RESEARCH REPORT

The Last Mile: Why AI Projects Fail on the Way to Production

The demo worked. The pilot impressed everyone. Then the project quietly stalled. This report explains why, and what the successful minority do differently from day one.

SDTC Digital · What We Think

September 2026 · Approx. 5,000 words

PART 1

Content Metadata

Content type	Research Report
Title	The Last Mile: Why AI Projects Fail on the Way to Production
Slug	ai-projects-production-readiness
Meta title	Why AI Projects Fail Before Production | SDTC Digital
Meta description	Most AI projects die between the demo and daily use. Our research report explains why, what the failure statistics really measure, and how to build AI that ships.
Excerpt / card summary	The demo worked. The pilot impressed everyone. Then the project quietly stalled. This report examines why most AI projects never survive the journey to production, and what the successful minority do differently from day one.
Hero image direction	An editorial photograph of a relay race at the moment of a baton exchange, slightly blurred with motion, shot in muted tones. The report's core idea is that AI projects fail at the handover, from demo conditions to real operations. Avoid robots, brains, and glowing circuitry entirely.
Card image direction	A simple flat graphic of a road that is fully paved except for its final stretch, rendered in SDTC's brand palette. Clean and legible at thumbnail size beside the "Research Report" category label.
Primary audience	Founders, CTOs, CIOs, VPs of engineering, product leaders, SaaS operators, and enterprise leaders who have AI pilots running and need them to become dependable systems.
Search intent	Informational and evaluative: "why AI projects fail," "AI pilot to production," "AI production readiness," "MLOps," "AI project failure rate," "AI proof of concept to production."
Suggested related articles	1) The Adoption Gap: Why Enterprise AI Works in Some Companies and Stalls in Most (Research Report). 2) A Guide to running an AI pilot with owners, metrics, and decision rules. 3) A Field Note on AI data readiness assessments. 4) A Perspective on governing agentic AI. 5) A Case Note on taking an AI feature from prototype to production.

PART 2 · REPORT BODY

The Last Mile: Why AI Projects Fail on the Way to Production

Executive Summary

Ask how many AI projects fail and you will get a different number depending on who you ask. RAND Corporation says more than 80%, roughly twice the failure rate of ordinary IT projects. MIT's Project NANDA found that 95% of organisations saw no measurable financial return from their generative AI pilots. Gartner expects 30% of generative AI projects to be abandoned right after proof of concept. S&P Global found that 42% of companies abandoned most of their AI initiatives in a single year, up from 17% the year before.

> **CENTRAL FINDING**
>
> The numbers disagree. The diagnosis behind them does not: AI projects almost never fail because the technology doesn't work. They fail in the unglamorous stretch between a successful demo and a dependable system that real people use in real work every day.

Across every credible study reviewed for this report, and across the practitioner accounts that accompany them, the same finding repeats. Projects fail because the data underneath was never ready, because nobody owned the outcome, because success was never defined in numbers, because governance arrived as a surprise at the end, and because the pilot was quietly mistaken for a finished product.

That last point may be the single most expensive misunderstanding in enterprise technology today. A pilot is built to prove an idea can work under friendly conditions. A production system has to work under unfriendly ones: messy data, impatient users, security reviews, edge cases, and costs that scale with every transaction. They are not two stages of the same thing. They are two different things, and treating the first as evidence that the second is nearly done is how organisations end up in the failure statistics.

The encouraging news is just as consistent. The minority that succeed, roughly one in five projects by most counts, are not doing anything exotic. They write down the business outcome before they build. They fix their data before they train anything. They design for production from day one rather than after the demo. They put governance in early, name a permanent owner, and are willing to stop weak projects quickly. None of this requires a bigger budget. It requires a different order of operations. This report explains that order.

The Context

Enterprise AI spending has reached a scale that demands results. Gartner forecast worldwide spending on generative AI alone at $644 billion in 2025, a 76% jump on the prior year. McKinsey research found that 92% of companies plan to increase AI investment. Boards have approved budgets, teams have been hired, and pilots have multiplied across nearly every large organisation.

What has not multiplied is production value. The same McKinsey research found that only 1% of companies describe themselves as mature, meaning AI is fully woven into how they work and producing substantial results. IBM's 2025 CEO study reported that only 25% of AI initiatives delivered the return that was expected of them, and only 16% managed to scale across the whole enterprise.

