such considerations are especially critical to ensure secure and effective model deployment in regulated environments.

7.1. Regulatory Constraints

Compliance-related applications are subject to strict privacy and governance requirements. While public models offer convenience, they are often unsuitable due to limited control over data residency and processing. Instead, private LMs or deployments within regulated, on-premise, or sovereign cloud environments are essential. This approach ensures full data privacy, supports explainability, and aligns with the data protection mandates of regulatory bodies such as FINMA or under frameworks like the European AI Act, which classifies many financial compliance applications as high-risk.

7.2. Infrastructure Sizing

The infrastructure required for deploying LMs in compliance contexts is highly dependent on both technical and operational parameters. Key factors include:

- **Model type and size**: Larger models (e.g., *Mixtral-8x22B Instruct v0.1*) typically require significantly more memory and compute capacity, especially when running in parallel or under time constraints. In contrast, smaller models (e.g., *TinyLlama-v1.1* and *Phi-4-mini-instruct*) allow for a minimised infrastructure and are more suitable for resource-constrained environments.

- **Operation mode (Mode A vs. Mode B)**: The selected mode has a direct impact on system demands. Mode A (LM-only) relies purely on the model's internal knowledge and is computationally less demanding. Mode B (Retrieval-Augmented Generation), by contrast, requires additional resources for real-time search queries, web scraping, and dynamic context processing. This mode increases demands not only in terms of compute and memory, but also network bandwidth and I/O throughput.

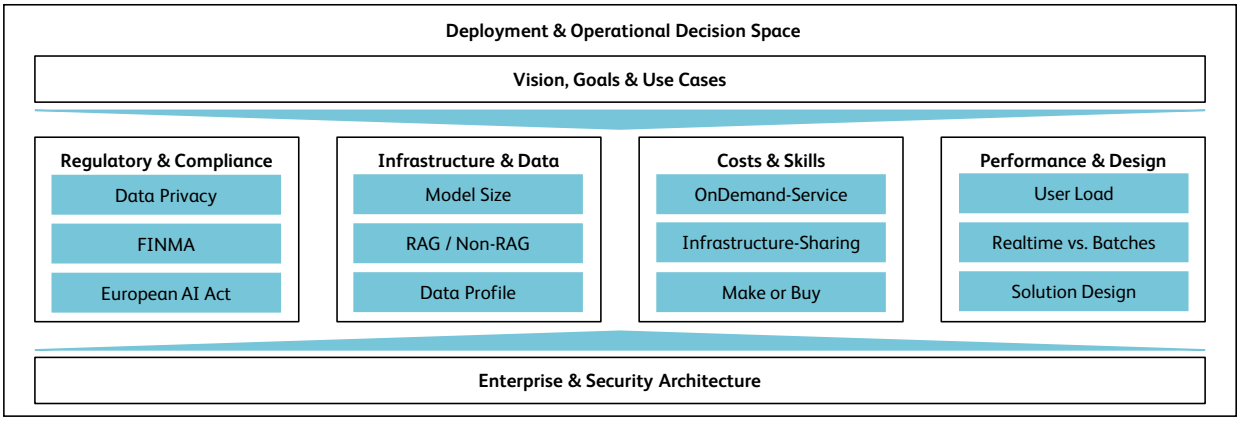

Figure 7.1: Overview of deployment and operational factors for integrating LMs in regulated environments

- **Data characteristics**: The dataset to be processed also significantly influences infrastructure requirements. As shown in the empirical evaluation, datasets from large international companies are more information-rich in RAG-mode. Conversely, datasets with smaller or less-known entities may result in shorter contexts but may require broader and deeper searches to ensure effective and adequate recall.

- **Parallelisation and throughput goals**: If the use case involves high-volume data processing, such as screening thousands of names daily, then the ability to scale horizontally (e.g., multiple concurrent inference jobs) becomes essential. This further impacts infrastructure sizing and may require specialised orchestration frameworks or dedicated GPU clusters.

Ultimately, infrastructure design must balance model capabilities, data characteristics, and performance expectations. Organisations should not only benchmark models under realistic workloads but also consider provisioning strategies that support flexible scaling as requirements evolve.

