



2026 EDITION · VSTORM

The LLM Book

Why most AI pilots still do not reach production —
and what mid-market companies do differently when
they get there. Every figure sourced.

What is inside

00

How to read this edition

- 0.1 What changed, and why this edition is shorter
- 0.2 Who this is for

- 0.3 What we do, so you can discount for it
- 0.4 The sourcing rule

01

The gap that did not close

- 1.1 Where the money went
- 1.2 The measurement gap
- 1.3 Deployment is not scale

- 1.4 One number we will not print
- 1.5 The diagnosis

02

What actually changed since 2024

- 2.1 Capability stopped being the constraint
- 2.2 The open-weight picture inverted

- 2.3 What did not change

03

Agents, mechanically

- 3.1 What an agent is
- 3.2 Tool use, and the standard that settled it

- 3.3 Where adoption actually sits
- 3.4 What agents are not

04

Retrieval, still the workhorse

- 4.1 Why a million tokens did not kill RAG
- 4.2 Grounding is a design decision, not a feature

- 4.3 What validation looks like when the stakes are real

05

Reliability is the product

- 5.1 The loop is the architecture
- 5.2 Governance is not overhead, and there is a number

- 5.3 What we evaluate before a system ships

06

What it actually costs

- 6.1 The headline rate is not the price
- 6.2 Nobody is watching the meter

6.3 The cost question worth asking

07

Governance you can actually run

- 7.1 The EU AI Act timeline changed, and most content has not caught up
- 7.2 Ownership is the variable nobody assigns

7.3 A governance minimum that fits a mid-market company

08

How we deliver

- 8.1 TriStorm
- 8.2 Ownership transfer

8.3 Open source

09

The next 90 days

- 9.1 If you have pilots that have not shipped
- 9.2 If you are choosing what to build next

- 9.3 If you already have agents in production
- 9.4 What to ask any vendor, including us

10

Sources

- 10.1 Surveys and industry research
- 10.2 Vendor and standards documentation

- 10.3 Regulation
- 10.4 Vstorm project evidence

How to read this edition

0.1 What changed, and why this edition is shorter

The first edition of this book was written in 2024. It explained what a language model was to readers who had mostly met one through a chat window. That book named BERT twenty-eight times and GPT-4 once.

Two years later the explaining is largely done. Mid-market companies are not asking what a language model is. According to RSM's Middle Market AI Survey, published July 2026, **86% of middle-market organisations have partially or fully integrated AI into operations** and 73% use generative AI. The question moved.

The question now is why so little of it reaches production, and what to do about that.

This edition is shorter than the last one, deliberately. Every number in it carries a publisher and a year. Where we could not source a figure, we cut it rather than round it. Several widely repeated statistics about AI adoption did not survive that rule and do not appear here — including one you have almost certainly seen quoted this year, which we discuss in section 1.4.

0.2 Who this is for

Operations, technology and finance leaders at companies between roughly \$25M and \$500M in revenue who have run at least one AI pilot and are deciding what to do next.

You do not need to write code to use this book. You do need to be the person who signs off on where the next budget goes.

0.3 What we do, so you can discount for it

Vstorm is an applied agentic AI engineering consultancy. We build production-grade agents for mid-market companies in Europe, the US and the Middle East, and we are the first AI consultancy in the Agentic AI Foundation, the Linux Foundation body that now governs the Model Context Protocol.

That gives us a view, and it gives us an interest. Where this book describes a pattern we sell, we say so. Where our own projects are the evidence, they are named, with the client's permission and the published figure.

0.4 The sourcing rule

Three rules governed the writing:

Every statistic names its publisher and year. Where two credible sources disagree, we show both and say which we trust and why. Where a number is ours, it names the client and the validation method.

Numbers without a benchmark are marketing. We have tried not to make any.

The gap that did not close

1.1 Where the money went

Gartner predicts that **over 40% of agentic AI projects will be cancelled by the end of 2027**, citing escalating costs, unclear business value and inadequate risk controls (Gartner, June 2025).

