SmarterFi.ai
AI for Community Banking

WHITEPAPER · NOVEMBER 2026

Why LLMs Are Not Enough for Community Banks

The case for a private large quantitative model

> How SmarterFI pairs language models with a private LQM back end, so the numbers a community bank relies on are repeatable, traceable and examinable

By James (Jamie) Donahue, Founder, President & Chief AI Officer, SmarterFI.ai

Cite as: Donahue, J. (2026). Why LLMs Are Not Enough for Community Banks: The case for a private large quantitative model. SmarterFI.ai. Version 1.1. https://doi.org/10.5281/zenodo.23884876. CC BY 4.0

Words from a language model
reading, extracting, drafting

Numbers from a quantitative model
built on your own records

Same inputs, same answer
repeatable by design

Every figure traced
back to its source records

1 Executive Summary

> **Our position**
> SmarterFI isn't just selling another SaaS widget. It has packaged **enterprise-grade quantitative AI, bank-specific security controls and regulatory risk mitigation** into a single platform designed specifically for community banks. The platform is designed to help banks manage regulatory risk; it does not replace a bank's own compliance, model risk or third-party risk programs.

Large language models (LLMs) have earned a place in community banking. They can read a long loan agreement in seconds, draft a credit memo and answer a staff question in plain English. SmarterFI uses them for exactly that work, and they are good at it.[1][2]

The decisions that matter most at a community bank, though, run on numbers: cash flow, coverage, concentrations, loss rates and liquidity. That is where language models fall short in ways a risk committee should care about. Their quantitative reasoning is fragile.[4] The same request can return different answers on different runs.[5] They can state wrong content with confidence, including explanations and citations that sound right and are not.[3] And on their own, they cannot show an examiner which records produced a figure.

SmarterFI's answer is “and”, not “or”: **the language model does the talking; the quantitative model does the math.**

- **Words from the language model.** LLMs read, extract and draft, and every extracted value must cite its source or it is discarded. LIVE BUILT
- **Numbers from tested code and from a private LQM.** Every calculation runs in versioned, tested, deterministic code.[2] LIVE The large quantitative model (LQM) adds forecasting and risk estimation. It is trained for quantitative work on an institution's own records and runs in SmarterFI's private cloud. RUNNING
- **Repeatable and traceable by design.** The LQM is designed so that the same inputs give the same answer, every output links to the records behind it, and it states its confidence or says “not enough data” instead of guessing. RUNNING
- **Private by design.** The platform and the LQM run in SmarterFI's private cloud. Data the LQM uses is obfuscated and protected with an encryption key the bank owns, every model request passes through a governed model gateway, and bank data is not used to train models for any other customer.[2] RUNNING
- **A person decides.** The platform prepares the work; an authorized person at the bank makes every consequential decision, and it is logged.[2] BUILT

There is a regulatory reason this design matters. The agencies' revised model risk guidance defines a model as a quantitative method that turns input data into quantitative estimates, and it places generative AI outside its scope.[6][7] An LQM is a model in that familiar sense. It can be documented, validated, monitored and governed with the discipline community banks and their examiners already know. We believe SmarterFI is among the first to design a private LQM back end specifically for community banks and credit unions.

How to read the status labels

Every claim about SmarterFI carries a status so reviewers can weigh the evidence. LIVE deployed in SmarterFI's production environment on Microsoft Azure or published as a platform commitment on smarterfi.ai. BUILT implemented in SmarterFI's application code and demonstrated in prototypes that run on synthetic data only. RUNNING built and operating in SmarterFI's private cloud. SPECIFIED required by SmarterFI's product requirements or defined in the customer agreement. **The LQM is running.** This paper explains its design and the evidence an institution should expect before relying on it. It does not report accuracy or validation results.

2 What Language Models Do Well

An LLM is software trained on very large amounts of written text so it can predict and produce language. That training makes it excellent at work that is mostly about words, and a great deal of banking work is about words. Credit is due before we discuss limits.

