November 2025

A Year of FinLLM

Building the UK's first large language model for financial services

Leading the safe, ethical and responsible adoption of AI in financial services

UK financial services entered 2025 with a clear goal. Firms wanted the benefits of generative AI, but they needed it to work inside a regulated environment where accuracy, privacy and auditability shape every decision. General models could help with broad tasks, yet they struggled with the precise, structured work that advisers, analysts and compliance teams handle every day.

One year ago, Aveni began working with Lloyds Banking Group and Nationwide to address a shared challenge: **creating AI that meets the standards of UK financial services**. Both institutions partnered with us to shape a model trained on trusted sources and tested against real workflows. Their teams provided invaluable input throughout development, helping to validate outputs, challenge assumptions and strengthen our approach to governance, risk and model behaviour.

The **first version of FinLLM** was delivered in May 2025. It captured the foundations: strong financial reasoning; reliable extraction; and the ability to work with long, complex documents. By September, we released an **improved model** with tighter safety controls and clearer performance on the structured data tasks that drive most operational processes.

This report reflects what we learned over the last twelve months. It explains how we built the UK's first financial services specific language model, why we took an iterative path, and what this means for firms that want AI they can trust in regulated settings. It is the story of a model shaped by real financial work, tested in production, and designed for the standards the industry requires.

The limitations of general models in financial services

Most large language models are trained on large collections of internet text. That breadth helps them recognise general patterns, but it creates clear limitations for financial services. Because the training data includes unverified, biased and sometimes toxic material, these models can carry those patterns into their outputs. In practice, this often leads to misinterpreted terminology, inaccurate statements and missed regulatory nuance. These gaps make it difficult for them to handle the structured, detail-heavy tasks that financial teams rely on.

> "Train on the whole internet and you get everything. Including the bits that don't belong in financial services. Internet data includes internet problems"
>
> - Mouna Samrout, Senior Product Manager

To support the sector properly, the model needed to understand UK financial language, treat sensitive information responsibly and produce outputs that firms could rely on. That became the foundation for everything that followed.

Building from trusted foundations

Reliable intelligence depends on reliable data. We assembled a curated, financial-specific corpus built from lawful, relevant and traceable sources that reflect the real knowledge used across the sector.

What went into it

The dataset focused on high-quality UK financial materials. This included FCA and PRA guidance, company filings, educational resources, and reputable reference content. Each source provided authoritative information that mirrors the structured documents professionals use in their daily work.

How we prepared it

Raw data is only the starting point. We refined every component, removing duplicates, filtering noise and improving linguistic quality to ensure the model learned from clear, consistent English.

Privacy was addressed with a risk-based approach. Public, non-sensitive material remained unchanged. Financial identifiers were redacted when appropriate, while names of public figures were retained for context. Content drawn from sensitive community spaces, such as forums, was fully pseudonymised.

The outcome is a well-structured and carefully governed dataset that gives the model strong factual grounding while supporting firms' data protection requirements.

How we built and improved the model

Developing a reliable financial model required an approach that proved value early and reduced unnecessary risk. Instead of starting with a large system and hoping it would perform well, we built in stages, learning from each iteration before moving to the next.

Start small, learn fast

We began with smaller prototypes. These early versions allowed us to test data combinations, training methods and evaluation criteria without committing to full-scale training runs. Each experiment revealed what contributed to accuracy, what introduced noise and where the model needed deeper financial grounding.

Scale up with purpose

As we moved to larger models, we targeted the capabilities that matter most in practice. Models needed to read long documents, work confidently with tables and numbers, and hold multi-turn conversations without losing accuracy.

Filling the gaps with synthetic content

Some financial skills are simply under-represented in public data. Tasks like reasoning over tables and performing calculations, for example, rarely appear in sufficient volume. We addressed this by generating additional training material from high-quality sources such as regulator guidance and textbooks. We also transformed web tables into clear analytical write-ups.

Right-sizing for deployment

We produced a high-performing flagship model, then derived a smaller, more efficient version. The smaller model retains most of the important capabilities whilst meeting real-world constraints around budget, latency, and infrastructure.

