findic

WHITE PAPER

Agents in Production in the Financial Industry

Agentic AI in the financial industry: what the published accounts demand, and how an agentic GraphRAG platform meets them



Contents

About this white paper 3
1 · The industry is shifting: what people are talking about 3
2 · Five requirements that keep recurring 4
3 · Where agents deliver impact in the business 5
4 · The architectural answer: a supervisor in front of the knowledge graph 7
5 · What published benchmarks show 9
6 · Lessons from building it 10
7 · Limits and what comes next 11
8 · Sources and further reading 12
9 · Contact 14

About this white paper

The financial industry has left the experimentation phase of agentic AI behind. Banks, asset managers and financial services providers now run agents in daily operations, and the accounts that have become public formulate remarkably uniform requirements: multi-stage agent architectures with explicit planning, end-to-end observability, systematic evaluation, governance in the runtime environment, and the human on the last mile. Hardly any of these requirements is a property of the language model. They are all made achievable only by the platform surrounding the model.

This white paper places where agents have their strongest substantive impact in the financial industry and shows, using a GraphRAG platform we have built and operate, how each of these requirements can be answered architecturally. We deliberately do not name the institutions we work with. Instead, we lay open the building blocks, point to published benchmarks, and report what building the platform has actually taught us in practice.

1 · The industry is shifting: what people are talking about

As recently as two years ago, generative AI in the financial industry was above all a topic for innovation labs. That has changed. Morgan Stanley provides around 16,000 advisers with an assistant that draws on more than 100,000 internal research documents, and has publicly documented its handling of systematic model evaluation.[1] BBVA has expanded access to a company assistant from an initial figure of just over 3,000 to more than 100,000 employees.[2]

The most instructive material at present is the accounts in which institutions describe how they got there. LangChain's financial services report collects three of them.[3] J.P. Morgan built Ask DAVID, a research agent for investment advice, whose answer accuracy initially fell below 50 per cent, rose to around 80 per cent through better prompting and retrieval, and reached the required figure of around 90 per cent only through the orchestration of specialised subtasks. Bridgewater reports an analysis agent whose data retrieval climbs from 50 to 90 per cent accuracy, while the same analysis plan produces identical code in 95 per cent of cases across two independent runs. The agent works with up-to-date data for several hundred investors, each time strictly confined to their data domain. And the financial services provider Chime operates a member agent for 9.5 million customers whose quality is monitored via evaluation dashboards with scores between 87 and 96 per cent.[4]

These accounts do not read like press releases. They contain information about failures, about intermediate states, about the effort behind the final percentage points. That is precisely what makes

[1] OpenAI & Morgan Stanley (2024): Morgan Stanley uses AI evals to shape the future of financial services. openai.com/index/morgan-stanley

[2] BBVA & OpenAI (2026): Expansion of ChatGPT Enterprise to more than 100,000 employees; starting point 2024: 3,300 licences. bbva.com / openai.com, company announcement

[3] LangChain (2026): Financial Services Agents in Production: Lessons building agents from J.P. Morgan, Chime, and Bridgewater. LangChain Inc., San Francisco

[4] LangChain (2026), ibid.

them valuable. They describe at first hand everything an agent must be capable of in a regulated institution before it may go into production.

2 · Five requirements that keep recurring

Reading the accounts, five points recur, regardless of whether the subject is investment advice, wealth management or customer service.[5]

First, multi-agent architectures emerge through iteration, not on the drawing board. None of the institutions reported designed its target architecture in advance. All of them began with a single, well-instrumented agent and decomposed its task only when the measurements demanded it. Second, observability is a precondition for go-live, not a retrofitted add-on. Without complete traces of every single agent step, none of the systems went into production. Third, evaluation has by now differentiated itself methodically with rigor. Leading institutions strictly separate rigorous offline evaluation (regression tests on golden datasets before deployment) from continuous online evaluation at runtime (for example through evaluator agents acting as LLM-as-a-judge). These evaluations became the shared language between business function and engineering. The dashboards on which the business side and development agree bundle traces and metrics in exactly the same instrument. Fourth, governance is moving into the runtime environment. Spending limits, approval points, the masking of personal data and the mapping to threat catalogues such as the OWASP Top 10 for LLM Applications[6] are enforced not in concept papers but in the code of the runtime. Fifth, the human remains in the last mile. The reported accuracy figures of around 90 per cent arise only from the combination of machine and expert final review.