Task	Why an LLM fits	SmarterFI example	Status
Reading unstructured documents	Tax returns, loan agreements, policies and vendor contracts are written for people, not databases	Lending Co-Pilot spreads tax returns with K-1 tracing; Contract & Invoice Assurance turns contract terms into structured data[1]	BUILT
Extracting terms with citations	The model can point to the page and clause it read	Extraction is schema-constrained with page and clause citations; a value whose citation does not check out is discarded[2]	BUILT SPECIFIED
Answering staff questions	Plain-English questions against a known library	Policy Assistant answers only from the bank's own policies, with the passage it came from[1]	BUILT
Drafting and explaining	Memos, letters and summaries in the institution's voice	Source-cited credit memos; drafted letters receive numbers injected from data, not written by the model[1][2]	LIVE BUILT

Notice what these tasks have in common. The model reads and writes. It does not decide, and it does not produce the number of record. In SmarterFI's applications today, dollar figures, ratios and variances are computed in versioned, tested code, and totals are precomputed before any model sees them.[2] LIVE BUILT

3 Where Language Models Fall Short for Banking

A language model learns how people *write about* the world. It did not learn a bank's ledgers, payment histories or loss records, and it produces answers by predicting likely words. Five gaps follow from that, and each one maps to a question a risk committee or examiner will ask.

Gap	What the public evidence shows	Why it matters at a community bank
Quantitative reasoning	Apple researchers found that model performance on grade-school math declined when only the numbers in a question were changed, and that adding one plausible but irrelevant clause caused drops of up to 65% across state-of-the-art models.[4]	A covenant test or a coverage ratio is a chain of steps. Fragile reasoning produces a wrong number inside fluent, convincing prose.
Determinism and repeatability	Thinking Machines Lab sampled 1,000 completions of the same prompt at temperature 0 and received 80 different completions. Its analysis traces the main cause to server load changing how requests are batched.[5]	If an examiner reruns the analysis next quarter, the answer should be the same. With a hosted LLM alone, that is not guaranteed.
Explainability	NIST notes that generative AI outputs may include confabulated logic or citations that purport to justify the answer, and that models sometimes give reasoning steps even when the answer is wrong.[3]	A model's account of its own reasoning is more generated text. It is not evidence of which records produced a figure.
Confabulation (hallucination)	NIST defines confabulation as confidently stated but erroneous or false content, calls it a natural result of how generative models are designed, and flags it as especially important in consequential decisions.[3]	In lending, a confident wrong number is worse than "insufficient data".
Data privacy	NIST notes that models may leak, generate or correctly infer sensitive information.[3] The interagency information security guidelines require banks to protect customer information and to oversee service providers by contract.[11]	Whatever reaches an outside model is a third-party exposure that the bank must diligence, contract for and monitor.

> **Isn't a calculator enough?**
> It is a fair question. An LLM can be given a calculator, a database or code to run, and SmarterFI already works this way: the model extracts and cites, and tested code does the arithmetic.[2] Formulas, however, handle known arithmetic. They do not forecast a borrower's cash flow from transaction history, or show how a loan compares with similar loans in the bank's own book. Once a separate engine does that quantitative work, that engine is the part that must be built, documented and validated to bank standards. That is the job of the LQM.

4 What a Large Quantitative Model Is

A **large quantitative model (LQM)** is a model built for quantitative work. It learns directly from a domain's numerical records in their original form (for a bank: accounts, transactions, loans, payments and losses) rather than from text that describes them. The category is drawing analyst attention: Gartner's March 2026 research positions LQMs as the coming standard for finance and science problems.[10]

A simple way to picture the split: the LLM is the friendly banker at the front desk who explains things clearly. The LQM is the back office that runs the numbers and keeps the records. You want both, and

you would not let the front desk set loan terms from memory.