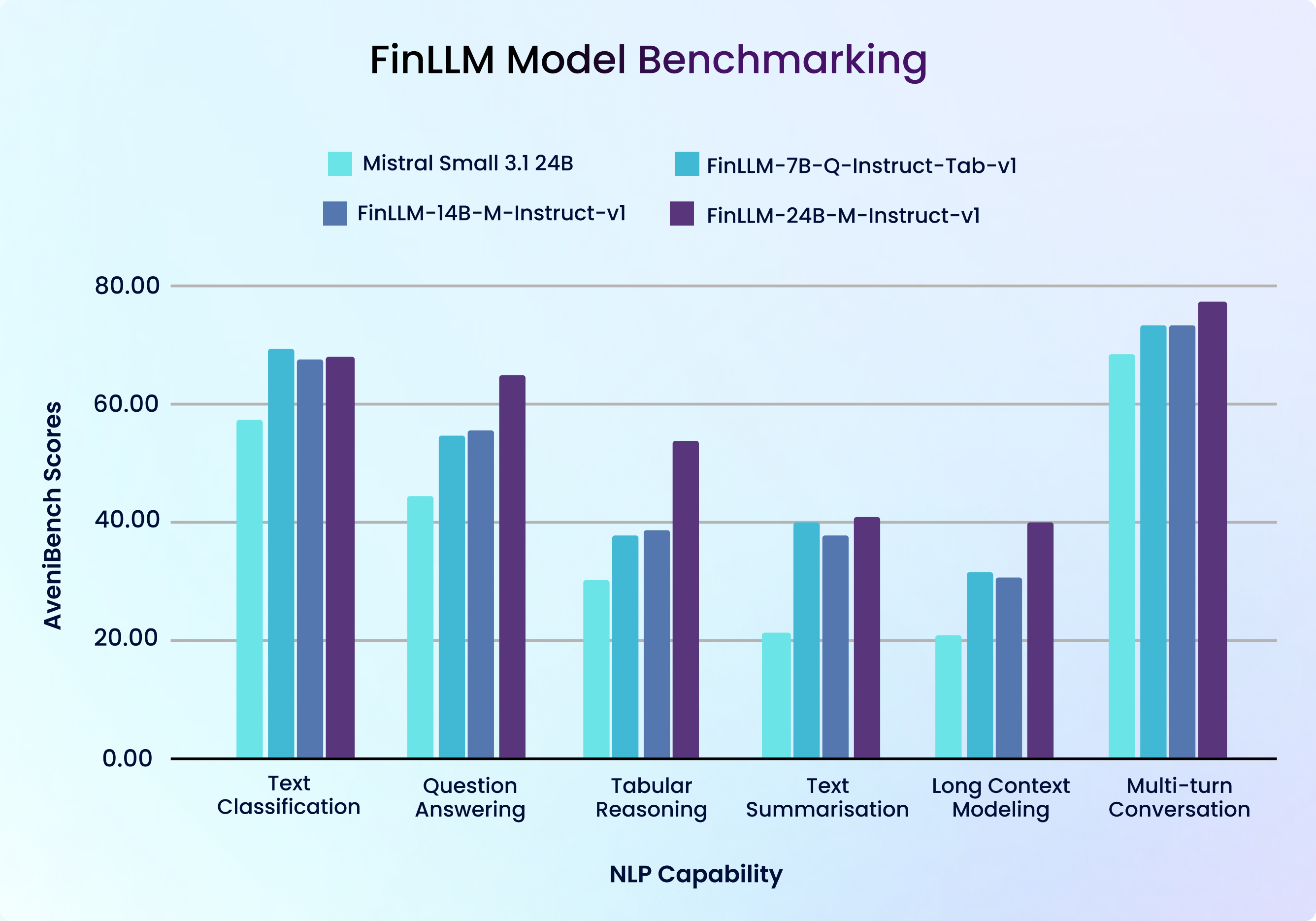
Accelerated training with NVIDIA

Training at scale required robust infrastructure. We used **NVIDIA NeMo** and NVIDIA accelerated computing to optimise training efficiency. This approach aligns with the growing need for domain-specific models that deliver predictable, safe behaviour in regulated sectors. By focusing the training data on relevant financial knowledge and removing unnecessary content, we achieved better performance and more responsible deployment.

