



CORA RESEARCH NOTE · OCTOBER 2026

Beyond the Prompt

Making AI Repeatable in Lending Workflows

Prompt-based and embedded AI: what the evidence says about repeatability, auditability and scale.

Alex Parisi
Co-Founder, CORA



ABSTRACT

General-purpose AI assistants are now on nearly every credit desk. The same underlying models can be put to work in two different ways: through prompts, where a person composes a request and interprets the answer, or embedded inside a purpose-built workflow, where inputs, outputs, validation and logging are fixed by the system rather than by the user. This note reviews the published evidence on both modes, drawing on field experiments, enterprise surveys, technical studies of model behavior and U.S. supervisory guidance, and applies it to portfolio monitoring in asset-based and private credit lending, where the same borrowing base certificates, covenant packages and financial reports arrive every period across hundreds of borrowers. The evidence supports three conclusions. Prompt-based use produces real but uneven gains that have largely not translated into institutional value. Several of its limitations, including variability across users and model versions, degradation on long documents and the absence of an audit trail, are structural rather than matters of user skill. Embedded systems address those limitations by design. The appropriate choice depends less on which model is used than on whether the task is recurring, structured and must be defended after the fact.

IN THIS NOTE

01	Executive Summary
02	Two Ways of Putting a Language Model to Work
03	What the Evidence Says About Prompt-Based Use
04	Why Ad-Hoc Prompting Resists Repeatability
05	What Embedding Changes, and What It Does Not
06	The Lending Case: Portfolio Monitoring as a Test
07	A Decision Framework for Institutions
08	Conclusion
—	References

SECTION 01

Executive Summary

This note asks a narrow question: when an institution puts a large language model to work on recurring, high-stakes document review, does it matter whether the model is used through prompts or embedded in a system? The published evidence, summarized below and cited in full in the References, points to seven findings.

80%+ of firms using gen AI report no tangible enterprise-level EBIT impact [8]	76 pts accuracy swing from prompt formatting alone, same model, same task [17]	84→51% one model’s accuracy on the same questions, three months apart [19]	94% of operational spreadsheets found to contain errors in field audits [33]
--	--	--	--

- **Prompt-based AI works, for individuals.** Controlled experiments report time savings or output gains in the range of 12 to 40 percent on writing, consulting and customer-support tasks, with the largest gains among less experienced workers [2, 3, 4].
- **Those gains have not reliably reached the P&L.** More than 80 percent of McKinsey’s 2025 respondents reported no tangible enterprise-level EBIT impact from generative AI [8]; MIT’s 2025 study found that general-purpose chatbots “primarily enhance individual productivity, not P&L performance” [7]; BCG found 60 percent of firms achieving no material value [11].
- **Variability is a property of the technology, not of the user.** Public model endpoints return different answers to identical queries even at temperature zero [16]; prompt formatting alone moves accuracy by up to 76 points [17]; one model’s accuracy on a simple task fell from 84 to 51 percent in three months [19]; providers retire models on as little as 60 days’ notice [20, 21].
- **The binding constraint is learning and workflow integration, not model capability.** MIT: “The core barrier to scaling is not infrastructure, regulation, or talent. It is learning” [7]. McKinsey: of 25 attributes tested, redesign of workflows has the biggest effect on EBIT impact [8].
- **Embedded systems are not a free lunch.** Enterprise tools are also rejected for “brittle workflows” [7]; Gartner expects more than 40 percent of agentic AI projects to be canceled by 2027 [1]; Treasury flags dependence on a small number of AI vendors [27]; over-reliance applies to any tool that is usually right [6].
- **For lenders, the supervisory frame now distinguishes stochastic generative AI from deterministic rule-based processes.** Revised interagency model risk guidance issued in April

2026 excludes “deterministic rule-based processes” from the definition of a model and places generative AI outside formal scope while directing institutions to govern it through their own risk-management practices [25]. Fed Governor Barr has said bank decisions must be “well-controlled, numerically and legally precise, explainable, and replicable” [15].

- **A practical rule follows.** Use prompts for one-off, exploratory, human-reviewed work. Embed the model for recurring, structured, high-volume work whose output must be consistent across people and periods and traceable afterward.

SECTION 02

Two Ways of Putting a Language Model to Work

The distinction this note draws is functional, not technical. The two modes often use the same models from the same providers. What differs is where the knowledge about the task lives.

Prompt-based use. A person opens a general-purpose assistant such as ChatGPT, Claude, Copilot or Gemini, composes a request, supplies whatever context they judge necessary and interprets the response. The instruction, the input and the judgment about the output all sit with the individual. The interaction is a conversation, and the product is a single answer at the end of it.

