

Why Most Back-Office AI Programs Stall Before They Scale

Almost every large bank now runs an AI program in its operations, usually with public targets attached: fewer manual hours, faster loan decisions, lower cost per case. Far fewer can point to results at scale, and the wider data suggests why.

In MIT's 2025 study of enterprise generative AI, 95% of the organizations in the research saw no measurable return from their initiatives. The researchers traced this to how the tools were deployed rather than to the underlying AI models: generic tools that never adapted to the workflows they were placed in [1]. Gartner, meanwhile, expects more than 40% of agentic AI projects to be cancelled by the end of 2027, as costs escalate, business value stays unclear or risk controls fall short [2].

For a bank's back office, a stalled program is expensive in a very particular way. The pilot budget is gone, but the operations team is the same size, and the review queues are as long as before, so the next funding request starts from a weaker position. The irony is that operations are exactly where MIT found the strongest returns: back-office automation paid off more than the front-office tools that attracted most of the budget [1]. And the programs that do reach production often stop after the first or second use case.

Successful initiatives have been observed to pass through three levels:

1. Making it into production.
2. Generating real EBIT impact.
3. Scaling across the organization.

What gets such automation programs to level 3?

Reaching production is hard enough. But once it is reached, the real test begins, and the two levels above it are governed by different rules.

Level 2: Interestingly, the value of GenAI solutions is close to binary. It is either there or it is not. A solution can be fully scaled into production and still deliver close to zero impact: while automation performance sits below a critical threshold, every case still passes through a human and the cost per case barely moves. Above that threshold the return climbs exponentially. Crossing it is not a matter of one metric: it demands leading performance simultaneously in extraction and validation accuracy, functionality, processing cost, processing and validation speed, delivery method and

timeline, change management and human adoption. Only composite AI is built to meet all of these at once: deterministic wherever meaningful, LLMs for what conventional software cannot automate. (explored in Chapter 1)

Level 3: the second and third use cases decide the program. Having earned a return on the first, most banks risk stopping there. Scaling depends entirely on the ease, time and cost of adding the next one, and on using the solution with confidence across all FTE-heavy, critical back-office workflows. If each new process means restarting development, the program quietly ends after use case one or two. Reaching level 3 means automating the entire lifecycle of setting up operations transformation, so the bank has its agentic automation engine ready within days for any workflow, fully configured, tested and reported, without compromising the regulatory requirement for an explainable and auditable approach. (explored in Chapter 2)

The perspective in this article comes from the successful deployment of OpsAI at financial institutions across Europe. OpsAI automates operations tasks end-to-end, from input processing (collecting and understanding everything a case needs, such as unstructured documents), through operations intelligence (assessing the case against the bank's own policies or expert analysis), to action taking and generating output (writing a message or triggering the next workflow), while the bank's experts keep ownership of the decisions.

Chapter 1

Value stays near zero until a threshold is crossed

In banking and insurance operations, value does not grow in proportion to performance. A solution at 80% accuracy or 60% coverage does not deliver 80% or 60% of the value, because the operations team still must touch every case. If nobody knows which fifth of the cases is wrong, all of them get checked. The team stays the same size, trained on the full rule set and staffed for peak volume, and customers still wait for the slowest cases in the queue. Value stays close to zero until the system is good enough that people can stop reviewing every case, and then it rises quickly.

The same threshold applies well beyond accuracy. Take functional coverage: if the system has major gaps in its capabilities while processing a mortgage application or a corporate account opening, operations colleagues still need to touch every single case, just as they did before, and the cost barely moves.

Processing cost per case, meaning what the bank pays for LLM and OCR usage, has a threshold too. A workflow that is affordable in a pilot of a few hundred files can become uneconomic at full volume, which is often why a scaled business case can only work with highly optimized token usage. Speed has one as well, since a system that needs many minutes, or even hours, for a comprehensive file cannot support the same-day decisions the business case was built on. Clearing the bar on one of these while missing another leaves the bank in much the same place as missing all of them. Across OpsAI deployments, the business case only holds once every dimension reaches roughly 95% or better.