7.3. Cost Efficiency

Given that compliance screening tasks are not performed continuously, it is essential to avoid over-provisioning resources. A pay-as-you-use infrastructure model, where compute resources are allocated on demand, can provide financial flexibility and operational efficiency. Where such elasticity is not available (e.g., in on-premise deployments), institutions should consider multi-use infrastructure strategies: sharing GPU and memory resources across several AI use cases to maximise utilisation and return on investment.

7.4. Timing and Process Mode

The architecture and deployment of the system should be closely tied to how the solution is used in practice. Two major usage patterns are relevant:

- **Real-time queries**: Real-time queries (e.g., during onboarding or transactions) require low-latency responses and high availability. This places stricter demands on infrastructure sizing, monitoring, and rigorous testing for performance under load.

- **Scheduled background processes**: Scheduled background processes (e.g., daily or weekly batch PEP screenings) are less time-sensitive. These can be executed using a smaller infrastructure footprint, allowing for more efficient resource allocation, even if total processing time is longer.

Choosing between real-time versus scheduled execution affects not only the infrastructure but also the design of logging, alerting, and scaling mechanisms. Time-critical systems require high fault tolerance and redundancy, whereas scheduled processes allow more relaxed operational parameters.

7.5. Enterprise Architecture

LM applications represent only a small fraction, often one out of more than 150 to 200 enterprise-wide applications, within an organisation's IT landscape. Consequently, LMs must be embedded into the broader IT architecture rather than treated as standalone solutions. Operating concepts, support processes, and compliance incur fixed costs at platform level. To avoid unnecessary complexity, it is crucial to leverage and scale existing platform concepts, aligning LM deployments with the enterprise's platform and operating strategy.

8. Conclusion and Outlook

This study investigated the potential of language models (LMs) for automated identification of individuals in executive and board-level roles at listed companies. Based on empirical results across multiple configurations, the following theses summarise key findings and outline implications for design, deployment, and operational strategy.

The prototype shows promise for the use case, though model performance is subject to limitations. The results of the presented prototype suggest that LMs can support the automated identification of individuals in executive and board-level roles. However, performance varies significantly depending on the model architecture and underlying data. In particular, individuals associated with larger and more prominent companies, about whom more extensive public information is available, are identified with greater reliability. By contrast, the detection of individuals from smaller organisations remains challenging. Moreover, the reported results are based on a standardised evaluation pipeline without optimisation and would likely differ if models were fine-tuned or adapted to the task.

Retrieval-Augmented Generation (RAG) improves performance, but increases complexity and introduces additional dependencies. The integration of RAG has been shown to enhance classification accuracy by providing LMs with relevant external context. However, this improvement comes at the cost of increased system complexity. In particular, context retrieval and preparation introduce additional operational complexity and require decisions regarding the selection, filtering, and structuring of contextual information, all of which can substantially influence outcomes. Furthermore, in the present prototype, the use of RAG creates an additional dependency on a third-party provider for web search capabilities. This introduces considerations regarding privacy, control, and long-term operational independence. While RAG approaches are well suited to dynamic and knowledge-intensive tasks, their performance and effectiveness depend on careful design and tuning, and critically, on the availability and quality of contextual data.

False negatives remain a critical limitation, particularly given compliance requirements. Despite encouraging overall results, the occurrence of false negatives, especially the failure to identify true decision-makers, represents a significant concern for practical application. In the compliance context, such omissions are particularly problematic, as they may result in the non-detection of relevant individuals, thereby exposing institutions to regulatory, legal, and reputational risks. Minimising false negatives is essential to ensure sufficient recall and meet the high standards of compliance. A key lever is improving data quality, since models perform well only if training and context data are accurate, complete, and relevant.

Performance and operational costs are inherently linked and require careful trade-offs. Larger models tend to achieve higher classification performance but also incur longer inference times and increased infrastructure requirements. However, this relationship is not universal, and model efficiency varies significantly across scenarios (e.g., datasets, tasks). The same model may perform differently depending on input complexity and context availability. For fast or high-volume processing, smaller and more efficient models may offer advantages. Model selection should therefore balance accuracy against operational factors such as speed, scalability, and cost.

Model selection should be aligned with use case-specific requirements and constraints. Beyond performance and cost, model selection must reflect domain-specific needs. Compliance applications, for example, emphasise privacy, explainability, and controllability. Larger public models offer superior reasoning and frequent updates but may fall short of strict data governance standards. In contrast, private models provide more control and confidentiality but demand greater maintenance and tuning. There is no universal solution, as each use case requires a holistic evaluation of accuracy, privacy, regulatory alignment, maintainability, and process integration.