The most striking numbers describe abandonment. S&P Global Market Intelligence found that 42% of companies abandoned most of their AI initiatives in 2025, a sharp rise from 17% a year earlier, and that close to half of AI proofs of concept were scrapped before they ever reached production. Gartner projects that 60% of AI projects that lack what it calls "AI-ready data" will be abandoned through 2026, and that 40% of agentic AI projects, the newer kind where AI carries out multi-step tasks on its own, will be cancelled by the end of 2027, mostly for governance reasons.

It is worth pausing on why the headline failure rates differ so much, because the differences are informative rather than contradictory. RAND's "more than 80% fail" is drawn from in-depth interviews with 65 experienced AI practitioners and describes projects that failed to deliver their intended value. MIT's "95%" applies a stricter test: a project only counts as successful if it produced sustained, documented profit-and-loss impact confirmed by both users and executives. Gartner's "30% abandoned after proof of concept" measures one specific drop-off point in the journey. Each number is a different camera angle on the same scene: a large majority of AI efforts, however you count them, are not becoming durable business systems.

One more piece of context matters. This is not a story about failed experiments in a niche technology. The same period produced clear evidence of what production AI can do when the journey is completed. To take one documented example from the World Economic Forum's 2025 reporting, a machine-learning system for hospital discharge planning, embedded properly into the daily operations of 12 hospitals, increased daily discharge rates by 4.6% in six months. The value is real. The road to it is what this report is about.

The Core Problem

Underneath every failure statistic sits one confusion, and it is worth stating in the plainest possible terms: **a demo that works is not a system that works.**

A proof of concept is built under conditions chosen to be friendly. The data has been hand-picked and cleaned. The users are a small, motivated group. The integrations with other systems are mocked up or skipped. The edge cases, the strange inputs, the angry customer, the half-filled form, are politely

excluded. None of this is dishonest; it is exactly what a proof of concept is for. It answers one question: could this idea work at all?

Production asks entirely different questions. Can the system handle live data that is messier, later, and stranger than anything in the test set? Does its output arrive inside the tools where people actually work, or in a separate window they must remember to visit? What happens when it is wrong, and who notices? Has security reviewed it? Will the costs still make sense at ten thousand transactions a day instead of a hundred? Who is responsible for it a year from now, after the team that built it has moved on?

The research is unambiguous that this gap, not model quality, is where projects die. RAND's interviews with practitioners identified five leading root causes of AI project failure: the wrong problem being solved because business leaders and technical teams misunderstood each other; data that was inadequate for the job; organisations chasing the latest technology instead of a real problem; missing infrastructure to manage data and deploy models; and, least often, problems genuinely too hard for AI. Notice what tops the list. In RAND's study, 84% of interviewees pointed to leadership-driven causes, misunderstood goals, unrealistic expectations, shifting priorities, as the primary reason projects fail. The technology is fifth on a list of five.

MIT's research reaches the same destination by a different route. Its "GenAI divide" is not between companies that use AI and companies that don't. It is between the roughly 5% whose deployments produce measurable financial results and the 95% whose pilots never translate into business impact. One chief operating officer quoted in the MIT research put the experience of that 95% bluntly:

> *"The hype on LinkedIn says everything has changed. Nothing fundamental has shifted."*

The core problem, then, is a mismatch of expectations and engineering. Organisations fund a demo, celebrate it, and then expect a production system to emerge from it with a launch decision. What actually stands between the two is a set of missing layers: ready data, real integrations, monitoring, governance, trained users, and a named owner. Building those layers is the majority of the work. Most project plans, and most budgets, treat them as an afterthought.

What Leaders Often Get Wrong

The sources reviewed for this report, spanning formal research and hands-on practitioner accounts, describe a remarkably consistent set of leadership mistakes.

Starting with the technology instead of a problem. The worst-performing projects begin with "we should do something with AI" and go hunting for a use case. The best begin with a specific, painful operational bottleneck and work backwards, sometimes concluding that AI isn't even the right tool. RAND's interviewees described being instructed to apply machine learning to problems that a few simple rules could have solved, projects that "succeed" technically and fail in every way that matters, because they were never necessary.