Testing what matters: AveniBench

Evaluating a specialised model requires tests that reflect how financial services work in practice. Generic benchmarks offer useful signals, but they rarely measure the tasks that influence suitability reviews, case assessments or quality assurance. To close that gap, we created an evaluation suite designed around real financial workflows.

AveniBench brings these requirements together in one framework. It assesses financial tasks such as classification, fact finding and tabular reasoning, alongside essential skills like comprehension, mathematics and safe behaviour under pressure. We also released a public subset so organisations can compare performance using the same criteria. This provides a more practical picture of how the model behaves when accuracy, reliability and regulatory expectations shape every decision.



The FinLLM models demonstrate a clear advantage in all six NLP capabilities in comparison to Mistral Small 3.1 24B.

Proof in production

Lab tests tell you something. Production tells you everything.

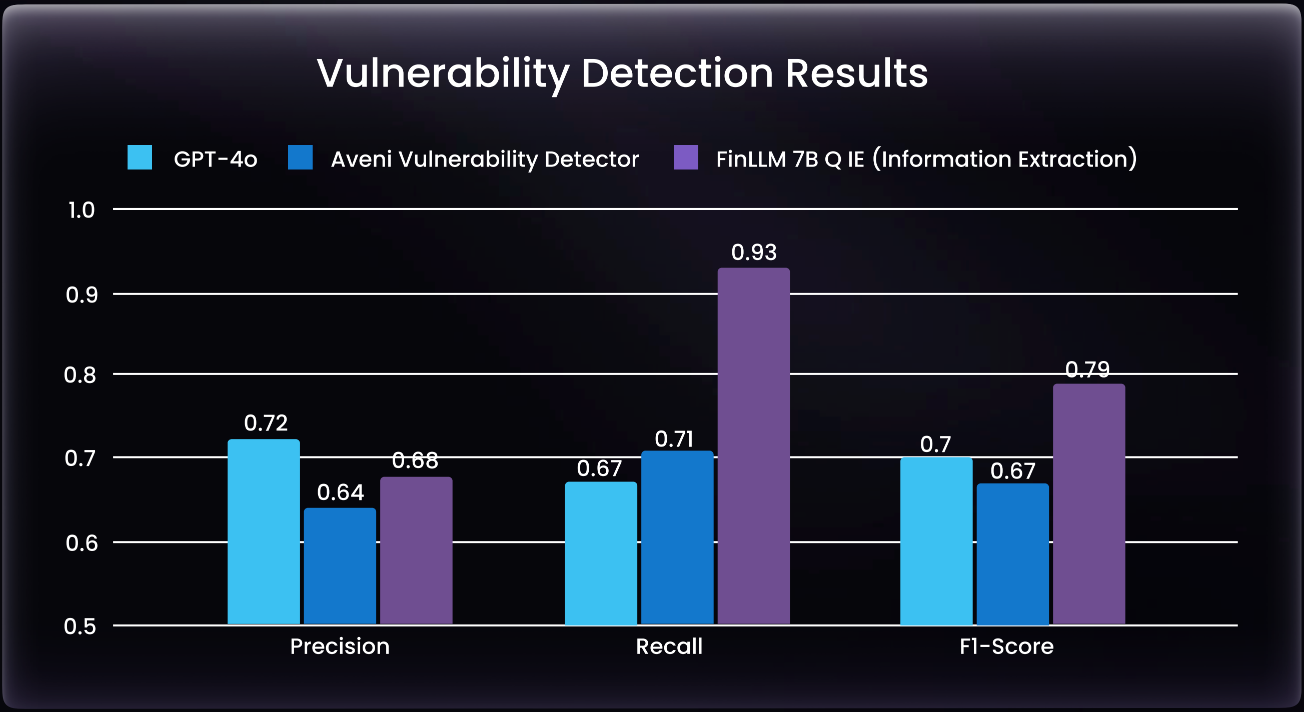
The only way to understand how a model behaves in regulated environments is to deploy it into real workflows and judge its performance across accuracy, consistency and operational efficiency. We introduced the model into two live Aveni products to see how it responded to genuine customer interactions and adviser tasks.