SmarterFI's LQM is designed around four properties. RUNNING

1. **Learn from the records, not descriptions of them.** The model is trained for quantitative work on the institution's own records, mapped and normalized through Data Mapping & Discovery, and obfuscated and protected with an encryption key the bank owns.
2. **Keep a structure that can be inspected.** Reviewers can see how the model groups customers or loans, rather than the knowledge being hidden.
3. **Trace every output to its source records.** Each figure carries references to the records that produced it.
4. **State how sure it is, and decline when evidence is thin.** This is calibrated uncertainty: when the model says an outcome is 90% likely, it should be right about 90% of the time. When there is not enough history, the answer is "not enough data".

	Large language model (LLM) Does the talking	Large quantitative model (LQM) Does the math RUNNING
Learns from	Text: how people write about the world	The institution's own numerical records in their original form, obfuscated and protected with a key the bank owns
Best at	Reading, extracting, drafting and conversation	Forecasts, risk estimates and patterns in numbers
Same inputs, same answer?	Not guaranteed on hosted services	By design: repeatable
Where did a number come from?	Cannot show on its own; only if wired to records	By design: references to the source records
When evidence is thin	May answer confidently anyway	By design: a range with stated confidence, or "not enough data"
Under SR 26-2	Generative AI: outside the guidance's scope	A model in the guidance's ordinary sense
At SmarterFI	Azure OpenAI LIVE	Private LQM in SmarterFI's private cloud RUNNING

Figure 1. Two engines, two jobs. The language model reads and writes; the quantitative model produces and explains the numbers. LQM properties are design properties; this paper reports no measured accuracy results.

5 Why an LQM Fits Model Risk Management

On April 17, 2026 the Federal Reserve, OCC and FDIC issued Revised Guidance on Model Risk Management (Federal Reserve SR 26-2; OCC Bulletin 2026-13), which supersedes SR 11-7 and OCC Bulletin 2011-12.[6][7] Three points in it matter for this discussion.

- **What counts as a model.** The guidance defines a model as "a complex quantitative method, system, or approach that applies statistical, economic, or financial theories to process input data into quantitative estimates," and excludes simple arithmetic and deterministic rule-based processes.[7]
- **Generative AI is out of scope.** The agencies state that generative and agentic AI models are novel and rapidly evolving and are not within the scope of the guidance, and they plan a request for

information on banks' use of AI.[7]

- **Community banks.** The guidance is expected to be most relevant to larger banking organizations, is applicable to all community banks subject to its limitations, and does not set enforceable standards.[7] The OCC has also clarified that its guidance does not require community banks to perform annual model validation.[7]

Put together, this is good news for a community bank. A language model sits in territory the guidance leaves open. A quantitative model that produces forecasts and risk estimates is a model in the guidance's ordinary sense, so the bank can govern it with practices its examiners already recognize: sound development, validation that tests conceptual soundness and outcomes, ongoing monitoring, and governance and controls, including considerations for vendor products.[7] SmarterFI's deterministic formulas remain what they are, software rather than models, and are inventoried that way.[2]

Guidance theme	What the bank will ask	SmarterFI LQM design response	Status
Development and use	Intended use, limits, data and testing	Per-feature model documentation: intended use, scope and known limitations, data categories, testing and results, monitoring metrics and change log[2]	SPECIFIED
Validation	Can a validator rerun it and test outcomes?	Repeatable outputs so a validator can rerun any result; cooperation with the bank's own or independent validator, including test environments and results[2]	SPECIFIED
Ongoing monitoring	Is it still performing as intended?	Monitoring for degradation and drift; prompt notice of issues likely to cause materially inaccurate outputs; acceptance, override and reviewer-edit metrics[2]	SPECIFIED BUILT
Governance and controls	Inventory, change control, accountability	Model inventory per application; pinned versions; advance notice before a material change, as defined in the customer agreement[2]	LIVE SPECIFIED
Vendor products	What is inside, and who stands behind it?	Components and configuration disclosed on request and on the subcontractor list; detailed materials available under confidentiality[2]	SPECIFIED
Fair lending	Can adverse-action reasons be supported?	Principal factors and source data shown where an output supports a credit decision; no prohibited-basis inputs or known proxies; disparate-impact testing before release[2]	SPECIFIED

Third-party risk management applies alongside model risk management. The interagency guidance on third-party relationships applies to all banks with third-party relationships and scales oversight to the risk and criticality of the activity,[8] and the FFIEC IT Examination Handbook's Outsourcing Technology Services booklet sets out how examiners evaluate service-provider oversight.[9]

6 Why the Back End Must Be Private

A quantitative model is only as useful as the records it learns from, and for a community bank those records are among its most sensitive assets. That is why SmarterFI's LQM back end is private by

design. Three reasons stand out.