Declaring victory at the pilot. A pilot that performs well creates a dangerous kind of confidence. It proves the model can produce useful output under ideal conditions. It proves nothing about whether the organisation can operate it: feed it reliable data, integrate it, monitor it, retrain it, and get people to use it. Practitioner frameworks reviewed for this report are emphatic that the move from pilot to production is additional engineering, not a launch decision.

Never writing down what success means. Projects routinely run for a year without anyone able to say, in numbers, whether they are working. The fix costs almost nothing: a one-page statement, agreed before any building starts, of the business measure being improved, by how much, by when, and who owns it. Its absence is one of the most reliable predictors of failure in the entire evidence base. As one practitioner framework puts it: if you cannot state the success metric in one sentence with specific numbers, stop and define it first.

Underestimating the data work, every time. Data preparation consumes the majority of real AI project effort, commonly estimated at 60 to 80% (a widely used practitioner figure; verify before publication). Yet it is consistently the least planned and least funded part of the work. RAND's interviewees were vivid about this. One said: "80 percent of AI is the dirty work of data engineering. You need good people doing the dirty work, otherwise their mistakes poison the algorithms." Business leaders, another noted, believe their data is fine because they get weekly sales reports, without realising that data collected for reporting is often unusable for training an AI system.

Treating governance as the final step. Security, compliance, and legal reviews held until the end of a project regularly send teams back to redesign the architecture they have just finished. In regulated industries, involving risk teams at the start compresses approval from months to weeks; leaving them until the end can add half a year. Gartner's prediction that 40% of agentic AI projects will be cancelled by 2027 names governance failure as the leading cause.

Expecting immediacy. Leaders primed by vendor demos expect weeks where honest work takes months, and expect certainty from technology that is inherently probabilistic, meaning it deals in likelihoods, not guarantees, and will sometimes be wrong. When results are incremental instead of dramatic, faith collapses and priorities shift. RAND's interviewees described promising projects abandoned mid-flight because leadership attention had already moved on; as one put it, "often, models are delivered as 50 percent of what they could have been."

The Evidence and Patterns

The numbers, and what each one actually measures

The headline statistics are best read together, as measurements of different points along the same pipeline.

At the front of the pipeline, nearly everyone is building: 92% of companies plan to increase AI investment (McKinsey). In the middle, attrition is severe: roughly 30% of generative AI projects are abandoned immediately after proof of concept (Gartner), and close to half of all proofs of concept never

reach production (S&P Global). At the end, where value is counted, the funnel narrows brutally: only 25% of initiatives deliver expected returns (IBM's 2025 CEO study), and by MIT's strict standard, documented and sustained profit-and-loss impact, only about 5% of organisations qualify.

Two findings deserve special weight because of how they were produced. RAND's study is unusual in that it interviewed the people who build these systems, 65 practitioners with at least five years of experience, rather than surveying executives. Its conclusion that leadership and data problems dwarf technology problems therefore comes from the engine room, not the boardroom. And BCG's 2025 research across more than 1,250 companies quantified the gap between the disciplined and the rest: organisations with mature, production-oriented AI operating models report deployment success rates above 60%, while laggards report around 12%. Same technology, five times the success rate.

The failure pattern, stage by stage

Synthesising the practitioner frameworks in the source material, the journey from idea to production has predictable danger zones.

At the start, the trap is framing: the goal is "use AI" rather than "fix this measurable problem," and no feasibility check is done on whether the data exists or the value justifies the cost. In the middle, the trap is discovery: production data turns out to be messier, more scattered, and more restricted than the pilot's curated extract, and integration with the systems where work actually happens, the CRM, the ERP, the legacy platforms, absorbs far more engineering than anyone budgeted. Late in the journey, the trap is review: security and compliance teams see the system for the first time and find problems that require redesign. And after launch, the trap is silence: nobody was assigned to watch the system, so when its accuracy slowly degrades, a phenomenon called drift, where changing real-world data gradually makes a model stale, no one notices until the business feels the damage.

One more pattern deserves a name because it is the most expensive of all: the project that fails late. A project stopped at the planning stage costs little and teaches a lot; researchers and advisers reviewed for this report describe it as the healthiest kind of failure. A project that collapses after deployment has consumed budget, credibility, and organisational goodwill. Worst is the lingering project, delivering just enough to avoid being killed but never enough to justify its cost, quietly draining resources for years. The willingness to stop the wrong projects early is one of the clearest behavioural differences between the successful minority and everyone else.