For institutions in the European supervisory sphere there is an additional point: these five requirements are not voluntary excellence. The regulation on digital operational resilience requires traceable control of ICT risks wherever third parties and new technologies are involved,[7] and the European AI Act demands documentation, supervisability and human oversight for high-risk applications.[8] Anyone who takes the five requirements seriously is in effect already building the evidential capability that supervisors and internal audit will later request.

The decisive point: none of these requirements can be solved merely by choosing a better language model. They all concern the platform around the model, such as planning, logging, evaluation, limitation, handover. That is where production systems differ from demonstrators.

---

[5] LangChain (2026), ibid.

[6] OWASP (2025): OWASP Top 10 for Large Language Model Applications. owasp.org/www-project-top-10-for-large-language-model-applications

[7] Regulation (EU) 2022/2554 of the European Parliament and of the Council on digital operational resilience for the financial sector (DORA)

[8] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence

3 · Where agents deliver impact in the business

Agents have the strongest effect where two conditions come together. The work consists of chained decisions over documents and data, and every intermediate decision must be explainable later. Figure 1 compares the publicly reported figures for the market: agent operations and safeguards, the maturity development of the fifty largest banks, and the cancellation forecast for agent projects. The publicly documented production examples concentrate on a few fields.

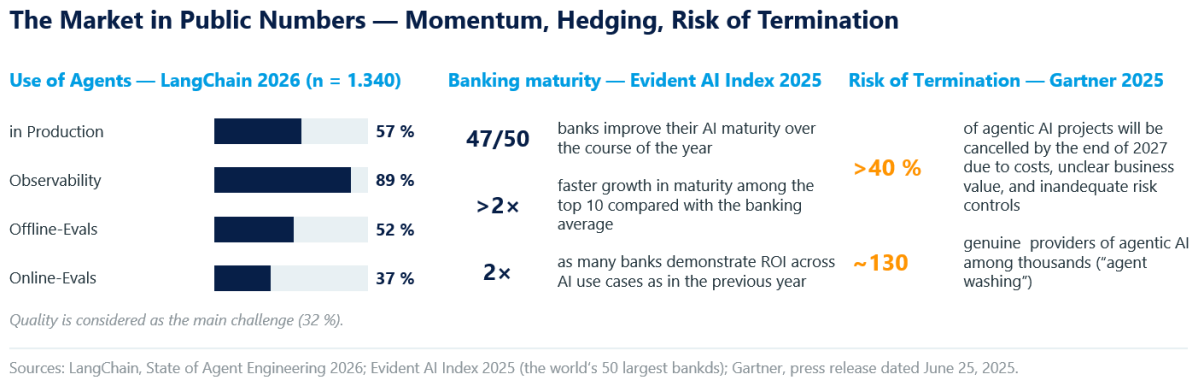


Figure 1: The market in public figures. Left: agent operations and safeguards (LangChain, State of Agent Engineering 2026). Centre: maturity development of the 50 largest banks worldwide (Evident AI Index 2025). Right: cancellation forecast and vendor field for agent projects (Gartner 2025).

Adviser assistants and service agents are the most widespread because they forgive errors. An imprecise research summary is corrected by the adviser, a wrong service answer by the next customer contact. Investment research agents work closer to the decision and therefore need the iteratively earned accuracy figures the accounts describe.[9] The most demanding field is the one in which we work ourselves: regulatory and compliance analysis. Here, an answer without evidence is not an answer but a risk. Every statement must point to a statute, a line in the body of rules or a contract section, and the path to it must withstand an audit.