Aveni Detect: spotting customer vulnerability

Traditional pipelines break call transcripts into small chunks and run them through multiple separate models. FinLLM can analyse the full transcript in one pass, then hand off to a human for a final verification step.

In testing on real customer calls, our model matched or exceeded larger general-purpose models on key measures for identifying vulnerable moments. At the same time, it simplified the pipeline and reduced processing costs.

What this means: Faster triage. More consistent results. Fewer moving parts. Better oversight.



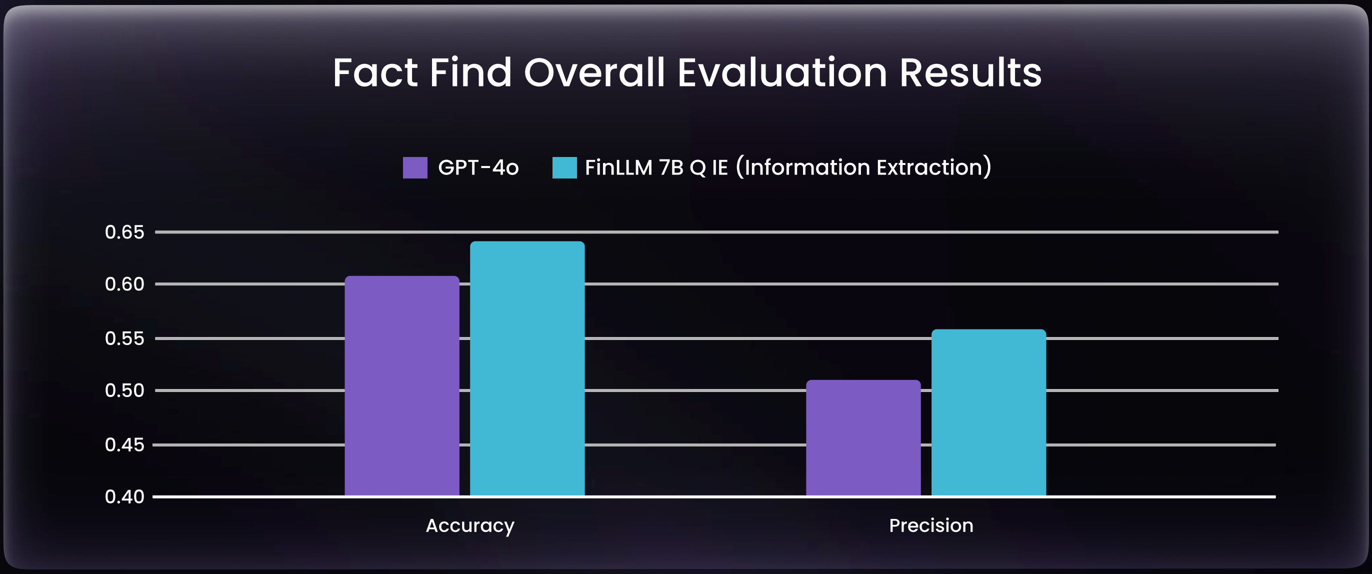
Results of vulnerability detection in Aveni Detect comparing an external LLM (GPT-4o), the existing Aveni Vulnerability Detector, and the new FinLLM 7B Q IE (Information Extraction) model. FinLLM 7B outperforms GPT-4o despite being a much smaller model, specifically we excel in Recall which is vital for correctly detecting vulnerabilities.