Embedded use. The model operates inside a purpose-built application. The system, not the user, determines which documents are accepted, what is extracted from them, into which schema, against which validation rules, and what happens when something does not reconcile. The person reviews exceptions rather than composing instructions. Outputs are written to a system of record and logged.

The boundary is a spectrum rather than a line. Prompt libraries, custom assistants, shared “projects” and retrieval tools add structure to prompting; some products marketed as embedded are little more than a prompt behind a button. Gartner has warned of “agent washing”, which it describes as “the rebranding of existing products, such as AI assistants, robotic process automation (RPA) and chatbots, without substantial agentic capabilities” [1]. Labels are unreliable. The test used throughout this note is: who fixes the task definition, and is the result reproducible and logged?

TABLE 1 The two modes compared

DIMENSION	PROMPT-BASED	EMBEDDED
Who defines the task	The individual user, each time	The system designer, once

DIMENSION	PROMPT-BASED	EMBEDDED
Input	Whatever the user pastes or uploads	Fixed document types mapped to a schema
Output	Free-form answer	Structured fields, flags and exceptions
Consistency across users	Depends on prompting skill	Enforced by design
Consistency across time	Changes with model updates	Versioned; changes tested before release
Validation	By the user, if at all	Rules and reconciliation built in
Audit trail	Chat history, if retained	Logged inputs, outputs and overrides
Learning	Starts fresh each session	Context retained; corrections persist
Best suited to	Drafting, exploration, one-off questions	Recurring, structured, high-volume work
Primary risk	Variability, data leakage, key-person dependence	Vendor dependence, brittle integration, over-reliance

SECTION 03

What the Evidence Says About Prompt-Based Use

3.1 The gains are real at the task level

Three widely cited experiments establish that general-purpose models improve individual output on bounded knowledge tasks. Noy and Zhang randomized 453 college-educated professionals to ChatGPT on occupation-specific writing tasks and found that “the average time taken decreased by 40% and output quality rose by 18%” [2]. Dell’Acqua and colleagues ran a field experiment with 758 consultants at Boston Consulting Group; on 18 realistic tasks within the model’s capabilities, consultants using GPT-4 “completed 12.2% more tasks on average, and completed task 25.1% more quickly” and produced “more than 40% higher quality compared to a control group” [3].

Brynjolfsson, Li and Raymond studied more than 5,000 customer-support agents and found that access to an AI assistant raised issues resolved per hour “by 14% on average, including a 34% improvement for novice and low-skilled workers but with minimal impact on experienced and highly skilled workers” [4].

One detail of the third study matters for everything that follows. The tool was not a chatbot. It was embedded in the agents' own console: it "monitors customer chats and provides agents with real-time suggestions for how to respond" [4].

3.2 The gains are uneven, and users misjudge them

The same experiments show that the benefit depends on who is using the tool and on what. In the consulting study, those below the average performance threshold improved by 43 percent and those above by 17 percent [3]; in the writing study, "inequality between workers decreased" [2]. The tool compresses the distribution, which is good news for new analysts and modest news for experienced ones.

The consulting study also included one task deliberately chosen to sit outside the model's capability. On that task, "consultants using AI were 19 percentage points less likely to produce correct solutions compared to those without AI" [3]. The authors describe a "jagged technological frontier" in which "some tasks are easily done by AI, while others, though seemingly similar in difficulty level, are outside the current capability of AI" [3]. The practical problem is that a user cannot tell, from inside a chat window, which side of the frontier a given task is on.

Nor are users reliable judges of whether the tool helped. In a 2025 randomized trial by METR, experienced open-source developers using AI tools took 19 percent longer to complete tasks; they had expected to be 24 percent faster, and after the study "still believed AI had sped them up by 20%" [5]. An earlier experiment with professional recruiters found that "as AI quality increases, humans have fewer incentives to exert effort and remain attentive, allowing the AI to substitute, rather than augment their performance" [6].

For an institution, the implication is that in a prompt-based regime outcomes vary by person, by task and by day, and the people doing the work are not well positioned to notice.

3.3 Individual productivity has not become institutional value

Enterprise surveys from 2024 and 2025 converge on the same picture: wide adoption, narrow return.

Tools like ChatGPT and Copilot are widely adopted. Over 80 percent of organizations have explored or piloted them, and nearly 40 percent report deployment. But these tools primarily enhance individual productivity, not P&L performance.