Classical retrieval works like a very good similarity search: it splits documents into sections, compares their meaning vectors with the question and returns the most similar hits. For the question "What does Section 25c of the German Banking Act say?" that is sufficient. GraphRAG goes one step further: when the index is built, it stores not only text passages but entities and their relationships, that is, which statute refers to which, which circular specifies which regulation, which internal policy implements which external requirement. The result is a knowledge graph in which every document knows what it is connected to. The decisive difference appears with questions no single passage can answer: "Does our outsourcing agreement meet the requirements of the current regulation?" requires walking from the contract to the policy, from the policy to the statute, and from the statute to its specification, supplying every station along the way.

Agentic GraphRAG combines this structure with a decision-maker. Rather than merely querying the graph, an agent navigates it: it enters via vector search at the substantively nearest node, examines the outgoing references, follows the relevant one, discards the dead end, and logs every step with its

[9] LangChain (2026), ibid.

rationale. Compared with a one-off retrieval this has advantages that are decisive in regulatory work. Multi-step questions that demand a chain of evidence are answered considerably more reliably, because the agent searches along the reference structure instead of hoping for the similarity of a single passage. Since the agent reassesses after every step whether the body of evidence holds up, the answer quality depends less on whether the question happened to be phrased in the words of the rulebook.

It is precisely in this field that a property pays off which classical search architectures do not have. Regulatory knowledge is by nature a network. Articles refer to paragraphs, circulars specify regulations, internal policies implement external requirements, registers and evidence records document the implementation. An agent that has this network available as a graph can navigate along the references, step by step, with a rationale for each step. The ReAct paradigm, which interleaves reasoning and acting,[10] is the methodological basis for this. Current research on agentic RAG shows that such navigating methods are considerably more accurate precisely for multi-step questions than one-off search.[11] Graph-based approaches confirm this in independent studies. For the global understanding of large document collections, measured by summarizing queries rather than regulatory reviews, graph-supported answers are preferred in 72 to 83 per cent of direct comparisons with classical vector RAG,[12] and for multi-step questions the retrieval hit rate rises by up to 20 percentage points over established dense methods when the system also uses associations in the graph.[13]

Knowledge graphs are considered expensive to build and hard to keep current. That is true of the classical design, in which a predefined schema determines which entities and relationships may exist at all and in which every amendment of a directive threatens a rebuild. Our platform therefore builds the graph not by schema but by verification: relationships between passages are proposed automatically and verified individually before they become an edge. Meaning arises not at the named node but at the verified connection. New documents and amendments grow incrementally into the collection without the graph being rebuilt. A distilled verification model keeps the cost per edge decision in ongoing operation orders of magnitude below those of the build. How this method works and what we have tested it on, credit and collateral documentation, insurance portfolios, regulatory documents, is something we are happy to discuss with you.

[10] Yao, S. et al. (2023): ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023

[11] Singh, A., Ehtesham, A., Kumar, S. & Talaei Khoei, T. (2025): Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. arXiv:2501.09136

[12] Edge, D. et al. (2024): From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130

[13] Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M. & Su, Y. (2024): HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. NeurIPS 2024

4 · The architectural answer: a supervisor in front of the knowledge graph

We have built a platform that masters exactly this form of navigation. Complexity is not an end in itself here. A single prompt or a single retrieval step cannot provably resolve chains of references across document and rulebook boundaries. Multi-step regulatory reviews are the case in which the simpler solution no longer suffices. At its core is not a single agent but an orchestration. A supervisor breaks every request down into an explicit review plan with a few clearly named steps. For each step it starts a specialized sub-agent: (1) one for traversing the internal document graph, (2) one for the regulatory side, (3) one for hybrid keyword search, and for public questions one for the web. The sub-agents write their intermediate findings into a shared scratchpad, which is passed on to the next agent as numbered notes. An evaluator agent then checks whether the collected body of evidence answers the question. If it does not, the evaluator formulates the missing aspects and the supervisor replans. At most twice, however; after that the matter is escalated.

Figure 2 shows what this looks like for a typical outsourcing question. Via vector search and reranking the agent arrives at the initial article, selects the substantively most fitting among the neighbouring nodes, discards an obvious but irrelevant trail, and stops only when the finding, a contradiction between the contract template and the regulation, is evidenced with high confidence.

