SUMATOSOFT

Industries: Fintech, digital banking, and consumer lending

Region: Lithuania

PROJECT TYPE

Custom LLM development, production build (retrieval-augmented generation, fine-tuning, and agentic workflows)

TECHNOLOGIES

Python, FastAPI, Llama 3, LoRA/PEFT, Sentence, Transformers, LangChain, LlamaIndex, PostgreSQL, pgvector, Qdrant, Azure OpenAI, AWS Bedrock, NeMo Guardrails, Langfuse, Docker, Kubernetes

DURATION

8 months

TEAM

1 Solution Architect
2 LLM and ML Engineers
1 Data Engineer
2 Backend Engineers
1 QA Engineer
1 Project Manager, working with a compliance subject-matter expert on the client side

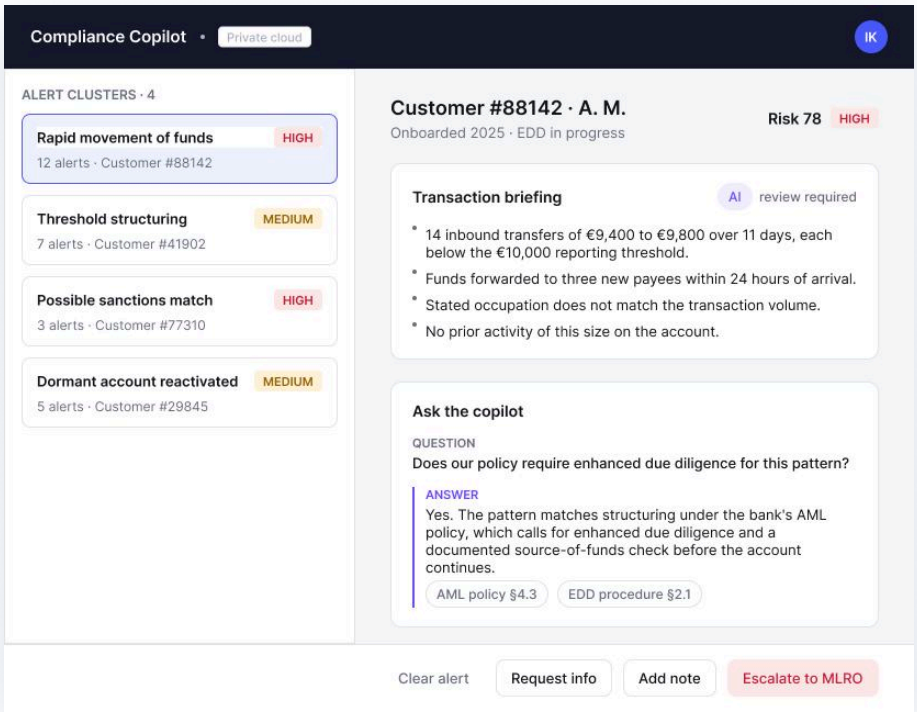
A private LLM compliance copilot for a European digital bank

A licensed European digital bank offers accounts, cards, consumer loans, and payments across the EU. As its customer base grew, the compliance team spent most of its day reading policy by hand and writing cases from long transaction histories. SumatoSoft built a private large language model copilot that runs inside the bank's own cloud, grounds every answer in the bank's policies, and cites the clause behind it. Analysts now clear alerts and onboard customers faster, and every decision carries an audit trail.

Project background

A licensed European digital bank had grown quickly, and its compliance operation had not scaled with it. Transaction monitoring generated a high volume of alerts, the large majority of them false positives, and each one took an analyst's time to clear. Onboarding customers who needed enhanced due diligence meant reading across several long policy documents that changed often, which slowed legitimate applications and frustrated the people the bank most wanted to keep.

The work was also hard to audit. Analysts wrote case narratives from memory and from scratch, and a regulator's request for evidence could take weeks to satisfy. The 2025 round of supervisory attention to AI and model risk added another requirement: any tool the bank adopted would be judged on whether it could explain its reasoning and leave a trail.