MIT NANDA, The GenAI Divide: State of AI in Business 2025 [7]

MIT's headline finding, that 95 percent of organizations are "getting zero return" on an estimated \$30 to 40 billion of generative AI investment, drew wide coverage [7]. It should be read with its method in mind: 52 structured interviews, 153 survey responses and a review of 300 public initiatives, released as a preliminary report. The larger surveys are less dramatic but consistent. McKinsey's March 2025 survey found that "more than 80 percent of respondents say their organizations aren't seeing a tangible impact on enterprise-level EBIT from their use of gen AI" [8]; its November 2025 edition found that 39 percent attribute any EBIT impact to AI, "and most of those respondents say that less than 5 percent of their organization's EBIT is attributable to AI use," with nearly two-thirds not yet scaling AI across the enterprise [9]. BCG reported in 2024 that "only 26% of companies have advanced beyond the proof-of-concept stage to generate value" [10], and in 2025 that "fully 60% of companies are not achieving material value at all" [11]. Deloitte found that "over two-thirds of respondents said that 30% or fewer of their current experiments will be fully scaled in the next three to six months" [12].

Meanwhile the prompt-based mode has spread on its own. MIT found that "while only 40% of companies say they purchased an official LLM subscription, workers from over 90% of the companies we surveyed reported regular use of personal AI tools for work tasks" [7]. Microsoft's 2024 Work Trend Index reported that "78% of AI users are bringing their own AI tools to work" and that "52% of people who use AI at work are reluctant to admit to using it for their most important tasks" [13]. The cost of that pattern is now measurable: IBM's 2025 breach study found that "one in five organizations reported a breach due to shadow AI, and only 37% have policies to manage AI or detect shadow AI," with such organizations observing "an average of \$670,000 in higher breach costs" [14].

The gap between individual gains and institutional value is the central puzzle. Sections 4 and 5 argue that it has a structural explanation.

SECTION 04

Why Ad-Hoc Prompting Resists Repeatability

The limitations described in this section are not failures of training or discipline. They are properties of how general-purpose models are built and served, documented in the technical literature and acknowledged by the providers themselves.

4.1 Same question, different answer

Because Gen AI usually involves stochastic processes, answers can differ in response to the same query asked at different times or to similar queries. This is tough to square with the requirements of banking, where decisions must be well-controlled, numerically and legally precise, explainable, and replicable.

Michael S. Barr, Governor, Federal Reserve Board, April 2025 [15]

Governor Barr's observation is borne out at the engineering level. Thinking Machines Lab showed in 2025 that "even when we adjust the temperature down to 0 (thus making the sampling theoretically deterministic), LLM APIs are still not deterministic in practice"; in their test, 1,000 identical requests to a model at temperature zero produced 80 distinct completions, because server load changes the batch size and therefore the arithmetic [16]. Variation also enters through phrasing. Sclar and colleagues found several widely used open-source models "extremely sensitive to subtle changes in prompt formatting," with "performance differences of up to 76 accuracy points" from formatting choices that do not change the meaning of the request [17]. Errica and colleagues describe the consequence for anyone building on these models: "a dreadful challenge: debugging LLMs' inconsistent behavior across minor variations of the prompt" [18].

On a credit desk this is the familiar experience of two analysts asking the same model the same question about the same borrowing base certificate and receiving different ineligible calculations. It is not a training problem. It is a property of the tool as deployed.

4.2 The model changes underneath the workflow

Chen, Zaharia and Zou tracked the behavior of two OpenAI models across a single quarter in 2023 and found that "GPT-4 (March 2023) was reasonable at identifying prime vs. composite numbers (84% accuracy) but GPT-4 (June 2023) was poor on these same questions (51% accuracy)" [19]. Their conclusion applies generally: "the behavior of the 'same' LLM service can change substantially in a relatively short amount of time, highlighting the need for continuous monitoring of LLMs," particularly because "when and how these models are updated over time is opaque" [19].

Providers are explicit that this is the operating model. OpenAI's documentation states: "As we launch safer and more capable models, we regularly retire older models. Software relying on OpenAI models may need occasional updates to keep working," with a minimum of six months' notice for generally available models [20]. Anthropic's equivalent commitment is at least 60 days [21]. A prompt that produced a correct covenant calculation in the first quarter may not in the third, and in a prompt-based workflow no one is assigned to find out.

4.3 Long documents, lost details

Lender reporting is long. Compliance certificates, financial statement packages and credit agreements run to dozens or hundreds of pages, and the detail that matters, such as a changed reserve definition or a one-time add-back, is rarely on the first page. Liu and colleagues showed that model performance “is often highest when relevant information occurs at the beginning or end of the input context, and significantly degrades when models must access relevant information in the middle of long contexts, even for explicitly long-context models” [22]. In the worst cases they tested, a model given 20 to 30 documents performed worse than the same model given no documents at all [22]. Pasting an entire reporting package into a chat window places the material in exactly the position where recall is weakest.

4.4 Confident errors

A chat interface delivers an answer with the same fluency whether or not the underlying figure was in the document. Researchers at OpenAI argue that “language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty” [23]. Acting Comptroller Hsu put the supervisory version plainly: “Data sets can be biased, algorithms can hallucinate, and reinforcement learning from human feedback can yield mistakes” [24]. The risk is not that the model is often wrong; it is that there is no signal in the output distinguishing the times it is wrong from the times it is right.