Aveni Assist: extracting data for FactFinds

Populating a FactFind is tedious work. We replaced a step in the key fact extraction pipeline with FinLLM and judged outputs using an independent model that grades semantic correctness.

The FinLLM variants trained specifically for information extraction achieved higher accuracy than generic models, particularly on the structured details that advisers care about.

What this means: Less manual rework. More reliable records. Better productivity for advisers.



Results of fact extraction in Aveni Assist comparing an external LLM (GPT-4o), and FinLLM 7B Information Extraction model. The FinLLM model outperforms GPT-4o in both accuracy and precision of identifying facts.

Safety, governance, and transparency from the start

You cannot add safety after deployment. Effective AI in financial services requires clear controls from the beginning, with every decision shaped by privacy, regulatory expectations and operational risk. Our approach focused on designing safety into the model, the data and the deployment process rather than relying on late-stage filters.

> “Safety is not a filter you apply at the end. It has to be embedded in every decision, from data sourcing to model design to deployment.”
>
> - Nicole Nisbitt, Senior NLP Engineer

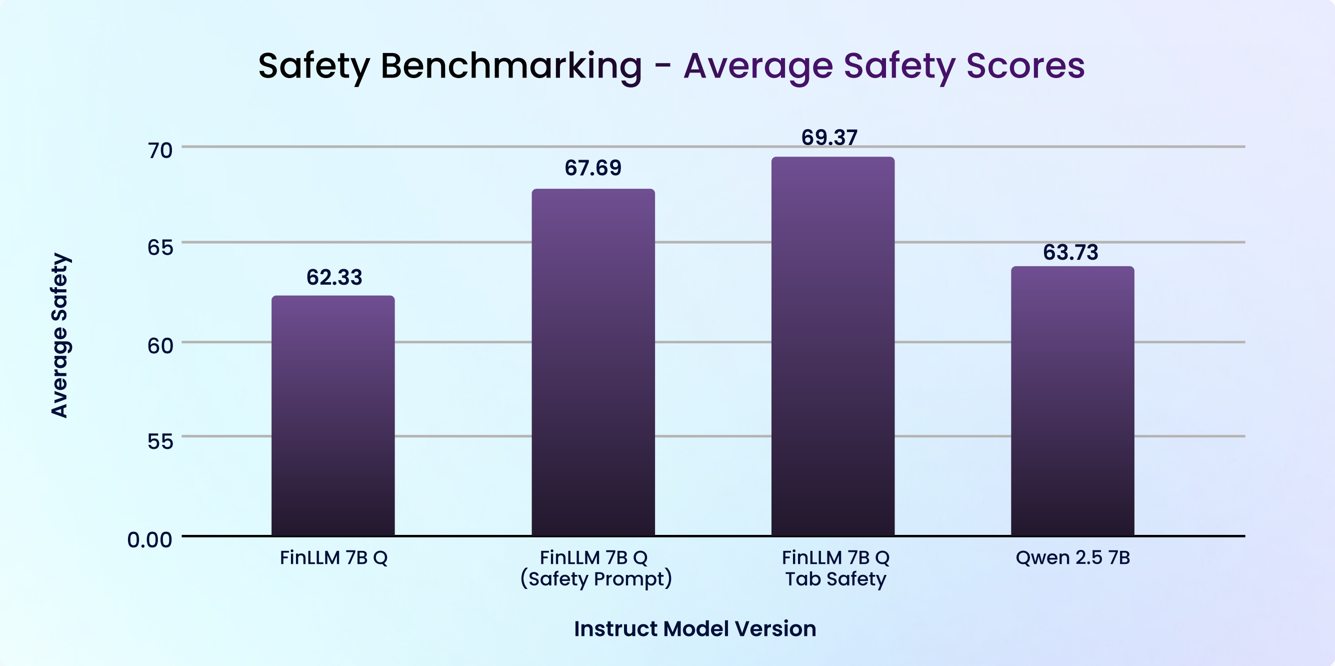
Governance

Strong governance starts with documentation. We maintain clear records of data sources and licences, conduct Data Protection Impact Assessments where required, and align with emerging regulation including the EU AI Act. We track environmental impact across training, with plans to expand this reporting to cover inference as operations scale.