That prediction is doing well. S&P Global Market Intelligence found that **42% of companies abandoned most of their AI initiatives** in 2025, up from 17% the year before, with **46% of proofs of concept scrapped** before production (S&P Global Market Intelligence, 2025, n>1,000, North America and Europe).

Deloitte's *State of AI in the Enterprise 2026*, fielded across 3,235 respondents in 24 countries, puts it from the other side: **only 25% of organisations have moved more than 40% of their AI pilots into production** (Deloitte, January 2026).

Two years of capability improvement did not move this number. That is the fact worth sitting with.

1.2 The measurement gap

BCG surveyed 152 CEOs at companies above \$500M revenue in July 2026 and found the sharpest articulation of the problem yet published.

89% of those CEOs report a cost or revenue benefit from AI in targeted areas. Only 14% have a clearly defined P&L impact for all of their AI initiatives.

That is a 75-point gap between believing something works and being able to show it. BCG also found that **more than half of CEOs cite the missing link between AI and P&L as a barrier to scaling** (BCG, July 2026).

McKinsey's *State of AI*, fielded across 1,993 respondents in 105 nations, gives the same shape in different numbers: **39% of organisations report EBIT impact at enterprise level, and most of those attribute under 5% of EBIT to AI** (McKinsey, November 2025).

The benefit is real. It is also local, and it does not aggregate. A pilot that saves a team four hours a week is a genuine result and an unfundable one, because nobody can show it in the accounts.

1.3 Deployment is not scale

Adoption statistics in this market are quoted carelessly, because three different bars get reported as one.

WHAT WAS MEASURED	FIGURE	SOURCE
Organisations scaling an agentic AI system	23%	McKinsey, Nov 2025
Executives whose organisation has deployed AI agents	52%	Google Cloud / NRG, Sep 2025
Enterprises with agents deployed	53%	KPMG, Q2 2026
Maximum scaling of agents in any single business function	no more than 10%	McKinsey, Nov 2025

These are not contradictory. They measure different things on different populations, and the honest summary is this: **a majority have deployed something, roughly a quarter have scaled anything, and no single business function anywhere is above about 10%.**

KPMG is the only quarterly series among them, and it carries the most uncomfortable finding in this section. Between Q1 and Q2 2026, enterprise agent deployment fell from 55% to 53%, and the share scaling agents across multiple functions fell from 33% to 29% (KPMG AI Quarterly Pulse Survey, June 2026, n=204 US leaders at \$1B+ revenue).

Deployment went down. Not dramatically, and one quarter is not a trend. But it is the first published reversal in this series, and it does not look like a market running out of capability.

1.4 One number we will not print

You have probably seen the claim that 95% of enterprise AI projects deliver zero return. It comes from a July 2025 working paper by MIT NANDA, and it is the most-quoted and most-misquoted figure in this field.

Three things about it are worth knowing before you repeat it in a board pack. It measures **zero P&L return**, not pilots that failed to reach production. The pilot-to-production figure in the same report is 5% for task-specific tools. Its evidence base is 52 interviews, 153 survey responses and 300 public initiatives — not a large-sample survey. And it is a non-peer-reviewed working paper that attracted substantial methodological criticism on publication. We cite S&P Global’s 42% instead. It is survey-grade, it has a sample above a thousand, and the year-over-year jump from 17% is a stronger story than a headline percentage nobody can reproduce.

This is not pedantry. If you build an investment case on a number your CFO can dismantle in one search, you lose the case and the credibility together.

1.5 The diagnosis

The pilot-to-production gap is not a model problem.

Capability rose sharply over the two years in which these abandonment rates also rose. If the constraint were the model, better models would have fixed it. They did not.

What the surveys actually name, consistently, across different publishers and populations, is a short list: data that is not ready, ownership that is not assigned, controls that do not exist, and measurement that cannot connect the work to the accounts. We take those in turn in chapters 5, 6 and 7.

What actually changed since 2024

2.1 Capability stopped being the constraint

A reader returning from 2024 would find four things different.