4.5 Knowledge that lives in chat histories

In a prompt-based regime, the institution’s process is a collection of personal habits. The analyst who has refined a prompt for borrowing base review over six months holds that refinement in a chat history. MIT’s interviews captured the problem from the user’s side; one corporate lawyer said of a general-purpose assistant: “It’s excellent for brainstorming and first drafts, but it doesn’t retain knowledge of client preferences or learn from previous edits. It repeats the same mistakes and requires extensive context input for each session. For high-stakes work, I need a system that accumulates knowledge and improves over time” [7]. The question a credit committee should ask is the one from the original panel outline: who owns the prompt library, and what happens when that person leaves?

SECTION 05

What Embedding Changes, and What It Does Not

5.1 Retained context and learning

The most direct evidence on why embedding matters comes from the same MIT study that documented the failure of generic tools.

The core barrier to scaling is not infrastructure, regulation, or talent. It is learning. Most GenAI systems do not retain feedback, adapt to context, or improve over time.

MIT NANDA, 2025 [7]

MIT identifies “the learning gap, tools that don’t learn, integrate poorly, or match workflows” as the primary factor separating organizations that see value from those that do not, and reports that “external partnerships with learning-capable, customized tools reached deployment ~67% of the time, compared to ~33% for internally built tools” [7]. The customer-support study points the same way from the other direction: the embedded assistant’s gains came from encoding what the best agents did, which the authors attribute to the fact that “generative AI systems work by capturing and disseminating the patterns of behavior that characterize the most productive agents” [4]. An embedded system turns one analyst’s correction into everyone’s default; a chat session turns it into one analyst’s memory.

5.2 Workflow redesign is where value concentrates

The enterprise surveys agree on where returns come from. McKinsey: “out of 25 attributes tested for organizations of all sizes, the redesign of workflows has the biggest effect on an organization’s ability to see EBIT impact from its use of gen AI” [8]. BCG: “big value comes not from AI pilots or isolated use cases, but from reshaping and reinventing core business workflows end-to-end” [11], and in its diagnosis of why programs fail, “about 70% of the challenges relate to people and process, about 20% are technology issues, and only 10% involve AI algorithms” [10]. Embedding forces the redesign, because the workflow has to be specified before the system can be built. Prompting lets the existing workflow persist with an assistant alongside it.

5.3 Bounding the model

The strongest argument for embedding is not that the model is better. It is that the model is asked to do less. In a well-designed embedded system the model’s role is narrowed to reading and

mapping: identifying that a figure on page 23 is the eligible receivables balance and placing it in the field that holds eligible receivables. Everything downstream, including eligibility rules, concentration limits, advance rates and reconciliation to the prior period, is deterministic logic that the institution can read, test and version. Exceptions are surfaced as they arise, with the source passage attached, rather than being folded into a single answer at the end.

This bears on a question the original panel outline raised: whether the interaction model, a single answer versus questions and exceptions surfaced along the way, is a real difference or a preference. The evidence in Section 4 suggests it is a difference. It changes what the reviewer sees (exceptions with provenance, rather than a fluent summary), what is recorded (every extraction, rule result and override, rather than a conversation), and what can be tested (the system's output on last year's files, rather than an individual's recollection of how well it went).

The design also aligns with how supervisors now draw lines. The revised interagency model risk guidance issued in April 2026 states that the term "model" "excludes simple arithmetic calculations, such as those found within spreadsheets, as well as deterministic rule-based processes and software where there are no statistical, economic, or financial theories underpinning their design or use" [25]. NIST's framework holds that "validity and reliability for deployed AI systems are often assessed by ongoing testing or monitoring that confirms a system is performing as intended" [26]. A versioned system that can be run against historical files is testable in that sense; a chat session is not.

5.4 The honest caveats

A neutral reading of the evidence requires acknowledging where embedded systems fail, and they do.

- **Enterprise tools are rejected too.** MIT found that 60 percent of organizations evaluated enterprise-grade or task-specific tools, 20 percent reached pilot and 5 percent reached production; "most fail due to brittle workflows, lack of contextual learning, and misalignment with day-to-day operations" [7]. An embedded system that does not match how the desk actually works is worse than a chatbot, because it cannot be worked around.
- **The category is over-hyped.** Gartner predicts that "over 40% of agentic AI projects will be canceled by the end of 2027, due to escalating costs, unclear business value or inadequate risk controls" and estimates that "only about 130 of the thousands of agentic AI vendors are real" [1].
- **Dependence shifts to the vendor.** Treasury's 2024 review found that "financial firms face challenges because they may not receive from model vendors and developers access to the

type of information needed to assess risks and develop controls,” and that firms “will need to rely on AI models and systems developed by others” [27]. The repeatability of an embedded system is only as good as the vendor’s change control, and the institution has to be able to see it.