Operationalisation requires alignment with infrastructure, regulatory, and process requirements, and must remain adaptable. Deploying LMs for compliance use cases demands balancing regulatory constraints, infrastructure needs, cost efficiency, and process integration. Regulatory requirements often necessitate private or sovereign deployments to ensure data privacy and explainability. At the same time, infrastructure and process designs must flexibly accommodate different usage patterns, from real-time queries to scheduled background processing. Finally, given the rapid evolution of LMs, continuous monitoring and adaptation are essential to maintain performance, regulatory alignment, and operational effectiveness over time.

Authors

This condensed study was prepared in collaboration with the following individuals who contributed in the form of text, discussion, document reviews, and other forms of feedback (in alphabetical order):

Authors HSLU

Thomas Ankenbrand
Head Competence Center Investments

Denis Bieri
Lecturer

Joël Ettlin
Research Associate

Angelo Gattlen
Research Associate

Jovana Milojevic
Lecturer

Guest Author (Chapter 7)

Carla Caspar
Strategic Innovation Manager
Inventx Lab AG

We would also like to thank the research partners of this study, namely e.foresight, Finnova, InvetxLab, Kanton Zug, SFTI / Swiss Fintech Innovations, SIX, Swiss Bankers Prepaid Services, and Zürcher Kantonalbank, for their monetary and content-related support.

Contact

For more information about this study, please contact us at:

Thomas Ankenbrand
Lucerne University of Applied Sciences and Arts
thomas.ankenbrand@hslu.ch

References

Ankenbrand, T., Bieri, D., & Gattlen, A. (2025). *IFZ FinTech Study 2025. An Overview of Swiss FinTech.* Retrieved 06/05/2025, from https://www.hslu.ch/-/media/campus/common/files/dokumente/h/1-medienmitteilungen-und-news/2025/w/ifz-fintech-studie-2025.pdf

Bank of England. (2024). *Artificial intelligence in UK financial services 2024.* Retrieved 19/03/2025, from https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024

Basel Committee on Banking Supervision. (2005). *Compliance and the compliance function in banks.* Retrieved 25/03/2025, from https://www.bis.org/publ/bcbs113.htm

Castellum AI. (2024). *Best Practices for PEP Screening.* Retrieved 18/03/2025, from https://www.castellum.ai/insights/best-practices-for-pep-screening

CBinsights. (2025). *Enterprise AI agents & copilots: Our growth projections for the $5B+ market.* Retrieved 05/05/2025, from https://www.cbinsights.com/research/enterprise-ai-agents-market-size/?utm_campaign=newsletter_general_HN_hs&utm_medium=email&_hsenc=p2ANqtz-8-IVfRV04wNdPO-xh3ln1413vVcMzrtSt3StQk1islLBypg8Ja2N7nhFMNoPWGDbbXRTZYVWdKAeO29pMRkQEb4viGuQ&_hsmi=359532406&utm_content=359532406&utm_source=hs_email

Clairo AI. (2023). *Private LLMs vs Public LLMs.* Retrieved 21/03/2025, from https://www.clairo.ai/blog/privatellms-vs-publicllms

Collibra. (2024). *The fading flame: Why data governance under BCBS 239 needs your attention now.* Retrieved 07/05/2025, from https://www.collibra.com/blog/the-fading-flame-why-data-governance-under-bcbs-239-needs-your-attention-now

CompFidus. (2024). *How to Conduct Enhanced Due Diligence on Politically Exposed Persons (PEPs).* Retrieved 25/03/2025, from https://compfidus.com/risk-management/how-to-conduct-enhanced-due-diligence-on-politically-exposed-persons/

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending various regulations and directives (Artificial Intelligence Act).* Retrieved 20/03/2025, from https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

FFIEC. (2023). *Assessing Compliance with BSA Regulatory Requirements.* Retrieved 20/03/2025, from https://bsaaml.ffiec.gov/manual/AssessingComplianceWithBSARegulatoryRequirements/04?utm.com