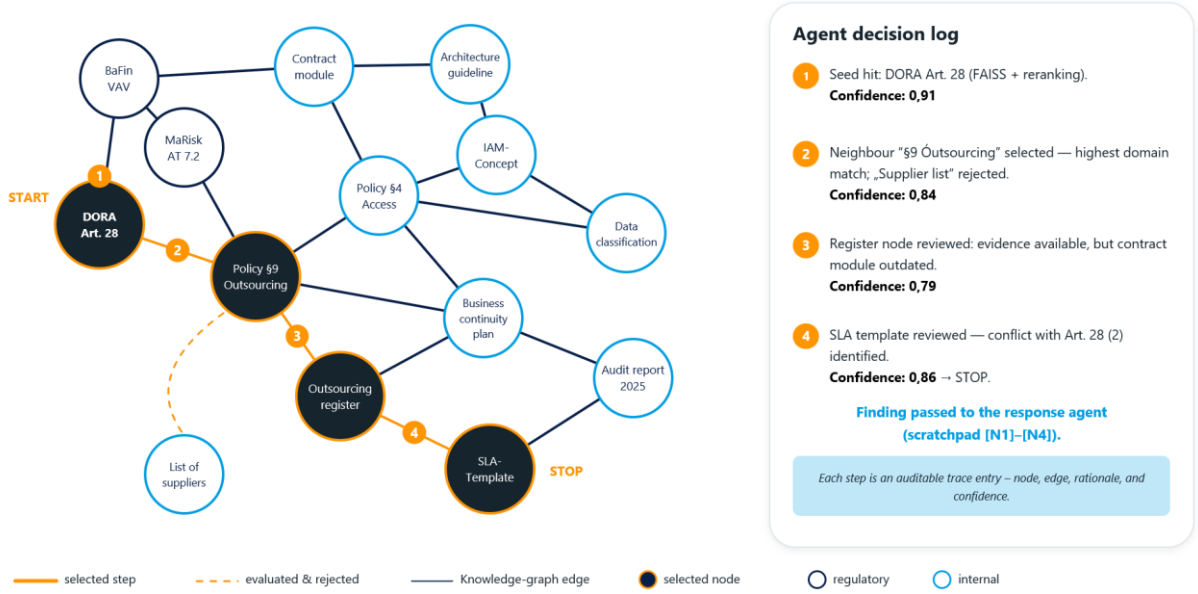


Figure 2: Agentic traversal of a knowledge graph using the example of an outsourcing question. Left: the graph with the chosen path and the discarded alternative; right: the agent's decision log. Illustrative example from platform operation.

Two decisions in this architecture were not a given for us. First, the plan is an artefact. It exists in writing before the first step runs, and it is therefore verifiable, both by the business function assessing the approach and by internal audit tracing the path later. Second, every step is limited. What counts as sufficient coverage is decided not in a configuration file but in conversation with those responsible for the subject matter. We calibrate these values anew for each engagement. We likewise document who set them, when and for what reason. What sounds like bureaucracy is in truth the implementation of

what the accounts describe as governance in the runtime[14] and what research into the stepwise operation of large models has prepared methodologically.[15] In its engineering guidelines, Anthropic observes that the most successful implementations manage without complex frameworks and instead combine simple, composable patterns, with the express reservation of building only as complex as the task requires.[16] The building blocks of this chapter, supervisor with a written plan, specialized sub-agents, evaluator as a step of its own, are such patterns: individually measurable, individually replaceable.

Mapping the five requirements from Chapter 2 against the platform yields the mapping in Figure 3. We present it here in full deliberately. Anyone wishing to check whether a service provider meets the requirements of the accounts should ask exactly these five questions and demand an architectural answer to each, not a slide.