- **Over-reliance does not go away.** The recruiter experiment’s finding that better AI reduces human attention [6] applies to a clean exception queue as much as to a chat window. An embedded system should make the reviewer’s job narrower, not optional.
- **Validation is still hard.** Treasury’s 2024 cybersecurity review recorded that financial institutions found generative AI “very difficult to validate for high-assurance applications” [28]. Embedding reduces the surface that needs validating; it does not eliminate it.
- **Narrowness cuts both ways.** The property that makes an embedded system repeatable, a fixed task definition, makes it useless for the novel question. The two modes are complements, not substitutes.

SECTION 06

The Lending Case: Portfolio Monitoring as a Test

6.1 Why monitoring is the hard case

Asset-based and private credit lending is large and reporting-intensive. Lenders participating in the Secured Finance Network’s survey reported \$310 billion of committed asset-based lines at year-end 2025 [29]; the Secured Finance Foundation’s market sizing puts total asset-based commitments at \$537 billion as of year-end 2024 [30]. Private credit as commonly measured “totaled nearly \$2 trillion by the end of 2023, roughly ten times larger than it did in 2009” [31], and reaches \$3.5 trillion under the broader definition used by the Alternative Credit Council [32].

Every facility in that market reports every period: borrowing base certificates, receivables agings, inventory reports, covenant compliance certificates and financial statements, across hundreds of borrowers per lender. The task differs from drafting a credit memo in three ways. The output must be consistent, so that the same ineligible treatment is applied to the same borrower each month. It must be auditable, so that an examiner, auditor or borrower can trace an availability figure back to its source. And it must be defensible, because advances are made and covenant breaches are called on the basis of it. A memo is read once by a human who judges it; a borrowing base calculation is relied upon.

6.2 The incumbent: spreadsheets

The relevant comparison for any new approach is not a perfect process. It is the one in place, which at most lenders is a spreadsheet fed by email. The research on spreadsheet reliability is long-standing and consistent. Panko's 2015 review of field audits of operational spreadsheets found that "collectively, these studies found errors in 94% of the spreadsheets studied" [33]. Powell, Baker and Lawson audited 50 operational workbooks, found errors in 47 of them, and measured "an average cell error rate over all 270,722 formulas audited" of 1.79 percent [34]. The most consequential illustration in finance remains JPMorgan's 2012 trading loss, where the task force found that the risk model "operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another," and that one formula "divided by their sum instead of their average," which "likely had the effect of muting volatility by a factor of two" [35].

A prompt-based workflow typically feeds this process rather than replacing it: the model's answer is pasted into a cell. It adds a variable step in front of an error-prone one. An embedded system replaces both the extraction and the spreadsheet with versioned logic and a log.

6.3 The supervisory frame in 2026

For fifteen years the baseline for model governance at U.S. banks was the 2011 interagency guidance known as SR 11-7, which required that documentation be "sufficiently detailed so that parties unfamiliar with a model can understand how the model operates, its limitations, and its key assumptions," and that implementing code "cannot be altered except by approved parties, and that all changes are logged and can be audited" [36]. On April 17, 2026, the Federal Reserve, OCC and FDIC replaced it with revised guidance that addresses generative AI directly.

Generative AI and agentic AI models are novel and rapidly evolving. As such, they are not within the scope of this guidance. Nonetheless, a banking organization's risk management and governance practices should guide the determination of appropriate governance and controls for any tools, processes, or systems not covered in this document.

Interagency Guidance on Model Risk Management (revised), April 2026 [25]

The revised guidance is stated to be most relevant to banking organizations above \$30 billion in assets and is not enforceable, and the agencies have announced a forthcoming request for information on banks' use of AI, including generative and agentic AI [25]. Non-bank private credit lenders are not bound by it at all. But it records the supervisors' current view, and that view is

consistent with Treasury’s 2024 findings that some banks hesitate to use language models where “inaccurate, inconsistent, or incomplete answers could lead to liability and reputational damage” [27], and that “if a higher level of explainability is appropriate for a use case, Generative AI may not currently be a viable option” [28].

Read together, the supervisory position is not that lenders should avoid generative AI. It is that whatever is used must be governed with the institution’s own controls, and that stochastic tools are hard to reconcile with replicability. An embedded design that confines the model to extraction and mapping, applies deterministic rules, logs each step and is versioned can meet that bar. An analyst’s chat session cannot, and the limited partners, auditors and bank counterparties of a non-bank lender will ask the same questions an examiner would.