Context windows went from a differentiator to a commodity. One million tokens is the frontier default in 2026. GPT-5.6 ships with a 1.05M window across all three tiers (OpenAI, July 2026). Claude Opus 5, Sonnet 5 and Fable 5 all carry 1M (Anthropic platform documentation, 2026). DeepSeek-V4-Pro and V4-Flash carry 1M with 384K maximum output (DeepSeek API documentation, 2026). Grok 4.5 carries 500K (xAI documentation, 2026).

Reasoning became a dial rather than a model class. OpenAI exposes six effort levels — none, low, medium, high, xhigh, max — on a single model (OpenAI reasoning guide, 2026). Anthropic ships thinking enabled by default on Opus 5 and adaptive thinking on Sonnet 5 (Anthropic release notes, 2026). You no longer choose a reasoning model; you choose how much reasoning to buy per request.

Multimodality became table stakes at the input layer. Text and image in, text out, is universal across frontier models. Generation of other modalities remains specialist: OpenAI keeps separate image, realtime-speech and transcription models, while Google has moved furthest toward a single omni model (vendor documentation, 2026).

Price per unit of capability fell hard. The cleanest single piece of evidence is Anthropic's own top tier: Claude Opus 4 and 4.1 were priced at \$15 per million input tokens and \$75 per million output. Opus 4.5 onward, including Opus 5, is priced at \$5 and \$25 — a threefold cut at the same product tier (Anthropic pricing documentation, 2026).

2.2 The open-weight picture inverted

In 2024 the open-weight story was Meta's. In 2026 it is not.

Meta launched Muse Spark in April 2026 — proprietary, cloud-only, no downloadable weights, built from scratch by Meta Superintelligence Labs. Existing Llama models remain open source but are described as receiving incremental updates and maintenance rather than frontier investment (The New Stack, April 2026).

Alibaba's Qwen went the same direction at the top of its range. Every Max-tier release since September 2025 has been proprietary, and the flagship Qwen3.8-Max is API-only (industry reporting, July 2026).

What replaced them is a broader field. Epoch AI measures the gap between open and closed models at **four months, or eight points on the Epoch Capabilities Index**, since January 2026 — a gap Epoch itself compares to the distance between two consecutive GPT releases (Epoch AI, May 2026).

Four months is not nothing. It is also not the two-year canyon that justified writing off open weights in 2024.

For a mid-market company the practical consequence is narrower than the headlines suggest. The choice between an open-weight model you host and an API you call is now mostly a question of data jurisdiction, unit economics at your volume, and whether you have anyone to operate it — not a question of whether the open model is good enough.

2.3 What did not change

Three things a 2024 reader would recognise immediately.

Retrieval is still how you ground a model in your own data, and the million-token context window did not remove the need for it. We take that up in chapter 4.

Evaluation is still the part teams skip, and still the part that decides whether a system survives contact with production.

And the models still do not know anything about your business that you have not given them. Every capability in section 2.1 is a capability in the abstract. None of it is knowledge of your order book.

Agents, mechanically

3.1 What an agent is

An agent is a language model given tools, a goal, and permission to loop.

That is the whole of it. The model receives a task, decides which tool to call, reads the result, and decides again. It plans, executes and delegates. It does not think or understand, and describing it that way makes the engineering harder, not easier, because it obscures where the failure modes live.

The loop is what distinguishes an agent from a prompt. A prompt produces one output from one input. An agent produces a sequence of actions, each conditioned on the result of the last, until it decides it is finished or something stops it.

Everything difficult about agents follows from that sentence. A prompt that goes wrong produces a bad paragraph. A loop that goes wrong produces forty minutes of unsupervised tool calls.

3.2 Tool use, and the standard that settled it

For most of 2024 and 2025, connecting a model to a system meant writing a bespoke integration per model per system.

That is now largely a solved problem, and it was solved by a standard rather than a vendor. Anthropic donated the Model Context Protocol to the **Agentic AI Foundation**, a Linux Foundation entity, in December 2025, alongside Block's goose and OpenAI's AGENTS .md (Linux Foundation, December 2025).

The membership list is the fact that matters. Platinum members include AWS, Anthropic, Block, Bloomberg, Cloudflare, Google, Microsoft and OpenAI. Gold includes Cisco, Datadog, Docker, IBM, JetBrains, Okta, Oracle, Salesforce, SAP, Shopify, Snowflake and Twilio (Linux Foundation, December 2025).