Your data stays yours

Each institution's data lives in its own Azure environment in U.S. regions, isolated from other customers, with institution-specific encryption keys.[2] LIVE Language-model requests go only to Azure-sold models through the institution's private endpoint, and production has no route to public AI services.[2] LIVE Quantitative work goes to SmarterFI's own LQM, which runs in SmarterFI's private cloud, so no public AI service is added to the data path. RUNNING The design keeps that path narrow:

- Customer records stay in the institution's dedicated environment. Data the LQM uses is **obfuscated and protected with an encryption key the bank owns**. RUNNING
- Every model request, to the language model or the LQM, passes through one model gateway that applies PII detection and redaction and records that redaction occurred.[2] RUNNING
- SmarterFI's requirements for the LQM are that it is hosted by SmarterFI in the United States, is reached only through the governed model gateway, is not trained on bank data for any other customer, and that any third-party component is disclosed on SmarterFI's subcontractor list. Hosting specifics are defined in the customer agreement. SPECIFIED
- Bank data is not used to train, fine-tune or improve any model used for another customer.[2] LIVE SPECIFIED

Your numbers are your advantage

Every vendor can rent the same general-purpose language model, so “we use model X” gives no bank a reason to choose anyone. A quantitative model built for an institution on its own records reflects its own borrowers, its own markets and its own history. Data Mapping & Discovery is the starting point for that work: the Data Mapper process maps a bank's data endpoints and can run inside the bank's environment or in SmarterFI's private cloud.[1][2] LIVE

Your examiners can see it

A private back end means change happens on the institution's schedule, not a provider's. Model versions are pinned; material changes come with advance notice, release notes and, where feasible, a test environment first; and the append-only audit log records prompts, outputs, model versions and user actions, exportable for examiners.[2] LIVE SPECIFIED Because the model never decides and its numbers carry their sources, the audit question reduces to three verifiable facts: which records a value came from, which model or rule produced it, and which named person made the decision.[2]

7 The SmarterFI Architecture: The LLM Talks, the LQM Does the Math

The SmarterFI AIOS is a multi-agent, policy-governed platform with governance between every agent and every action.[1] SmarterFI hosts the platform and the LQM for every institution in SmarterFI's private cloud. RUNNING Models plug in underneath it through one model gateway, so a model can be added or replaced without changing the controls around it. RUNNING Figure 2 shows where the private LQM fits.