The accuracy bar is higher than it sounds. A typical mortgage application is checked against more than 2,000 rules before it can be approved: income continuity, the required documents for the applicant's employment type, the valuation against the purchase price, the address across 4 documents filled in by 4 different people, among many other checks. A corporate credit file is no simpler: financial statements have to be spread and analyzed, ownership traced to the ultimate beneficial owner, and signatory rights checked against the company registry.

If each of the mortgage file's rule checks were independently right 99% of the time, the file would still carry about 20 errors on average ($2,000 \times 1\% = 20$), and the chance of it coming through with no error at all would be about 2 in a billion ($0.99^{2,000} \approx 0.000000002$). That is why accuracy has to be measured per case, and why a figure that sounds close to perfect can still leave every file in the review queue.

Why neither technology gets there alone

Instead of following the hype path of agent builders or fully GenAI-based solutions, where almost every step is powered by a language model, OpsAI combines the strengths of traditional software with what LLMs do best. It uses LLMs only where conventional software falls short, so every task is handled by the technology best suited to it. To see why, it helps to look at where each approach falls short against the threshold.

Language models struggle with accuracy on chained work. FinSheet-Bench, a March 2026 benchmark from Qubera and the University of Zurich, tested ten model configurations from OpenAI, Google and Anthropic on financial spreadsheets modeled on private equity fund data. The best, Gemini 3.1 Pro, reached 82.4% overall, roughly one wrong answer in six, and the three strongest models fell from about 94% on simple lookups to about 33% on complex multi-step calculations. The authors concluded that no standalone model was accurate enough for unsupervised use in professional finance, and that reliable results will likely need architectures that separate document

understanding from deterministic computation [3]. The Bank for International Settlements makes a related point about consistency: unlike traditional AI models, where the same input always gives the same result, generative AI can give different results from the same input [4]. LLM-only agents also suffer from compounding error: a mistake made early in the process, such as a misread income figure, is carried through the rest of the case and becomes the wrong basis for every step that follows.

They also fall short on cost. Token prices have fallen, but Gartner's modeling, reported in August 2026, suggests that an AI agent with reasoning capabilities needs on average 5 to 30 times more tokens to solve a task than a standard chatbot exchange, the simplest way of using a language model, and that a task costing one cent in a chatbot could cost an agent up to \$1.50 [5]. The more of a workflow is handed to a reasoning agent, the faster that multiplier adds up. Every step sends the accumulated context back to the model, and when a case changes halfway through (a co-debtor joins the application, the applicant switches property, a supplementary damage report arrives) the reasoning chain has to run again, on a longer context.

And they are slow at runtime. An agent working through a task makes tens to hundreds of chained model calls, each waiting on the one before, and research on agent inference finds that the way those calls are scheduled inflates end-to-end latency by three to eight times [6]. Accuracy and speed also pull against each other: in the FinSheet-Bench tests, the three most accurate configurations took 16 to 28 seconds on average to answer a single question, while the quickest answered in under three seconds but got no more than 58% of answers right [3]. A coded rule check, by contrast, runs in milliseconds, and a loan file needs more than 2,000 of them.

Rules-based software has the opposite profile. It applies a rule the same way every time, runs in milliseconds, and re-running a check costs the same as running it the first time. The problem is getting it built. Every rule has to be spelled out in full, and property area is a good example of what that means: on a valuation report it may differ from the title deed within a tolerance, because different people measure it in different ways, while between the title deed and the purchase contract it has to match exactly. Even comparing two addresses means splitting each into eight fields, normalizing them, matching them and scoring the differences. Multiply that by 2,000 rules and a deterministic build takes months, and every later change request queues behind the next release. On top of that, rules-based software cannot read what it was not built to read. A purchase contract has no fixed layout and a damage report contains whatever the claimant chose to write, so every new document format turns into another development project.