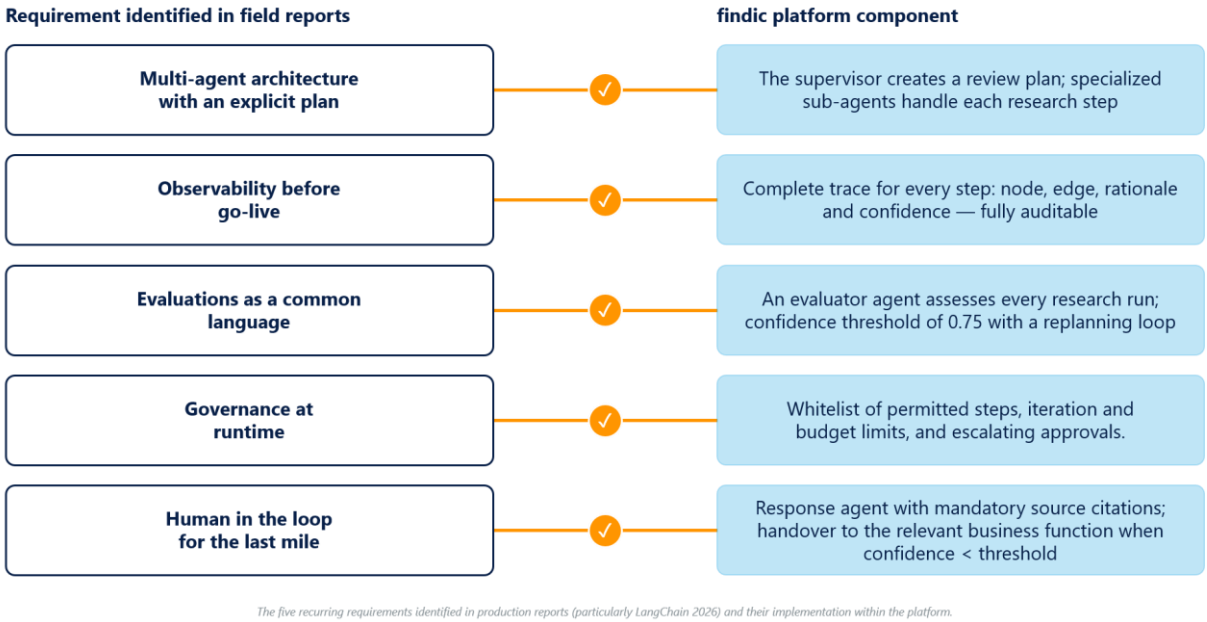


Figure 3: Mapping: the five recurring requirements of the accounts and the building blocks of the findic platform.

The answering agent at the end of the chain operates under a strict source-citation requirement. Every statement points to the collected notes, every note to nodes and passages. If the evaluator's confidence falls below the threshold, the request does not go out with a weak answer but with a handover to a human expert, with what is already evidenced and what is missing. That is the last mile the accounts speak of. With us, it is standard.

[14]LangChain (2026), ibid.
[15]Wei, J. et al. (2022): Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 35
[16]Anthropic (2024): Building Effective Agents. anthropic.com/research/building-effective-agents

5 · What published benchmarks show

The architectural questions we discuss here are now also evidenced by measurement. Two published benchmarks make visible exactly the two risks that matter in practice: insufficient evidence and unreliable repeatability.