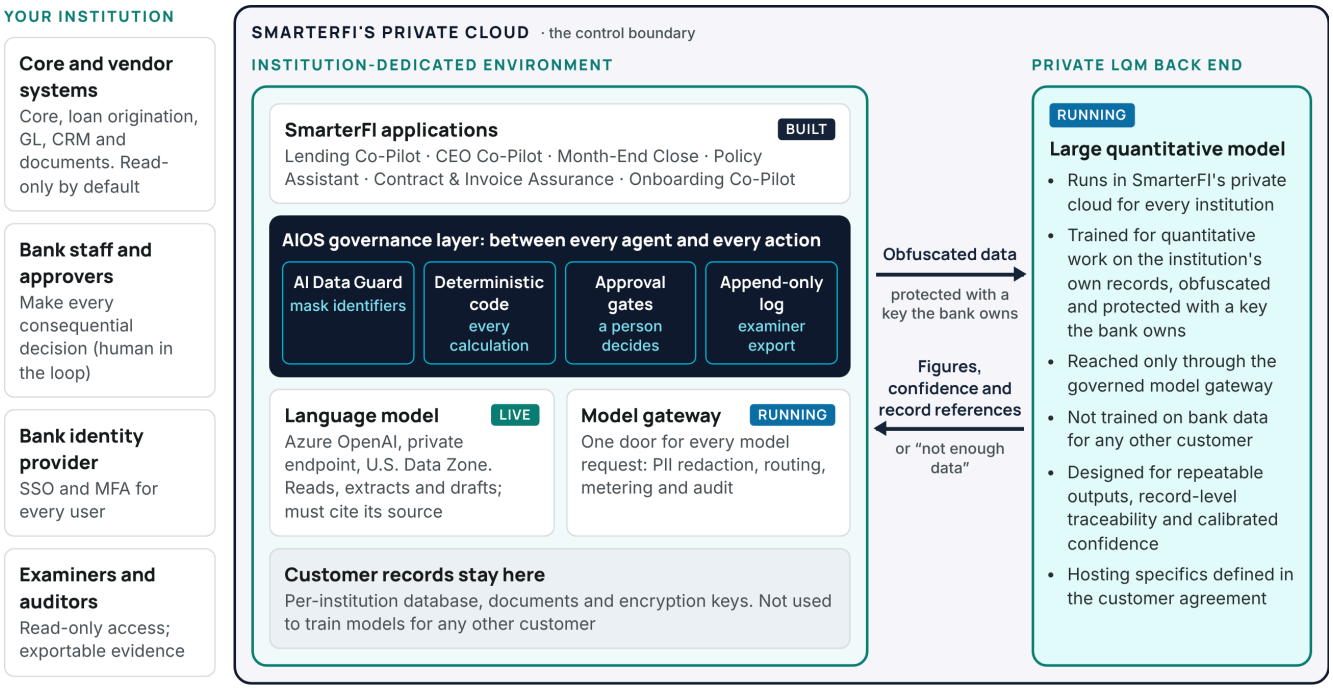


Figure 2. The private LQM back end inside SmarterFI's private cloud, which hosts the platform and the LQM for every institution. Customer records and every control stay in the institution's dedicated environment; data the LQM uses is obfuscated and protected with an encryption key the bank owns; every model request passes through the model gateway, and the LQM returns figures with stated confidence and record references. Hosting specifics are defined in the customer agreement.

Component	Role	Status
SmarterFI applications	Lending Co-Pilot, CEO Co-Pilot, Month-End Close, Policy Assistant and the rest of the catalog[1]	BUILT
AIOS governance layer	AI Data Guard masking before any model call, approval gates, feature enablement and the append-only audit log[2]	LIVE BUILT SPECIFIED
Deterministic computation	Versioned, tested code for every dollar figure, ratio and variance[2]	LIVE BUILT
Language model	Azure OpenAI through a private endpoint in the U.S. Data Zone: reads, extracts and drafts, with citations[2]	LIVE
Model gateway	One door for every model request: routing, PII redaction, metering and audit[2]	RUNNING
Private LQM	Runs in SmarterFI's private cloud: forecasts and risk estimates as ranges with stated confidence and record references, or "not enough data"	RUNNING
Data Mapping & Discovery	Maps the institution's data sources, names a source of truth for each kind of data and normalizes it into one common language[1]	LIVE

One request, end to end

Consider a small-business line of credit. The example is illustrative and carries no figures; the point is the division of labor.

1. **Gather.** Lending Co-Pilot collects the documents. The language model reads the tax returns and statements and extracts values with page citations; AI Data Guard masks identifiers before any model call. BUILT
2. **Calculate.** Tested code spreads the financials and computes ratios such as debt service coverage. LIVE BUILT
3. **Forecast.** Through the gateway, the LQM works from the borrower's deposit and payment history and comparable loans in the bank's own history, obfuscated and protected with an encryption key the bank owns. It returns a cash-flow range with stated confidence and references to the records behind it, or “not enough data” if the history is too short. RUNNING
4. **Write.** The language model drafts the credit memo, quoting the code's and the LQM's figures with their references. It does not create numbers. BUILT
5. **Decide.** The credit officer reviews, edits and decides. Every step, from extraction to approval, is written to the audit log. BUILT

8 What Is Proven and What Is Not Yet

Risk committees are right to separate design from evidence. Here is where things stand.