Every major vendor signed the same standard. That does not happen often, and it is the strongest available signal that tool integration is no longer where the risk sits.

At the December 2025 announcement the ecosystem counted more than 10,000 published MCP servers and over 97 million monthly SDK downloads. A registry snapshot in May 2026 counted 9,652 current server records and 15,926 GitHub repositories tagged as MCP servers.

The specification has moved with it. The current version, released 28 July 2026, made the protocol core stateless, added header-based routing so gateways can route without parsing message bodies, made list results cacheable, and hardened authorization (MCP project, July 2026). The published 2026 roadmap names audit trails, SSO-integrated authentication and gateway behaviour as the recurring enterprise pain points.

That roadmap is a useful tell. The unsolved problems in agent tooling are now operational, not conceptual.

3.3 Where adoption actually sits

Stacklok's *State of MCP in Software 2026* puts enterprise production adoption at **29% limited production and 12% broad production**, with 41% combined across planning, pilot and production.

A figure claiming that 78% of enterprise AI teams had adopted MCP circulated widely in early 2026. It has no traceable source and was retracted by at least one publication that had carried it. Do not use it.

3.4 What agents are not

They are not a replacement for a workflow you have not defined. An agent given an ambiguous goal produces ambiguous work faster than a human would.

They are not cheaper by default. Chapter 6 covers why.

And they are not autonomous in the sense the word implies in ordinary English. IBM's Institute for Business Value, surveying 2,000 C-level technology executives across 33 countries, found organisations averaging **54 agent incidents per year requiring human correction, of which 17% were high-severity** — and among those, 37% caused data exposure and 33% cascading failures (IBM IBV, June 2026).

That is the number to bring to anyone proposing an unsupervised agent in a customer-facing path.

Retrieval, still the workhorse

4.1 Why a million tokens did not kill RAG

The argument was straightforward and it was wrong. If a model can hold a million tokens, put the documents in the prompt and skip retrieval.

Three things stopped it.

Effectiveness degrades before the window does. Independent testing through 2026 finds synthesis quality falling away well before the advertised limit, and advertised context regularly exceeds usable context in shipped products — an open issue on OpenAI's Codex reports GPT-5.6 Sol capped near 372K against a 1.05M model specification.

Long context is priced as a surcharge. OpenAI applies a 2x multiplier on base pricing for long-context requests on most models (OpenAI pricing documentation, 2026). Filling a million-token window is not a rounding error at production volume.

And your corpus is bigger than any window. A mid-market company's document estate is not a million tokens. It is not ten million. Retrieval is not a workaround for small windows; it is how you select what matters from something that will never fit.

4.2 Grounding is a design decision, not a feature

Retrieval-augmented generation, described mechanically: when a question arrives, search your own corpus, select the passages most likely to answer it, and put those passages in front of the model as context before it generates.

The quality of the answer is decided by the selection step, not the generation step. This is the part that teams under-build, because the retrieval layer is unglamorous and the model is the thing on the slide.

Two failure modes account for most of what goes wrong. The system retrieves the wrong passages and the model answers confidently from them. Or the system retrieves nothing relevant and the model answers from its training data instead, without saying so.

Both are solvable, and both are solved in the same place: by deciding what happens when retrieval fails, before the system is live.

4.3 What validation looks like when the stakes are real

For Schmitt-Thompson Clinical Content we built an agentic retrieval system for clinical triage guidelines — the material a nurse follows when deciding whether a caller needs an emergency room, a same-day appointment or home care.

In that setting a misread is not an edge case to tolerate. It is a patient safety risk.

The architecture is staged retrieval with a validation pass, rather than one model answering in a single hop. The published result is **zero hallucination events across 329+ nurse-validated scenarios**.

The number that matters there is not the zero. It is the 329. A zero without a denominator is a claim; a zero across a named, human-validated scenario set is a measurement. When a vendor shows you the first, ask for the second.

Reliability is the product

5.1 The loop is the architecture

For Synera, an engineering platform whose customers include NASA, Airbus and BMW, we built a text-to-workflow agent: an engineer describes what they need in plain language, and the system produces a working node graph.