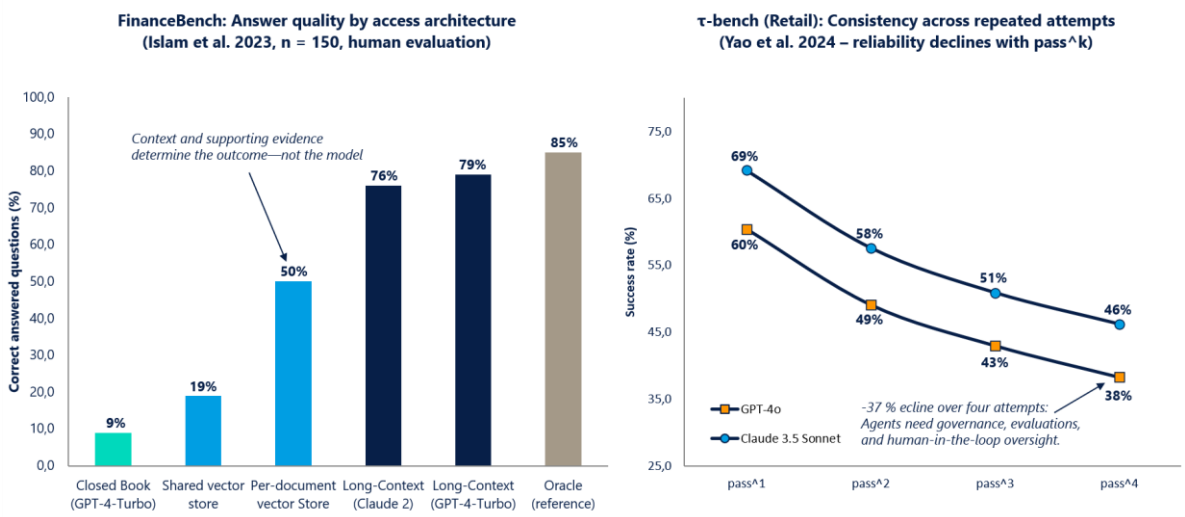


Figure 4: Published benchmarks. Left: answer quality by access architecture on FinanceBench, 150 questions with human evaluation (Islam et al. 2023); Closed Book and Oracle are reference points, not production methods. In the Oracle setting the model is given the annotated evidence page directly. Right: success rate across repeated runs on τ-bench Retail (Yao et al. 2024).

FinanceBench poses 150 questions about listed companies that are evaluated by humans and mostly demand numerical reasoning.[17] The result is a staircase. Without document access, even a strong model answers only 9 per cent correctly. A naively built shared vector store reaches 19 per cent. Poor retrieval is worse than none at all here, because it supplies wrong evidence with full conviction instead of declining. Document-sharp indexes, a dedicated index per annual report, reach 50 per cent. Loading the entire annual report into the context window instead yields 76 to 79 per cent, a result that holds for individual documents but scales neither to collections of thousands of rulebooks nor resolves their chains of references; the paper itself classifies this path as unsuitable for production because of cost, latency and document length.[18] The endpoint of the staircase is likewise not an access technique but a measurement reference: in the Oracle setting the model is given the exact evidence page annotated in the dataset directly; search is eliminated entirely. Even with this perfect evidence the model remains at 85 per cent. The remaining 15 per cent are pure reasoning errors over correct evidence.[19] The span between 9 and 85 per cent is not a model question. It is the difference between access architectures,[20] in other words the field in which GraphRAG with navigating agents operates.

τ-bench measures the second construction site: consistency. There, agents conduct realistic dialogues with tool use, and the success rate is measured across repeated runs of the same task. It falls steeply. GPT-4o solves a task on the first attempt in 60 per cent of cases. If the same task must succeed four

[17]Islam, P. et al. (2023): FinanceBench: A New Benchmark for Financial Question Answering. arXiv:2311.11944
[18]Islam, P. et al. (2023), ibid.
[19]Islam, P. et al. (2023), ibid.
[20]Lewis, P. et al. (2020): Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 33, 9459–9474

times out of four, 38 per cent remain.[21] Already these first τ-bench measurements from 2024 showed a massive consistency drop that manifests itself even more clearly in the successor work τ^2-bench under tightened conditions, namely cooperating users and systems accessing the same data in parallel. A system that acts differently in four out of ten repetitions is an annoyance in customer service; in regulatory review it is unusable. The answer cannot be to wait for the next model. The τ-bench authors themselves name the need for methods that make agents more consistent and more rule-compliant. Our answer lies in the architecture: limited step counts, evaluation of every piece of research, documented handover to the human. No benchmark demonstrates that this raises consistency in our own operation; only the operating metrics of our own platform show that.

For multi-step questions, answer accuracy rises by up to 20 percentage points over established dense methods when the system uses its associations like a memory network.[22] And in direct comparisons in which an evaluation model judges answer quality, graph-supported answers win against classical vector RAG in up to 83 per cent of cases.[23] That does not mean GraphRAG makes classical search obsolete. For targeted fact lookups, simple vector RAG often remains faster and more efficient. But in multi-step, networked regulatory work, the structure of the knowledge is decisive. An agent that traverses it extracts more than one that merely searches.