Item	Where it stands
Language-model controls	Citation checks, identifier masking, deterministic computation, human approval and the append-only audit log are live or built, as documented control by control in SFI-2026-044.[2]
The LQM	Running in SmarterFI's private cloud, for every institution. This paper reports no accuracy, validation or customer results.
Accuracy and validation evidence	Before an institution relies on LQM output, it should receive the model documentation and test results, and its validator should be able to rerun results in a test environment, consistent with SmarterFI's validation commitments.[2]
Model gateway and redaction	Running. Every model request, to either model, passes through the gateway, which redacts identifiers and logs the request.
LQM hosting and contract terms	Hosted in SmarterFI's private cloud. U.S. hosting, no training on bank data and subcontractor disclosure for any third-party component are requirements; hosting specifics are defined in the customer agreement.
Timing	This paper does not state a production date.

> **Our commitment**
> We will call the LQM proven when the evidence says so, and not before. Until then, every statement in this paper about how the LQM performs is a design property, not a measured result.

9 Questions to Ask Any AI Vendor About Numbers

Whether or not you work with SmarterFI, these questions help a risk committee see where the numbers in an AI system really come from.

- [ ] Which component produces each number we will rely on: a language model, tested code or a quantitative model?
- [ ] If we run the same request twice, do we get the same answer? Can you show us?
- [ ] Can each figure be traced to the records that produced it?
- [ ] What does the system do when the evidence is thin?
- [ ] Is the quantitative component a model under SR 26-2, and where are its documentation and validation evidence?
- [ ] What leaves our environment, in what form, to whom, and where is it hosted?
- [ ] Is our data used to train models for anyone else?
- [ ] Who approves each consequential decision, and is that approval logged?
- [ ] How are model changes announced and tested before they reach us?
- [ ] Can we export our data, outputs and audit logs, and leave?

10 Conclusion

The takeaway for a bank is a simple test: for every number an AI system produces, ask where it came from and whether you could defend it to an examiner. Words can come from a language model. The figures behind a credit decision, a reserve or a board report should come from tested code or from a model that is documented, validated and monitored like any other model the bank relies on.

That is the gap SmarterFI was built to close. Instead of one more point solution, it brings quantitative AI built for banking, security controls built around how banks operate and safeguards designed to help manage regulatory risk together on one platform made for community institutions.

What to do next: put the questions in section 9 to every AI vendor you are evaluating, including SmarterFI, and bring your risk, compliance and audit leads into that conversation early. The answers will tell you quickly which systems you can govern and which you would only be trusting.

Let's work through it together

I would welcome the chance to walk your team through this in a 30-minute working session: how the language model and the LQM divide the work, how your data is protected, and what your risk committee and examiners would review. Bring your toughest questions.

BOOK A SESSION WITH JAMIE
calendly.com/jamie-donahue-smarterfi/30min

LEARN MORE
www.smarterfi.ai

James (Jamie) Donahue, Founder, President & Chief AI Officer, SmarterFI.ai

A Appendix: Sources

Bracketed numbers in the text, such as [6], refer to the numbered sources below. All web sources were accessed and checked on October 29, 2026. SmarterFI sources describe SmarterFI's own design and status; public sources support general statements about language models and regulation and do not refer to SmarterFI.