Building one of those workflows by hand took up to two hours. The published figure after deployment is **three minutes**.

The interesting part is not the speed. It is what made the speed safe.

Synera's platform runs on a proprietary node interpreter, and the model generates pseudo-Python that the interpreter converts into nodes. Modern models are trained hard on producing correct Python — which is precisely the wrong output here, and they produced it confidently. The model was improving on the specification while making the result unusable.

The fix was a validator that inspects every output before it reaches the interpreter. Non-compliant code goes back to the model for regeneration, and the loop runs until it passes. Synera reports zero hallucinations in production, maintained through that multi-layer validation loop.

Note the shape. The reliability did not come from a better model. It came from refusing to let the model's output reach the system unchecked.

5.2 Governance is not overhead, and there is a number

The most useful finding in the 2026 literature, for anyone arguing this internally, is IBM's.

Organisations with governance controls embedded in their agent infrastructure, rather than applied manually, deploy **16 times more agents**, report **18% higher operating margins** and experience **25% fewer incidents** than those governing manually (IBM IBV, June 2026).

Read that as a scaling result rather than a compliance one. Embedded controls are what let you run many agents at all. Manual review does not scale past the first few, which is a reasonable explanation for why no business function anywhere is above 10% agent penetration.

The gap is wide. Deloitte finds **only 21% of organisations have governance mature enough to manage agentic AI risk** (Deloitte, April 2026). IBM finds **77% of CIOs and CTOs say AI adoption is outpacing their governance capability**, **66% are accountable for AI systems they do not fully control**, and **11% feel fully ready** for the scale of deployment they expect (IBM IBV, June 2026).

5.3 What we evaluate before a system ships

Four questions, asked before anything reaches a user.

What does the system do when it does not know? A system with no defined failure behaviour has one anyway, and it is usually to answer confidently.

What is the measured accuracy, against what set, validated by whom? If the answer has no denominator, there is no measurement.

What is the blast radius of a wrong answer? A misrouted internal ticket and a misrouted clinical triage call are not the same system even when they are the same architecture.

And who sees it when it fails? IBM's 54 incidents a year happen whether or not anyone is watching. The difference between a mature deployment and an immature one is whether they are counted.

What it actually costs

6.1 The headline rate is not the price

Published per-token pricing has fallen sharply at every tier. Here is the current frontier, per million tokens, input and output.

MODEL	INPUT	OUTPUT
GPT-5.6 Sol	\$5	\$30
GPT-5.6 Terra	\$2	\$12
GPT-5.6 Luna	\$0.20	\$1.20
Claude Opus 5	\$5	\$25
Claude Sonnet 5	\$3	\$15
Grok 4.5	\$2	\$6
Gemini 3.6 Flash	\$1.50	\$7.50
DeepSeek-V4-Flash	\$0.14	\$0.28

Vendor pricing documentation, fetched August 2026. Introductory and promotional rates excluded.

Four things make the real bill higher than that table implies, and none of them are hidden — they are simply not on the pricing page you looked at first.

Reasoning tokens bill as output. The model's internal reasoning is invisible in the response and charged at the output rate (OpenAI and xAI documentation, 2026). On a high-effort setting this can dominate the cost of a request.

Long context carries a multiplier. OpenAI applies 2x on base pricing for long-context requests on most models.

Tokenizers change. Anthropic's new tokenizer, shipped with Sonnet 5, produces roughly 30% more tokens for the same text. The rate did not move; the bill did.

And a premium tier appeared above the flagship. Pro-tier models are priced at \$30 input and \$180 output — six times the flagship rate.

6.2 Nobody is watching the meter

KPMG found that **only 26% of enterprises have full real-time cost visibility on AI spend**, at companies averaging \$202M in projected twelve-month AI investment (KPMG, Q2 2026).

Three quarters of large enterprises cannot see what their AI costs this week.

For a mid-market company the exposure is smaller in absolute terms and larger in proportion. An agent loop with no ceiling, called from a public page, is an uncapped liability against a budget that has one.