So the two technologies are not competing for the same job. Deterministic software is at its best wherever an answer has to be exact, repeatable and explainable: rules,

calculations, validations and system updates. Generative AI is at its best wherever the input is unstructured or the task calls for interpretation: reading documents, understanding free text, handling the unexpected and generating output such as documents, texts and messages, for example a credit memorandum or a credit risk rationale draft. A single mortgage or corporate credit file needs both inside one audited workflow, which is the idea behind composite AI. In a regulated banking environment, auditability is critical. A black-box agent that receives a task and works out a new way to solve it every time cannot be audited. A workflow built on deterministic logic, using traditional and agentic components side by side, can be audited at every step.

How composite AI clears the threshold

The team behind OpsAI reached this conclusion in 2023, before it had built a product, when its first client project showed that probabilistic output alone could not be shipped to a regulated client. OpsAI was built on that constraint. The team has delivered automation solutions for clients including Erste BCR, Raiffeisen, UNIQA and Allianz, and OpsAI now automates complex banking and insurance workflows, such as mortgages, claims, SME and corporate lending, KYC and many more. By the end of 2026, more than 840,000 annualized banking and insurance cases will run through it across Europe.

Composite AI is what lets OpsAI clear the threshold on every dimension at once rather than just one: 98%+ accuracy, coverage of the edge cases, under 0.01 USD per page and seconds per document.

This is also why OpsAI does not follow the operating principle of a typical agent platform. A typical agent gets a task, decides for itself how to do it, calls tools and then verifies the answers, which makes its output non-deterministic, slower, more expensive and prone to compounding errors. OpsAI works as an agentic digital co-worker instead: it completes every task along a deterministic, auditable workflow, uses AI only in the steps that require it, and judges every case against the same criteria, which keeps it faster, cheaper and accurate at 98 to 99% and above.

The way it reaches the threshold is easiest to see in how accuracy is measured and secured. OpsAI measures accuracy at two levels: for every extracted field, and for every case as a whole, where a case counts as correct only if its outcome matches what a competent operator would have produced from the same inputs. Several mechanisms keep both numbers high. Each field is extracted with multiple LLM calls that have to agree, every value carries its own confidence score, and anything below the threshold is routed to a reviewer automatically. The deterministic decision engine then checks every value for completeness, consistency and validity across the whole file, so a misread figure is caught by the rules before it reaches the operations team, and every value

stays traceable to its exact location in the source document. Because each workflow is also tested against ground truth before it goes live, high accuracy is something OpsAI can guarantee in the contract rather than only report afterward. Accuracy is measured in production, per workflow, on a sample agreed with the client, and measured again every time the configuration changes.

People stay in the loop and keep the decisions. The system owns the repetitive work end-to-end: reading the file, validating it against the bank's rules and preparing the outcome. Judgment calls such as the credit decision itself stay with people, by design and by regulation. Reviewers work through an interface that shows every extracted value alongside its exact place in the source document and the rule that judged it, and the bank sets how deep that review goes for each document type and product. What remains for people is judgment, not leftover data entry, and their review closes the last gap between very high accuracy and a fully correct outcome.

Getting there took serious investment. By the OpsAI team's own count, clients only saw the threshold crossed after around 8,000 man-days of product development, roughly a 35-person team working for a full year, and total investment has since passed 21,000 man-days over two and a half years. That scale of effort is one reason buying tends to beat building: in MIT's study, tools bought from specialized vendors and implemented through partnerships reached deployment about 67% of the time, while internally built tools succeeded far less often [1].

Chapter 2

Composite AI works. Scaling it is the hard part

A composite AI system can take a first workflow past the threshold. The harder question is what happens with the second, the fifth and the twentieth.