6.4 The five questions, revisited

The panel outline from which this note grew posed five questions about where prompt-based workflows reach their limit. Table 2 sets out how each mode answers them, on the evidence above.

TABLE 2 Panel questions and the two modes

QUESTION	PROMPT-BASED	EMBEDDED
Hand one borrowing base certificate to each. Where do they diverge?	Reads and summarizes; computes if asked; result depends on phrasing; no memory of last period’s certificate.	Maps the certificate to the facility’s schema, applies the eligibility rules in the credit agreement, compares to the prior period and flags variances.
Two people ask the same model the same question. Different answers?	Expected (Section 4.1). The difference is invisible unless someone compares.	Same inputs and rules produce the same output. Any difference is a logged override with a name on it.
Who owns the prompt library?	An individual. It leaves when they do.	Configuration owned by the institution, versioned and documented.
A provider ships an update. Does the workflow change?	Yes, silently (Section 4.2). No one is assigned to notice.	The vendor tests and releases under change control; the institution can see what changed.
Single answer at the end, or exceptions along the way?	A single answer, with no indication of which parts were uncertain.	Exceptions surfaced with the source passage; the reviewer’s job is narrower and the record is complete.

6.5 The borrower on the other side

There is no published evidence yet on how borrowers experience automated review of their reporting, and this note does not pretend to supply it. The questions are real: whether borrowers

know a system is reviewing their submissions, whether faster turnaround produces fewer surprises or more scrutiny, and whether an intake process that rejects a malformed submission immediately pushes quality upstream, as structured portals have in other domains. The last is the most promising and the least tested. It is a reasonable hypothesis that embedded review changes borrower behavior; it is, for now, only that.

SECTION 07

A Decision Framework for Institutions

The evidence does not support a blanket preference for either mode. It supports matching the mode to the task.

TABLE 3 Matching the mode to the task

PROMPT-BASED IS THE RIGHT TOOL WHEN	EMBEDDED IS THE RIGHT TOOL WHEN
The task is novel or one-off	The task recurs every period
The output is judged by a human reader	The output is relied upon for a decision
It does not need to match a prior result	It must match the prior period and other analysts
The user is expert enough to catch errors (inside the frontier)	Volume exceeds what careful human review can cover
Nothing needs to be traced afterward	An auditor, examiner or borrower may ask how the figure was reached
The data is non-sensitive or the tool is an approved enterprise tier	The data is borrower financials under confidentiality obligations

7.1 Build, buy or wait

Institutions that decide a task belongs in an embedded system face a second choice. The evidence favors buying over building for most of them. MIT found vendor partnerships reached deployment at roughly twice the rate of internal builds [7], and BCG’s finding that 70 percent of the difficulty is people and process rather than algorithms [10] identifies precisely the cost that internal teams building on a general model tend to underestimate: the workflow design, the exception handling and the maintenance as models change. Treasury’s observation that firms “will need to rely on AI models and systems developed by others, perhaps with some tailoring” [27] describes the realistic position of all but the largest lenders.

Waiting is also a choice, and the shadow-AI figures in Section 3.3 show what it means in practice: the prompt-based mode is already in use on the desk whether or not the institution has chosen it. Michael Barr, then the Federal Reserve's Vice Chair for Supervision, put the obligation this way: "Every financial institution should recognize the limitations of the technology, explore where and when GenAI belongs in any process, and identify how humans can be best positioned to be in the loop" [37].

7.2 Questions to ask before buying

The panel outline's final section asked what lenders who had been through a selection process wished they had asked. The evidence suggests a short list.

- Run it on your own historical files, not the vendor's demonstration set, and compare the output to what your analysts concluded at the time.
- Ask what the model does and what deterministic rules do, and insist on seeing the rules.
- Ask how changes are versioned, tested and communicated, and whether a model-provider update can reach production without the vendor's release process.
- Ask what is logged: inputs, extractions, rule results, exceptions and overrides, and who can see the log.
- Ask what happens when the system is unsure, and whether that uncertainty is visible to the reviewer.
- Involve IT, information security, model risk and the CFO before the pilot, not after; each will have a requirement the credit team did not anticipate.

SECTION 08

Conclusion

The debate about AI in lending is often framed as a question of which model is best. The evidence reviewed here suggests that for institutional work the model is the least important variable. General-purpose assistants improve individual output on bounded tasks and will continue to; they are the right tool for drafting, exploration and the one-off question. But their variability across users, phrasings and versions, their weakness on long documents, their confident errors and their absence of a durable record are structural, and no amount of prompt discipline changes that. For work that recurs every period, across hundreds of borrowers, and has to come out the same way and be explained afterward, the evidence favors embedding the model inside a system that bounds what it does, tests what it produces and records what happened.