6.3 The cost question worth asking

Not “what does the model cost per token”. The useful question is what a completed unit of work costs — one resolved ticket, one processed claim, one generated workflow — and how that compares to what it costs today.

That number is calculable before you build. It requires knowing your volume, your acceptable latency, the effort setting the task actually needs, and how often the loop retries. Most teams discover the retry rate in production. It is cheaper to discover it in a Proof of Value.

Governance you can actually run

7.1 The EU AI Act timeline changed, and most content has not caught up

If you read anything written in 2025 about the August 2026 compliance cliff, the dates in it are wrong.

Regulation (EU) 2026/1744, the Digital Omnibus on AI, was published in the Official Journal on 24 July 2026 and entered into force on 27 July 2026 — days before the deadline it moved.

In force now: prohibited practices and AI literacy obligations since February 2025; general-purpose AI model obligations since August 2025; and from 2 August 2026, the general application date together with the Article 50 transparency duty — the obligation to disclose that a user is interacting with an AI system. That one was explicitly not delayed.

Moved: standalone high-risk systems under Annex III — employment, education, credit scoring, critical infrastructure, law enforcement — from August 2026 to **2 December 2027**. High-risk systems embedded in regulated products from August 2027 to August 2028. Marking of AI-generated content, for systems already on the market, to December 2026. Regulatory sandboxes to August 2027.

The practical read for a mid-market company: the heavy high-risk obligations most 2025-vintage content warned about are now eighteen months further out. What binds this month is lighter — tell people when they are talking to an AI system, and mark generated content from December.

That is a smaller lift than the panic suggested. It is not nothing, and the Article 50 duty applies to exactly the customer-facing chat deployments most companies started with.

7.2 Ownership is the variable nobody assigns

BCG's CEO survey found governance concentrated where it is comfortable and absent where it matters: **82% of companies include technology in AI governance, 30% include HR**. More than half of CEOs named people and organisational redesign as a barrier to scaling (BCG, July 2026).

McKinsey found that organisations with explicit responsible-AI ownership scored materially higher on maturity — 2.6 against 1.8 on their index (McKinsey, March 2026).

IBM found **70% of organisations have business units deploying faster than IT can track** (IBM IBV, June 2026).

These describe one condition from three angles. The technology function owns the risk register. The business functions own the deployments. Nobody owns the join.

7.3 A governance minimum that fits a mid-market company

You are not going to staff a responsible-AI office. You do not need one. Five things carry most of the weight.

A register of every AI system in use, including the ones a business unit started without telling anyone. An owner named per system — a person, not a department. A defined failure behaviour per system, written down before launch. An incident count, reviewed monthly, because IBM's 54 a year is an average and yours is a fact you either know or do not. And a disclosure line on anything customer-facing, which Article 50 now requires anyway.

That is a morning's work to define and an hour a month to run. It is also the difference between the 21% Deloitte found with mature governance and the rest.

How we deliver

8.1 TriStorm

Vstorm runs delivery on the TriStorm methodology, in three phases.

Strategic alignment and planning. Before anything is built: which process, what it costs today, what the target is, how it will be measured, and what happens when the system is wrong. The output is a decision about whether to build at all. Some of these end with a recommendation not to.

Proof of Value. Not a proof of concept. A concept is proven by a demonstration; value is proven against the number agreed in phase one, on your data, in conditions close enough to production that the retry rate and the failure modes are real.

Process augmentation. The system goes into the workflow it was built for, with the controls, the runbooks and the handover that let your team operate it.

8.2 Ownership transfer

The engagement ends with your organisation running the system independently. Runbooks, knowledge transfer, no vendor lock-in.

We state that plainly because the alternative is common and expensive: a working system nobody internal can modify, which decays from the day the consultancy leaves.

8.3 Open source

We publish the patterns that survived production at oss.vstorm.co — Pydantic DeepAgents and a dozen other repositories. Vstorm is a contributor to the Pydantic AI ecosystem and the first AI consultancy in the Agentic AI Foundation.

That is partly how we hire and partly how we keep ourselves honest. Code that is public gets read.

The next 90 days

9.1 If you have pilots that have not shipped

Count them, and count what each one cost. S&P Global's 46% scrapped rate is an average made of exactly this, unexamined.