The baseline expectation in the market is that an automation platform works not just in one operations area but in all of them, and scales to each new one easily, quickly and economically. A bank that starts with retail lending expects to move on to SME and corporate lending, estate settlement, KYC and other areas without starting over. Scaling also means going deeper within one area. The rules that apply to a mortgage case depend on several things at once: the product, the applicant's employment type, their source of income and the type of collateral, and each of these multiplies the others. A salaried applicant buying an apartment on one income needs one set of documents and one rule path, while a self-employed applicant with rental income buying a plot of land needs a very different one. Four or five such dimensions produce thousands of valid configurations, each of which has to be defined somewhere, and

when a rule changes or a regulator moves a threshold, the change has to be traced through every configuration it affects and validated again.

Built by hand, that turns into a backlog that keeps growing. A composite system that can only be changed by writing new code inherits the slow build and change cycle of traditional software, spread across far more logic. Each new process tends to mean a fresh development effort, and over time the bank collects a set of isolated automations, each with its own maintenance cost, instead of one platform that grows.

OpsAI has already helped its clients lead the market in operations automation. The experience of the past year showed what it takes to go further: the work of setting up each new workflow, from configuring it to testing it and reporting on it, has to be automated as well.

This is what OpsAI 2.0, launching in Q4 2026, is built for. It makes OpsAI an agentic AI platform for document-intensive operations at regulated financial institutions that automatically builds the operations automation engine for the bank, with structural upgrades covering configuration, testing, monitoring and human review.

Configuration instead of development

When it comes to scaling, the client configurator is one of the most important advancements, as it decides how quickly a bank can change a live workflow.

In regulated lending, changing an automated workflow has usually forced a choice between moving fast and staying in control. Moving fast meant a vendor release and new logic that had to be validated again. Staying in control meant moving slowly. OpsAI ran into this in its own deployments, and its answer was to stop treating every change as an engineering project.

On its configuration platform, the bank's own business teams define document types and build or change workflows without writing code. For each document type they set the fields, extraction prompts and review display from sample documents. A workflow then combines those document types with the case's document requirements and its completeness, consistency and validity rules. These are built in a visual rule builder that also covers calculated fields and per-item checks over lists. Teams start from pre-filled templates rather than a blank page, and a document type added once carries across every workflow that needs it. OpsAI drafts the configuration. In some places it sets up a whole section; in others it helps with a single item, such as one rule or formula. Either way, the team reviews and corrects the result. A new document type therefore takes hours and a new workflow only a few days, with no engineering build

and no software release. Changes are made on versioned drafts and can be safely tested in-app before they are exposed to live cases. Ready versions go live immediately or on a scheduled date, and live workflows are never upgraded silently.

The working model reflects this: The OpsAI team builds the first workflow together with the bank, and the bank's own teams configure the ones after that. Document types that are already set up can be reused in every new workflow, and an existing workflow can be duplicated as a starting point, so each new workflow starts further ahead than the first.

For a regulated lender, the more important benefit is control. The platform configures; it does not generate code. A configured change runs inside the same audited rule engine as every other decision in the system, so OpsAI works as the bank configured it, with no new logic that nobody has reviewed. Where a step calls an LLM, as in extraction and some evaluations, the result cannot be guaranteed to be identical every time. Mass testing, detailed configuration and refinement make outcomes predictable in most cases. Hand-coded changes build up engineering debt, and model updates can shift outputs without anyone touching the workflow; both make past decisions harder to reproduce. That matters under the EU AI Act, which classifies AI systems used to evaluate the creditworthiness of individuals as high-risk and requires, among other things, automatic logging, technical documentation and human oversight for them [7]. The Digital Omnibus, in force since July 2026, moved the deadline for these obligations from August 2026 to December 2027, postponing them rather than removing them [8]. Being able to show why a past decision was made comes from the architecture, and it only survives change if the changes stay inside it.

Testing before go-live

Before a configuration reaches real cases, the Benchmarking module lets teams mass-test it. They build a set of synthetic documents, either generated or hand-written to mimic real ones with artificial information. Each document carries a ground truth that humans can review and update. One bulk run tests document categorization, data extraction and rule validation, and the expected results can be evaluated against the system's output. Deviations therefore show up in testing, not on real customer cases.