Three cautionary tales from the public record

Three well-documented cases, all cited in the source material, show what these abstractions look like in practice.

Amazon built a machine-learning recruiting tool to identify strong engineering candidates. It performed well in testing. In production, it turned out to be systematically downgrading women's CVs, because it had been trained on ten years of historical hiring data that reflected a decade of bias. The model was accurate about the past; the past was the problem. Amazon scrapped it in 2017. The lesson: a model optimised on history is not automatically aligned with the outcome you actually want.

IBM's Watson for Oncology was trained on data and clinical practice from one leading American cancer centre. Deployed into hospitals in other countries, its recommendations skewed toward American methods and demographics, and clinicians lost confidence. IBM sold off Watson Health in 2022. The lesson: pilot data does not automatically generalise to the real world's variety.

In early 2024, Air Canada's customer-service chatbot confidently told a passenger about a refund policy that did not exist. When the passenger claimed the refund, the airline argued the chatbot was "a separate legal entity responsible for its own statements." A Canadian tribunal rejected that argument and held the airline fully liable. The lesson: there was no gap in the model, there was a gap in governance, no human oversight for high-stakes answers, and no mechanism to catch confident nonsense before a customer acted on it.

What the successful minority do differently

The most useful evidence is the consistency in how the roughly one-in-five successful projects operate. Across MIT's, RAND's, and BCG's findings and the practitioner playbooks, the same behaviours recur: success is defined in numbers before approval; data infrastructure is made ready before use cases are layered on it; the pilot is designed with a production path from day one; end users co-design the workflow rather than receiving a finished tool; feedback and monitoring are built in from launch; and there is a formal decision gate, commonly around 90 days, where the project is scaled, redirected, or stopped on evidence.

None of these behaviours require frontier technology. That is the pattern's most important property: the gap between the 5% and the 95% is a discipline gap, not a capability gap.

Strategic Implications

For enterprise leaders, the portfolio matters more than any single project. With most initiatives statistically likely to underdeliver, the strategic skill is not picking one perfect project; it is running a portfolio with honest gates, so weak projects die cheaply at week 12 rather than expensively at month 18, and strong ones get resources faster. The organisations that do this well treat an early kill not as failure but as the system working. Boards should ask not "how many AI projects do we have?" but "how many have we stopped, and how quickly?"

The first production deployment is worth far more than it appears. Practitioner evidence shows that the infrastructure, governance, and lessons from a first successful deployment, when deliberately captured, cut delivery time for subsequent projects dramatically; in one documented delivery case, by around 40%. The alternative is an organisation with several isolated AI successes and no compounding capability, where every new project starts from scratch. Treating deployment one as a platform investment, not a one-off, is among the highest-return decisions available.

For SaaS companies, production readiness is becoming the product. Enterprise buyers have been burned by pilots; procurement teams increasingly ask about monitoring, audit trails, rollback, and integration before they ask about model quality. SaaS products that arrive production-shaped, with

governance features, observability, and clean integration paths built in, will beat more impressive demos that leave the operational burden to the customer. The failure statistics of your buyers are your sales objection; answering them is your differentiation.

Readiness is also a market signal to talent and partners. Teams that ship AI to production learn faster than teams that pilot endlessly, and skilled engineers increasingly sort employers accordingly. Meanwhile, the divide BCG measures, 60% versus 12% deployment success, compounds quarterly: production systems improve on real feedback while sandboxed pilots do not. Falling behind on the production discipline means falling behind at an accelerating rate.

Expectations set strategy more than most leaders admit. IBM's finding that 50% of CEOs say recent investments have left them with disconnected, piecemeal technology is a warning about pace. Adding AI on top of fragmented systems creates another fragment. The strategic sequencing that the evidence supports is patient: foundation first, focused use case second, scale third. It feels slower. Measured by value delivered rather than pilots launched, it is far faster.

Technical Implications

Data readiness is the foundation, and it is a continuous discipline, not a clean-up. Gartner's definition of AI-ready data is a practical checklist: data mapped to the specific use case; ownership and quality standards at the level of individual data assets; pipelines that are automated and monitored rather than manually refreshed; metadata, meaning the information that describes what each dataset is and whether it can be trusted, kept current and machine-readable; and quality checked continuously, not audited annually. The detail that matters most: traditional data management works on reporting rhythms, quarterly and monthly. Production AI needs quality signals in hours. Most data teams are not staffed or tooled for that cadence yet.