For each one still alive, answer BCG's question: what is the P&L impact, in the accounts, if it works. If you cannot answer for a pilot, you have found why it has not shipped. Only 14% of CEOs can answer it for everything they are running.

9.2 If you are choosing what to build next

Pick a process with volume, a measurable cost today, and a tolerable failure mode. Volume makes the arithmetic work. A measurable cost gives you a denominator. A tolerable failure mode buys you the room to be wrong while you learn the retry rate.

Avoid, for a first production system, anything where a wrong answer reaches a customer unreviewed. Not because it cannot be done — chapter 4 describes one that was — but because doing it properly requires the validation apparatus you do not have yet.

9.3 If you already have agents in production

Find out how many incidents you had last quarter. If the answer is none, that is not a result, it is a missing counter.

Then check the governance question that IBM priced: are your controls embedded in the infrastructure or applied by people reviewing things. The gap between those two is 16x in deployment capacity, on their numbers.

9.4 What to ask any vendor, including us

What is the measured accuracy, against what set, validated by whom.

What does the system do when it does not know.

What does one completed unit of work cost, at our volume.

Who owns it when you leave.

A vendor who answers those four with numbers is worth the meeting. A vendor who answers them with adjectives has told you something useful too.

Sources

10.1 Surveys and industry research

Gartner,
Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027
, June 2025.

Gartner,
Lack of AI-Ready Data Puts AI Projects at Risk
, February 2025 (n=1,203 data management leaders).

S&P Global Market Intelligence, AI abandonment findings, 2025 (n>1,000, North America and Europe).

Deloitte,
State of AI in the Enterprise 2026
, January 2026 (n=3,235, 24 countries).

Deloitte,
AI agents are scaling faster than governance
, April 2026.

McKinsey,
The State of AI
, November 2025 (n=1,993, 105 nations).

McKinsey,
State of AI trust in 2026
, March 2026 (n≈500).

BCG, CEO survey on AI cost and revenue benefits, July 2026 (n=152 CEOs, revenue ≥\$500M).

KPMG,
AI Quarterly Pulse Survey Q2 2026
, June 2026 (n=204 US leaders, \$1B+ revenue).

IBM Institute for Business Value, CIO/CTO AI control gap study, June 2026 (n=2,000 C-level technology executives, 33 countries).

RSM US,
Middle Market AI Survey 2026

, July 2026 (n=1,030, US \$30M–\$10B and Canada \$30M–\$1B revenue).

Google Cloud / National Research Group, AI agent deployment study, September 2025 (n=3,466 executives, 24 countries). Vendor-commissioned.

MIT NANDA,

The GenAI Divide: State of AI in Business 2025

, July 2025. Cited in section 1.4 as a caution, not as evidence.

Epoch AI, open–closed capability gap analysis, May 2026.

Stacklok,

State of MCP in Software 2026

.

10.2 Vendor and standards documentation

OpenAI model, pricing, reasoning and tools documentation, fetched August 2026.

Anthropic platform documentation, model overview, release notes and pricing, fetched August 2026.

Google Gemini API changelog and pricing, fetched August 2026.

DeepSeek API pricing documentation, fetched August 2026.

xAI model documentation, fetched August 2026.

Mistral models overview, fetched August 2026.

Linux Foundation,

Formation of the Agentic AI Foundation

, December 2025.

Model Context Protocol, specification release 2026-07-28 and 2026 roadmap.

10.3 Regulation

Regulation (EU) 2026/1744, Digital Omnibus on AI. Published in the Official Journal 24 July 2026, in force 27 July 2026.

Note: implementation timelines published before 24 July 2026 show superseded dates.

10.4 Vstorm project evidence

Schmitt-Thompson Clinical Content — agentic retrieval for clinical triage guidelines. Zero hallucination events across 329+ nurse-validated scenarios.

Synera — text-to-workflow agent. Workflow generation from up to two hours to three minutes, with zero hallucinations in production through multi-layer validation.

Both published in full at vstorm.co/case-studies.



APPLIED AGENTIC AI • [VSTORM.CO](https://vstorm.co)