Embedded systems bring their own failure modes, and institutions should hold vendors to the same standard of evidence this note has tried to apply. But the direction is clear enough. As Acting Comptroller Hsu observed, “trust not only sits at the heart of banking, it is likely the limiting factor to AI adoption and use more generally” [24]. Trust, in a lending operation, is built on repeatability. That is what embedding is for.

ABOUT THE AUTHOR AND DISCLOSURE

Alex Parisi is a co-founder of CORA (Collateral Oversight and Risk Analytics), which builds software that embeds document extraction, borrowing base and covenant logic in a lender’s portfolio monitoring workflow. The author therefore has a commercial interest in the conclusions of this note. All sources cited are publicly available and are listed in full below so that readers can weigh the evidence independently. This note is provided for general information and is not legal, regulatory or investment advice.

SOURCES

References

All sources were accessed and verified in October 2026. Quotations are reproduced verbatim from the primary document, except where bracketed letters mark a change of case at the start of a quotation. Where a report has been issued in more than one version, the version quoted is identified.

- [1] Gartner, Inc. "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027." Press release, June 25, 2025. <https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>
- [2] Noy, Shakked, and Whitney Zhang. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science* 381, no. 6654 (July 2023): 187–192. <https://doi.org/10.1126/science.adh2586>
- [3] Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Kraye, François Candelon, and Karim R. Lakhani. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013, September 2023. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4573321
- [4] Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. "Generative AI at Work." NBER Working Paper 31161, April 2023 (revised November 2023); published in *Quarterly Journal of Economics* 140, no. 2 (2025): 889–942. Percentages quoted are from the NBER revised version. <https://www.nber.org/papers/w31161>
- [5] Becker, Joel, Nate Rush, Elizabeth Barnes, and David Rein. "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." METR, July 2025. arXiv:2507.09089. <https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/>
- [6] Dell'Acqua, Fabrizio. "Falling Asleep at the Wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters." Working paper, Harvard Business School, 2022. <https://aiinstitute.hbs.edu/is-ai-making-your-team-lazy/>
- [7] Challapally, Aditya, Chris Pease, Ramesh Raskar, and Pradyumna Chari. The GenAI Divide: State of AI in Business 2025. MIT NANDA, July 2025 (preliminary release). https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf
- [8] Singla, Alex, Alexander Sukharevsky, Lareina Yee, Michael Chui, and Bryce Hall. The State of AI: How Organizations Are Rewiring to Capture Value. McKinsey & Company / QuantumBlack, March 2025. https://www.mckinsey.com/~media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai/2025/the-state-of-ai-how-organizations-are-rewiring-to-capture-value_final.pdf
- [9] Singla, Alex, Alexander Sukharevsky, Bryce Hall, Lareina Yee, Michael Chui, and Tara Balakrishnan. The State of AI in 2025: Agents, Innovation, and Transformation. McKinsey & Company / QuantumBlack, November 2025. <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-state-of-ai-2025>
- [10] de Bellefonds, Nicolas, et al. Where's the Value in AI? Boston Consulting Group, October 2024. <https://media-publications.bcg.com/BCG-Wheres-the-Value-in-AI.pdf>
- [11] Apotheker, Jessica, et al. The Widening AI Value Gap: Build for the Future 2025. Boston Consulting Group, September 2025. <https://media-publications.bcg.com/The-Widening-AI-Value-Gap-October-2025.pdf>
- [12] Sniderman, Brenna, Jim Rowan, Costi Perricos, Beena Ammanath, and David Jarvis. Now Decides Next: Generating a New Future. Deloitte's State of Generative AI in the Enterprise, Quarter Four Report. Deloitte AI Institute, January 2025. <https://www.deloitte.com/content/dam/assets-zone3/us/en/docs/services/consulting/2024/us-state-of-gen-ai-q4.pdf>
- [13] Microsoft and LinkedIn. 2024 Work Trend Index Annual Report: AI at Work Is Here. Now Comes the Hard Part. May 2024. <https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part>
- [14] IBM. "IBM Report: 13% of Organizations Reported Breaches of AI Models or Applications, 97% of Which Reported Lacking Proper AI Access Controls." Press release accompanying Cost of a Data Breach Report 2025, July 30,