FINMA. (2017). *FINMA Circular 2017/1: Corporate governance – banks.* Retrieved 20/03/2025, from https://www.finma.ch/en/~/media/finma/dokumente/dokumentencenter/myfinma/rundschreiben/finma-rs-2017-01-20200101.pdf?sc_lang=en&hash=40C9AA3758DA15953D000B3B0497146D

FINMA. (2024). *FINMA Risk Monitor 2024: Principal risks for the financial sector and uncertainties due to geopolitical tensions.* Retrieved 07/05/2025, from https://www.finma.ch/en/news/2024/11/20241118-mm-finma-risikomonitor-24/

FINMA. (2025). *FINMA survey: artificial intelligence gaining traction at Swiss financial institutions.* Retrieved 07/05/2025, from https://www.finma.ch/en/news/2025/04/20250424-mm-umfrage-ki/

IBM. (2024). *Small Language Models: Efficient AI for Targeted Applications.* Retrieved 25/03/2025, from https://www.ibm.com/think/topics/small-language-models

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., … others (2020). *Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems*, *33*, 9459–9474.

LexisNexis Risk Solutions. (2023). *LexisNexis Risk Solutions Study Reveals Global Financial Crime Compliance Costs for Financial Institutions Totals More Than U.S.$206 Billion.* Retrieved 20/03/2025, from https://risk.lexisnexis.com/about-us/press-room/press-release/20230926-global-financial-crime-compliance-costs

Lindemann Law. (2023). *How to clear unjust convictions and false information from World-Check, social media, and the internet.* Retrieved 20/03/2025, from https://lindemannlaw.ch/insights/how-to-clear-unjust-convictions-and-false-information-from-world-check-social-media-and-the-internet/

Priyanshu, A., Vijay, S., Kumar, A., Naidu, R., & Mireshghallah, F. (2023). *Are chatbots ready for privacy-sensitive applications? An investigation into input regurgitation and prompt-induced sanitization*. *arXiv preprint arXiv:2305.15008*.

RedHat. (2024). *Large language models (LLMs) vs Small language models (SLMs).* Retrieved 21/03/2025, from https://www.redhat.com/en/topics/ai/llm-vs-slm

Risk Management Association. (2024). *Explainability Challenges Are a Growing Concern for Bank Governance of AI.* Retrieved 25/03/2025, from https://www.rmahq.org/journal-articles/2024/june-july-2024/explainability-challenges-are-a-growing-concern-for-bank-governance-of-ai/

Saaradeey, S., Ghosh, D., Ray, R., Ganesan, S., & Rajagopalan, R. (2019). *Anti Money Laundering Disrupting Status Quo In AML Compliance.* Retrieved 20/03/2025, from https://www.oracle.com/a/ocom/docs/industries/financial-services/fs-disrupting-status-quo-aml-complaince-wp.pdf

Sanctions.io. (2025). *Switzerland's Anti-Money Laundering Regulations: A 2025 Guide.* Retrieved 20/03/2025, from https://www.sanctions.io/blog/switzerlands-anti-money-laundering-regulations-a-2025-guide#:~:text=Switzerland

SIX Group. (2024). *SIX Future of Finance 2024/25.* Retrieved 24/03/2025, from https://www.six-group.com/dam/download/company/report/Studies/six-fof-2024-25-web.pdf

Swiss FinTech Innovations. (2025). *A Scalable Framework for Implementing Artificial Intelligence in Swiss Financial Institutions.* White Paper. Retrieved 05/04/2025, from https://www.sfti.ch

The Alan Turing Institute. (2024). *The Impact of Large Language Models in Finance: Towards Trustworthy Adoption.* Retrieved 18/03/2025, from https://www.turing.ac.uk/news/publications/impact-large-language-models-finance-towards-trustworthy-adoption

Wang, B., Zhang, Y. J., Cao, Y., Li, B., McMahan, H. B., Oh, S., … Zaheer, M. (2023). *Can public large language models help private cross-device federated learning? arXiv preprint arXiv:2305.12132*.

Lucerne School of Business
Institute of Financial Services Zug IFZ
Campus Zug-Rotkreuz
Suurstoffi 1
6343 Rotkreuz

T +41 41 757 67 67
ifz@hslu.ch
hslu.ch/ifz

A study conducted by

HSLU Lucerne University of Applied Sciences and Arts

AACSB ACCREDITED

ISBN-Number
978-3-907379-52-3