6 · Lessons from building it

Not every one of these lessons was pleasant, but all of them were cheaper than the alternative of learning them in production.

The first lesson concerned the supervisor itself. Our first version planned too ambitiously: five, six steps where two would have sufficed, because the model preferred to answer uncertainty with further steps rather than with a decision. Today four is the hard upper limit within which the supervisor continues to plan freely; most requests it resolves in two steps. Less plan is more, because every step costs context, time and opportunities for error.

The second lesson concerns the evaluator. Its judgement is only as good as the professional definition of "sufficient", and that cannot be pulled from a prompt generator. We formulated the evaluation criteria together with auditors and business functions: which aspects of an outsourcing question must always be examined, from what point a reference is outdated, what constitutes a sustainable contradiction. That is expert work, embedded in a prompt.

Third: the scratchpad has proved to be the artefact that creates trust. It is not the final answer that convinces business functions and internal audit, but the numbered intermediate findings with a rationale and a confidence score for each step. We originally built the traces for ourselves as developers. Today they are the first thing we show in acceptance reviews.

[21] Yao, S. et al. (2024): τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045

[22] Gutiérrez, B. J. et al. (2024): HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. NeurIPS 2024

[23] Edge, D. et al. (2024): From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130

Fourth: every request starts with a fresh state. Sessions remain strictly separated; intermediate findings travel only within a request. That sounds self-evident, but it is a deliberate architectural decision and indispensable in an institution whose permission concepts are exactly that: permission concepts.

7 · Limits and what comes next

The system does not replace legal or audit assessment. It accelerates the body of evidence. The decision remains with the human. The cited benchmarks measure components. They say little about operation in a specific institution with specific permissions, data flows and legacy systems. Anyone who derives production readiness for you from benchmarks is selling slides. What counts are the operating metrics of your own platform, and those belong in a dashboard, not in a white paper.

The next stages of development lie in the strict confinement of every agent to the data domain of the person asking,[24] in the continuous monitoring of regulatory sources, where a finding arises the moment a regulator publishes, and in second agents that critically counter-read the findings of the first. Institutions that build this operational discipline now will in future manage agents like vetted colleagues: with a job description, a log and an approval process. The others will continue to pilot pilots.

If you would like to check whether your architecture meets the five requirements: we are happy to put the question to ourselves together with you in a conversation.

[24] LangChain (2026), ibid.

8 · Sources and further reading

Anthropic. (2024). Building Effective Agents. anthropic.com/research/building-effective-agents. Engineering guidelines on simple, composable agent patterns and on the deliberate limitation of architectural complexity; used in the text to classify supervisor, sub-agents and evaluator as production-proven patterns.

Barres, V., Dong, H., Ray, S., Si, X. & Narasimhan, K. (2025). τ^2-bench: Evaluating Conversational Agents in a Dual-Control Environment. arXiv:2506.07982. Follow-up work to τ-bench in which user and system access the same data cooperatively and in parallel; used in the text as evidence that the consistency drop becomes even more pronounced under tightened conditions.

BBVA & OpenAI. (2026). Expansion of ChatGPT Enterprise to more than 100,000 employees (starting point 2024: 3,300 licences). bbva.com / openai.com. Company announcement on scaling an enterprise-wide AI assistant; used in the text as an example of the transition from pilot operation to broad production use.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O. & Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. Microsoft Research / arXiv:2404.16130. Research on extending classical RAG approaches with knowledge graphs and community summaries; used in the text as evidence for graph-supported answers to complex queries.

Evident. (2025). Evident AI Index 2025. Annual ranking of the AI maturity of the fifty largest banks worldwide; used in the text (Figure 1) as evidence of the industry's maturity development: nearly all institutions improved on the previous year, and the top 10 are developing markedly faster than the average.

Gartner. (2025). Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027. Press release of 25 June 2025. Cancellation forecast for agent projects amid escalating costs, unclear benefit and inadequate risk controls, with only around 130 genuine vendors in the market (“agent washing”); used in the text (Figure 1) as a risk and market classification.

Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M. & Su, Y. (2024). HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), arXiv:2405.14831. Graph-based retrieval modelled on the hippocampus; used in the text as evidence that graph-based retrieval improves the retrieval hit rate for multi-step questions by up to 20 percentage points over state-of-the-art RAG methods (MuSiQue, 2WikiMultiHopQA).

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N. & Vidgen, B. (2023). FinanceBench: A New Benchmark for Financial Question Answering. arXiv:2311.11944. Benchmark with 150 finance-related questions and human evaluation; used in the text to show the influence of the access architecture on answer quality and evidence.

LangChain. (2026). Financial Services Agents in Production: Lessons building agents from J.P. Morgan, Chime, and Bridgewater. LangChain Inc., San Francisco. Accounts of the production systems of J.P. Morgan (Ask DAVID), Bridgewater Associates (PAT) and Chime (Jade); central basis for the five recurring requirements for production agents in financial services identified in the text.

LangChain. (2026). State of Agent Engineering. Survey of more than 1,300 professionals on the production level, observability and quality problems of agentic systems; used in the text as a market classification and frame of reference for production-ready agent architectures.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33. Foundational paper on combining parametric language models with an external, retrievable knowledge base; cited in the text as the methodological reference for the access architecture as the decisive variable.

OpenAI & Morgan Stanley. (2024). Morgan Stanley uses AI evals to shape the future of financial services. openai.com/index/morgan-stanley. Publicly documented practice report on the use of AI evaluations in financial services; used in the text as an example of production adviser assistance and systematic model evaluation.

OWASP Foundation. (2025). OWASP Top 10 for Large Language Model Applications. owasp.org/www-project-top-10-for-large-language-model-applications. Threat catalogue for LLM applications; cited in the text as a reference for governance, runtime controls and security requirements of agentic systems.

Sierra Research. (2024). τ-bench Leaderboard. github.com/sierra-research/tau-bench. Official results table (pass^1 to pass^4); basis for the right-hand curve in Figure 4.

Singh, A., Ehtesham, A., Kumar, S., Talaei Khoei, T. & Vasilakos, A. V. (2025). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. arXiv:2501.09136. Systematisation of the agentic RAG research direction; used in the text as evidence that navigating, multi-step retrieval methods are considerably more accurate than one-off search.

Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (DORA), OJ L 333 of 27.12.2022. Regulatory framework for ICT risk management and operational resilience; used in the text as the supervisory reference point for evidential capability, third-party risks and runtime governance.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act), OJ L of 12.7.2024. European regulatory framework for AI systems; used in the text for requirements on documentation, supervisability and human oversight.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, 35. Foundational text on stepwise reasoning of large language models; cited in the text as methodological groundwork for the reasoned, stepwise operation of agentic systems.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. Eleventh International Conference on Learning Representations (ICLR 2023). Foundational text on combining reasoning and acting in language models; used in the text as the methodological basis of agentic navigation.

Yao, S. et al. (2024). τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv:2406.12045; published at ICLR 2025. Benchmark for measuring success rates of agentic systems across repeated, realistic tool interactions; used in the text as evidence of consistency problems and repeatability of production agents.

9 · Contact



Udo Jacobasch
Partner
Udo.Jacobasch@findic.de
Office Berlin



Dr. Josef Dirnberger
Manager
Josef.Dirnberger@findic.de
Office München

This publication has been prepared for general information purposes only and does not take into account the individual investment objectives or the risk appetite of the reader. The reader should not take any action on the basis of the information contained in this publication without first obtaining specific professional advice. findic gmbh accepts no liability for damages arising from any use of the information contained in this publication. This document may not be reproduced, copied, distributed or transmitted in any form or by any means without the prior written permission of findic gmbh.

© findic gmbh 2026. ALL RIGHTS RESERVED.

Hamburg

Kurze Mühren 20,

D-20095 Hamburg

Phone +49 40/303740-0

E-Mail info@findic.de

Zurich

Gutenbergstrasse 1,

CH-8002 Zurich

Phone +41 44/56098-00

E-Mail info@findic.ch