2025. <https://newsroom.ibm.com/2025-07-30-ibm-report-13-of-organizations-reported-breaches-of-ai-models-or-applications,-97-of-which-reported-lacking-proper-ai-access-controls>
- [15] Barr, Michael S. "AI, Fintechs, and Banks." Speech at the Federal Reserve Bank of San Francisco, April 4, 2025. <https://www.federalreserve.gov/newsevents/speech/barr20250404a.htm>
- [16] He, Horace, et al. "Defeating Nondeterminism in LLM Inference." Thinking Machines Lab, September 10, 2025. <https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/>
- [17] Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting." International Conference on Learning Representations (ICLR), 2024. arXiv:2310.11324. <https://arxiv.org/abs/2310.11324>
- [18] Errica, Federico, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. "What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering." Proceedings of NAACL 2025, 1543-1558. <https://aclanthology.org/2025.naacl-long.73/>
- [19] Chen, Lingjiao, Matei Zaharia, and James Zou. "How Is ChatGPT's Behavior Changing Over Time?" Harvard Data Science Review 6, no. 2 (2024). arXiv:2307.09009. <https://doi.org/10.1162/99608f92.5317da47>
- [20] OpenAI. "Deprecations." OpenAI API documentation. Accessed October 6, 2026. <https://developers.openai.com/api/docs/deprecations>
- [21] Anthropic. "Model deprecations." Claude Developer Platform documentation. Accessed October 6, 2026. <https://platform.claude.com/docs/en/about-claude/model-deprecations>
- [22] Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranajpe, Michele Bevilacqua, Fabio Petroni, and Percy Liang. "Lost in the Middle: How Language Models Use Long Contexts." Transactions of the Association for Computational Linguistics 12 (2024): 157-173. https://doi.org/10.1162/tacL_a_00638
- [23] Kalai, Adam Tauman, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. "Why Language Models Hallucinate." arXiv:2509.04664, September 2025. <https://arxiv.org/abs/2509.04664>
- [24] Hsu, Michael J. "AI Tools, Weapons, and Accountability: A Financial Stability Perspective." Remarks at the 2024 Conference on Artificial Intelligence and Financial Stability, June 6, 2024. OCC News Release 2024-61. <https://www.occ.gov/news-issuances/speeches/2024/pub-speech-2024-61.pdf>
- [25] Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation, and Office of the Comptroller of the Currency. Supervisory Guidance on Model Risk Management (revised). Federal Reserve SR Letter 26-2, April 17, 2026 (superseding SR 11-7); OCC News Release 2026-29. <https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm>
- [26] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, January 2023. <https://doi.org/10.6028/NIST.AI.100-1>
- [27] U.S. Department of the Treasury. Artificial Intelligence in Financial Services: Report on the Uses, Opportunities, and Risks of Artificial Intelligence in the Financial Services Sector. December 2024. <https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf>
- [28] U.S. Department of the Treasury. Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector. March 2024. <https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf>
- [29] Secured Finance Network. Asset-Based Lending Survey, Q4 and Annual 2025. Figures reflect survey participants, both bank and non-bank lenders. <https://www.sfnetwork.com/docs/default-source/data-files-and-research-documents/sfnetwork-quarterly-non-participant-report-4q-annual-2025.pdf>
- [30] Secured Finance Foundation and Keybridge Research. 2025 Secured Finance Market Sizing Study. Released February 10, 2026 (data as of Q4 2024). <https://www.businesswire.com/news/home/20260210873163/en/>
- [31] McKinsey & Company. "The Next Era of Private Credit." September 24, 2024. <https://www.mckinsey.com/industries/private-capital/our-insights/the-next-era-of-private-credit>

- [32] Alternative Investment Management Association and Alternative Credit Council. Financing the Economy 2025. Press release, December 9, 2025. <https://www.aima.org/article/press-release-strong-growth-sees-private-credit-market-reach-us-3-5-trillion.html>
- [33] Panko, Raymond R. "What We Don't Know About Spreadsheet Errors Today: The Facts, Why We Don't Believe Them, and What We Need to Do." Proceedings of the European Spreadsheet Risks Interest Group (EuSpRIG), 2015. arXiv:1602.02601. <https://arxiv.org/abs/1602.02601>
- [34] Powell, Stephen G., Kenneth R. Baker, and Barry Lawson. "Errors in Operational Spreadsheets." Journal of Organizational and End User Computing 21, no. 3 (2009): 24-36. https://mba.tuck.dartmouth.edu/spreadsheet/product_pubs_files/Errors.pdf
- [35] JPMorgan Chase & Co. Report of JPMorgan Chase & Co. Management Task Force Regarding 2012 CIO Losses. January 16, 2013, Appendix A. Quotations as reproduced in contemporaneous published commentary on the report. <https://elischolar.library.yale.edu/ypfs-documents/454>
- [36] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory Guidance on Model Risk Management. Federal Reserve SR Letter 11-7 / OCC Bulletin 2011-12, April 4, 2011 (superseded April 2026). <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>
- [37] Barr, Michael S. (Vice Chair for Supervision). "Artificial Intelligence: Hypothetical Scenarios for the Future." Speech at the Council on Foreign Relations, New York, February 18, 2025. <https://www.federalreserve.gov/newsevents/speech/barr20250218a.htm>



COLLATERAL OVERSIGHT & RISK ANALYTICS

coraabl.com