Faster case review

When a reviewer opens a case in GuidedFlow, they start with a to-do list, not a pile of documents. A case snapshot (borrowers, property, loan product, amount, term) sits above status counters showing exactly what blocks the review: documents still processing, missing, failed to categorize, technical errors or invalid files. Once the

pipeline has run, the same screen shows only the rules that were actually broken, as short cards, so the reviewer works from issues rather than files.

The review itself is organized by cluster, not by document. Data extracted from every document is merged per entity (personal data per borrower, property, loan, other), and each field is confirmed once. Where the sources agree, the field arrives already validated and can be filtered out of view. Where they conflict, the candidate values appear side by side with their source documents (for example, place of birth on the ID card vs. the personal data sheet), one click from the original. Instead of opening twenty documents and cross-checking them from memory, the reviewer settles a short list of discrepancies.

How the most advanced buyers test for this

The banks furthest along no longer take a headline accuracy figure at face value. In a recent evaluation, a top 10 European banking group measured OpsAI's platform against its own internally built group solution, using a checklist of more than fifty concrete criteria, and selected OpsAI. The criteria covered whether the system is an extraction tool or a full operations automation platform, whether it handles all client edge cases, whether its decision logic is deterministic and auditable, whether the bank can make changes itself through a configurator, and whether it already runs in production across personal, business and wholesale banking. The group also set hard numbers: at least 97% accuracy, under 15 seconds processing time per document, and under 0.01 USD processing cost per page. In the proof of concept, OpsAI's platform measured 0.0066 USD per page and an average of 11.6 seconds per document.

Very few of those criteria are about the model itself. Most concern coverage, auditability, cost, speed and the ability to change. So when a vendor quotes an accuracy figure, the first question worth asking is what was counted.

The programs that scale are the ones that combine both technologies above the threshold on every dimension, keep people on the decisions that need judgment, and can keep changing the workflow as the bank adds use cases without reopening the audit each time.

The OpsAI team will show what that last part looks like in practice at the 5th Annual Back Office & Operational Excellence in Banking Summit in Prague on October 20 and 21, 2026. On the first day, CEO Máté Jendrolovics will take the stage to demo the OpsAI agents live and unveil OpsAI 2.0 for full back-office banking operations automation, the biggest product update in OpsAI's history.

Sources

[1] Fortune, "MIT report: 95% of generative AI pilots at companies are failing," August 18, 2025, reporting on MIT NANDA, "The GenAI Divide: State of AI in Business 2025."

<https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo>

[2] Gartner, press release "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027," 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>

[3] Ravník et al. (Qubera AG and University of Zurich), "FinSheet-Bench: From Simple Lookups to Complex Reasoning, Where LLMs Break on Financial Spreadsheets," arXiv preprint, March 2026. <https://arxiv.org/pdf/2603.07316>

[4] Bank for International Settlements, Financial Stability Institute, "Regulating AI in the financial sector: recent developments and main challenges," FSI Insights No. 63, December 2024. <https://www.bis.org/fsi/publ/insights63.pdf>

[5] Computer Weekly, "Gartner: Agentic AI won't benefit from economies of scale," August 19, 2026, reporting on Gartner's Inference Paradox research.

<https://www.computerweekly.com/news/366648782/Gartner-Agentic-AI-wont-benefit-from-economies-of-scale>

[6] "SAGA: Workflow-Atomic Scheduling for AI Agent Inference on GPU Clusters," arXiv preprint, May 2026. <https://arxiv.org/html/2605.00528>

[7] Regulation (EU) 2024/1689 (Artificial Intelligence Act), Annex III point 5(b) and Articles 11, 12 and 14, Official Journal of the European Union. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>

[8] Regulation (EU) 2026/1744 (Digital Omnibus on AI), Official Journal of the European Union, July 24, 2026. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>