1. SmarterFI.ai, Platform & Solutions and Technology & Architecture pages (accessed October 29, 2026). https://smarterfi.ai/platform ; https://smarterfi.ai/platform/technology
2. SmarterFI.ai, SFI-2026-044, Eliminating Frontier Model Risk: SmarterFI Technical Architecture Brief, Version 1.1 (November 2026). https://doi.org/10.5281/zenodo.23884799
3. NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (July 2024), Sections 2.2, 2.4 and 2.9. https://doi.org/10.6028/NIST.AI.600-1
4. Mirzadeh, I. et al. (Apple), "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models," arXiv:2410.05229. https://arxiv.org/abs/2410.05229
5. He, H. and Thinking Machines Lab, "Defeating Nondeterminism in LLM Inference" (September 10, 2025). https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/
6. Board of Governors of the Federal Reserve System, SR 26-2, Revised Guidance on Model Risk Management (April 17, 2026). https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm
7. OCC Bulletin 2026-13, Model Risk Management: Revised Guidance (April 17, 2026). https://www.occ.gov/news-issuances/bulletins/2026/bulletin-2026-13.html
8. OCC Bulletin 2023-17, Third-Party Relationships: Interagency Guidance on Risk Management (June 6, 2023). https://www.occ.gov/news-issuances/bulletins/2023/bulletin-2023-17.html
9. FFIEC, IT Examination Handbook InfoBase, including the Outsourcing Technology Services booklet (accessed October 29, 2026). https://ithandbook.ffiec.gov/
10. Gartner, "Emerging Tech: AI Vendor Race: Large Quantitative Models Usher In a New AI Era, Leaving LLMs in the Dust," research abstract (March 17, 2026). https://www.gartner.com/en/documents/7604265
11. 12 C.F.R. Part 364, Appendix B, Interagency Guidelines Establishing Information Security Standards (issued under GLBA, 15 U.S.C. 6801 and 6805(b)). https://www.ecfr.gov/current/title-12/chapter-III/subchapter-B/part-364/appendix-Appendix%20B%20to%20Part%20364

Document control SFI-2026-073 · Version 1.1 · November 3, 2026 · Author: James (Jamie) Donahue · Reviewer: Internal · Approver: James Donahue · Status: Approved for publication

Cite as Donahue, J. (2026). Why LLMs Are Not Enough for Community Banks: The case for a private large quantitative model. SmarterFI.ai. Version 1.1. https://doi.org/10.5281/zenodo.23884876. CC BY 4.0

Revision Version 1.1 (November 3, 2026): added citation and DOI line; licence notice aligned to CC BY 4.0; SFI-2026-044 source reference updated. No change to findings. Version 1.0 published November 2, 2026.

SmarterFi.ai
AI for Community Banking

ABOUT THE AUTHOR

James (Jamie) Donahue is Founder, President & Chief AI Officer of SmarterFI.ai (February 2026–present), where he is building AI for community banks and credit unions. His career spans banking, fintech and financial technology providers. He was previously President & Chief Digital Officer of First Carolina Bank's BM Technologies Division (2025–2026) and President & Chief Technology Officer of BM Technologies (2020–2025). Earlier, he served as Global Head of Cloud at Finastra and SVP of Cloud Engineering at First Republic, and held CIO, CISO and SVP technology roles, including more than ten years in life and annuity-related IT. That combination of executive, security and cloud leadership shapes SmarterFI's approach: governance first, with the bank in control of its data and its decisions. A US Navy veteran with an electrical engineering background, he holds a BS in Workforce Development from Southern Illinois University Carbondale and completed graduate coursework in Information Resources Management at Central Michigan University. He is based in Palm Beach, Florida. LinkedIn: linkedin.com/in/jamesmdonahue

WEB
www.smarterfi.ai

EMAIL
jamie.donahue@smarterfi.ai

CITE AS

Donahue, J. (2026). Why LLMs Are Not Enough for Community Banks: The case for a private large quantitative model. SmarterFI.ai. Version 1.1. https://doi.org/10.5281/zenodo.23884876. CC BY 4.0

© 2026 SmarterFI.ai. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0). Describes SmarterFI's architecture and design as of November 2026; the platform and the LQM run in SmarterFI's private cloud. Status labels indicate the evidence for each statement. Not legal advice; does not modify any agreement or imply endorsement by any organization cited.