Measured safety performance

We tested across multiple categories: toxicity, bias, hallucination, misinformation, intellectual property, and privacy. Simple operational measures, such as clear system instructions, improved safety scores. Targeted fine-tuning lifted both safety and task performance.

Some areas, such as specific bias categories, require deeper work and we will address these through preference-based alignment supported by human oversight.



Results of Aveni Safety benchmarking showing FinLLM models (safety prompt and safety SFT) outperforming Qwen 2.5 7B.

What this means for your organisation

FinLLM was built for the reality of running AI in financial services.

It understands UK financial language and regulation because that's what it was trained on. It works on the tasks your teams actually do because that's how we tested it.

Firms use FinLLM through the Aveni platform, where it powers defined features and workflows. This provides consistent performance, clear governance and the right level of control for regulated users. As the platform grows, organisations will gain more ways to use FinLLM through specific tasks and outcomes rather than direct model access.

Compliance and risk teams get what they need. They see documented data sources, clear controls and benchmarked performance they can verify.

The architecture is simpler than traditional multi-model pipelines, which reduces complexity and cost.

And as regulation evolves, the model evolves with it.

What we achieved this year

- **Built a UK financial dataset:** We created a curated collection of regulatory guidance, educational materials, and financial content. Every source was checked for quality, legal use, and privacy compliance.

- **Created tests that reflect real work:** We built an evaluation suite based on actual financial tasks, not generic benchmarks. We also released a public subset so other organisations can verify our claims.

- **Launched two versions in twelve months:** May 2025 brought our first model trained on financial use cases. By September, we had released an improved version that handles structured data better and includes stronger safety controls.

- **Partnered with NVIDIA for training infrastructure:** We used NVIDIA NeMo and accelerated computing to train the model efficiently. This approach aligns with best practices for building specialised models in regulated sectors.

- **Generated training material to fill gaps:** Some financial skills rarely appear in public data. We created additional training examples focused on tables and numerical reasoning. This helped our September model outperform larger general-purpose alternatives on financial benchmarks.

- **Proved the model works in production:** We deployed FinLLM in live Aveni products. It matched or beat larger general models on accuracy whilst simplifying our technical infrastructure.

- **Embedded safety at every stage:** We built a multi-layer safety and governance framework designed for regulated deployment. Every model iteration went through the same rigorous controls.

What comes next

From model to platform capability

The next phase is about turning FinLLM from a standalone model into the intelligence layer that powers the wider Aveni Agentic Platform. Instead of operating in isolation, it will support multiple workflows behind the scenes. Whether a firm is analysing a case, reviewing compliance activity or generating customer insights, the same underlying intelligence will deliver consistent reasoning and reliable outputs. Build it once, apply it across the platform.

Fine-tuning for your organisation

We are preparing capabilities that allow firms to adapt the model to their own requirements. This includes teaching it your tone of voice, training it on internal policies and giving it context from your operational data. All of this will sit within a governed framework that protects your information and preserves model integrity. The result is a version that reflects the way your organisation works without compromising safety or control.

Continuous improvement

The model will develop stronger reasoning, improved handling of long and complex documents and more robust safety behaviour. Every production deployment will feed structured insights back into the next iteration. Human oversight remains part of the loop to ensure that each improvement strengthens both performance and assurance.

Our purpose and vision

One year in, FinLLM has moved from concept to capability. It understands the work. It respects the rules. It performs reliably in production. The partnership with Lloyds Banking Group and Nationwide helped ensure those foundations were shaped by real financial processes, strengthening both governance and practical performance. With that first chapter complete, the work now shifts to scale and platform integration. The model will continue to evolve, and firms will gain the intelligence needed to automate with confidence. The future of financial services AI is specialised, assured and built on proven capability. That future has already begun.

The next chapter of financial AI is being written now.

Join us in shaping how responsible intelligence powers the industry.

hello@aveni.ai

aveni