sumatosoft.com
info@sumatosoft.com

SUMATOSOFT

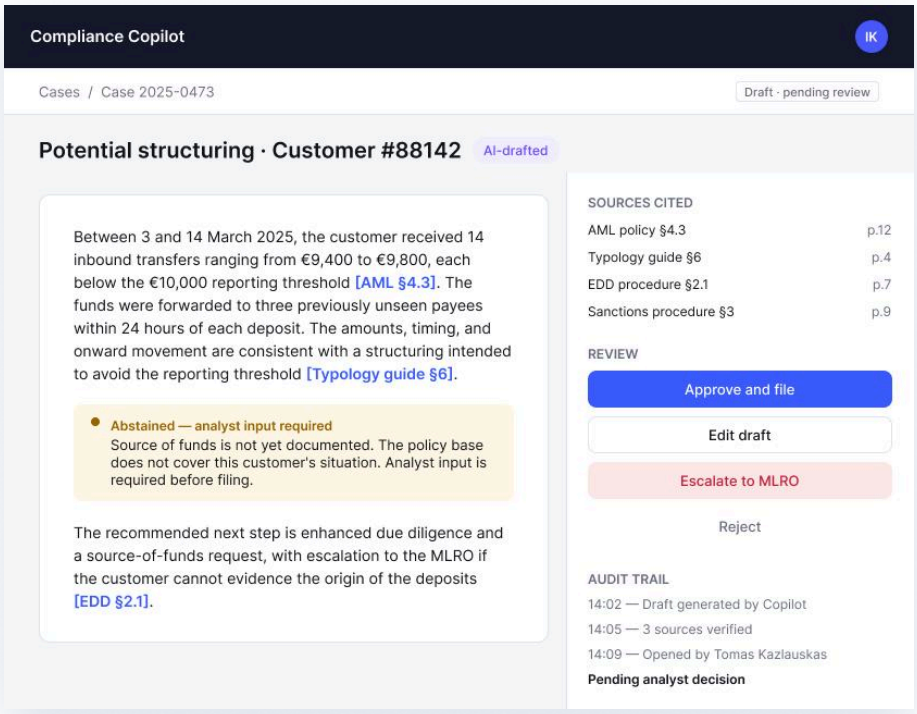
Project Distinctive Features

- Runs inside the bank's own cloud, with no customer data leaving the perimeter
- Cites the exact policy clause behind every answer, and abstains when the policy base does not cover the question
- Fine-tuned on the bank's own casework, so drafts match its house style and escalation rules
- Keeps a human analyst in control, with sign-off on every generated case
- Re-indexes policies the moment they change, so answers track the current rulebook

Business challenge

The bank wanted SumatoSoft to cut the manual load on its compliance analysts without weakening the control a regulated institution depends on. The goal was faster customer onboarding and faster alert handling, achieved by a system that an analyst stayed in charge of rather than one that decided on its own.

Four conditions framed the work. Every answer had to be grounded in the bank's own policies and cite the clause behind it, so an analyst and a regulator could both trace it. The system had to run inside the bank's own cloud, with no customer data leaving the perimeter. It had to fit the case management and monitoring tools the team already used. And it had to leave a complete audit trail behind every decision.



Our solution

SumatoSoft built the copilot as a private service within the bank's Azure cloud account, with AWS Bedrock as an alternative hosting option. No customer or transaction data leaves the perimeter, and no third party trains on it.

sumatosoft.com
info@sumatosoft.com

SUMATOSOFT

At the core sits a retrieval-augmented generation pipeline. The team indexed the bank's AML, KYC, sanctions, and fraud policies along with a history of resolved cases, splitting each document so retrieval returns the relevant passage, not a whole manual. A hybrid search over PostgreSQL with pgvector and Qdrant combines keyword and semantic matching. Every answer names the clause it drew on, and a reviewer can open the source in one click. When retrieval finds nothing that supports an answer, the copilot abstains and says so, which keeps it from inventing a policy that does not exist.

For drafting, SumatoSoft fine-tuned an open-source Llama 3 model with LoRA on the bank's own casework. The model writes alert briefings and case narratives in the bank's house style and follows its escalation rules, so analysts edit rather than rewrite. LangChain and LlamaIndex orchestrate the retrieval and the drafting, and a set of agentic steps handles the repetitive work: summarizing a transaction history into a briefing, grouping alerts that share a pattern, drafting the case narrative, and flagging gaps for the analyst. A human analyst signs off on everything before it leaves the system.

Guardrails screen for prompt injection, and role-based access ties each analyst to the data and actions their function allows. A full log records every prompt the system receives and every decision an analyst makes. Before launch, SumatoSoft red-teamed the system and built an evaluation harness that scores output quality on every release. The team modeled token costs up front and added cost and latency monitoring, so running costs stay predictable as volume grows.

Customer’s benefits

Within the first year of rollout, the compliance team handled a larger book of business without adding headcount in proportion. Complex onboarding reviews that used to take about five business days now take about two, roughly 55% faster, because analysts read a grounded summary instead of searching policy by hand. Alert triage moved from around 30 cleared alerts per analyst per day to between 70 and 80, since the copilot clusters related alerts and briefs the analyst on each group. Writing a case narrative dropped from about 90 minutes to about 25, with the governing clause already cited. Audit requests that once took three weeks are answered in about four days, drawn straight from the decision log. New analysts reach full productivity in about 6 weeks rather than 3 months because the copilot encodes the policy they would otherwise memorize.

What’s happening with the project right now?

The copilot runs in production across the bank's alert and onboarding workflows. SumatoSoft maintains the system and is extending it to more alert types and to a customer-facing FAQ that answers common KYC questions before they reach a human. The bank and SumatoSoft are also scoping a quarterly model review to ensure the fine-tuned model keeps pace with new typologies and regulatory changes.

sumatosoft.com

info@sumatosoft.com

One Boston Place, Suite 2602 Boston, MA 02108, United States