Design the production architecture before the model is finished, not after. The decisions that determine whether a system can live in production, how it will be deployed and updated, what latency it must meet, how it scales, how a bad version gets rolled back, are architectural, and retrofitting them after a pilot routinely adds months. Three properties matter most: the model must be deployable independently of the applications around it, so improvements don't wait for release windows; it needs its own kind of monitoring, watching prediction confidence and input drift, which ordinary application monitoring does not cover; and downstream systems must be built to consume probabilistic output gracefully, including a defined path for low-confidence answers.

Operations for AI, often called MLOps or LLMOps, is the difference between one deployment and a capability. Stripped of jargon, it means running AI models with the same discipline as any critical system: knowing which version is live, testing quality continuously rather than once at launch, tracking changes to prompts and configurations, watching costs, and rehearsing rollback before it is needed. Without this layer, every deployment is a bespoke effort and every incident is an investigation in the dark. With it, the second and tenth deployments get cheaper and safer.

Governance is an engineering deliverable, not a review meeting. The systems that clear compliance quickly are the ones where documentation, audit trails, human-oversight workflows for high-stakes outputs, and access controls were built alongside the model. This is doubly true under tightening regulation, the EU AI Act's risk classifications, GDPR's rules on automated decisions, and frameworks like NIST's AI Risk Management Framework and ISO/IEC 42001, where the absence of an audit trail can block deployment regardless of model quality. The Air Canada case shows the same logic from the liability side: courts will treat the AI's statements as yours.

Cost behaves differently in production, so build the meter early. Inference, the computing cost of the model answering, scales with usage in ways pilot budgets routinely understate. The practical disciplines: visibility into cost per use case, routing routine requests to smaller, cheaper models where quality allows, spending thresholds that trigger review before overruns rather than after, and a standing question of cost per business outcome, not just total spend.

Keep experimentation and production separate, deliberately. Exploration should be fast and loose; production must be reproducible, tested, and auditable. The test worth applying: if the production model cannot be rebuilt exactly from versioned code and versioned data, it is not production-ready. Conflating the two modes, deploying from research notebooks, is one of the most common and most fixable failure patterns in the record.

Decision Framework

Before an AI project is approved, and again before it is promoted from pilot to production, five questions determine its fate. We think of them as the Production Test. Each has a plain pass condition.

1. Worth doing: is there a numbered outcome and a named owner?

Pass: a one-sentence statement, "improve [specific measure] by [number] by [date]", signed by a business owner accountable for that result, with today's baseline measured. Fail: the goal is "use AI," the metric is vague, or the only owner is the technical team. Evidence across every source says this single page predicts more than any technology choice.

2. Able to feed it: is the data ready for production, not just for the demo?

Pass: the data the system needs is mapped, accessible, legally usable, and representative of live conditions, and someone owns its quality. Fail: the pilot ran on a hand-cleaned extract; production data lives in disconnected systems; privacy or licensing questions are unanswered. If the honest answer is "we'll fix data as we go," expect the timeline to double.

3. Able to run it: does the plan cover the system, not just the model?

Pass: deployment method, integrations into the tools where work happens, monitoring, rollback, and retraining are designed and costed, and production economics have been modelled at realistic volume.

Fail: the plan ends at "train the model," integrations are assumed, or nobody has calculated cost per transaction at scale.

4. Allowed to run it: has governance been designed in?

Pass: security, compliance, and legal have reviewed the architecture at design time; high-stakes outputs have a human review path; decisions are logged for audit; regulatory classification is known. Fail: review is scheduled "before launch." For regulated work, this question alone decides months of timeline.

5. Wanted: will the people it affects actually use it?

Pass: the users who will live with the system helped design the workflow, the output arrives inside their existing tools without extra steps, and there is a clear way to escalate to a human and to flag errors. Fail: users will "be trained at rollout." Systems that fail this question get built, deployed, and quietly routed around.

Two rules make the framework work in practice. First, sequence matters: the questions are ordered as dependencies, and weakness in an early one undermines everything after it; fixing data after architecture, or governance after build, is where the rework lives. Second, the gate must have teeth: set a formal checkpoint, around 90 days into any pilot, where the project is scaled, redirected, or stopped against the metric from question one. A project that cannot face that gate is not a project; it is a hope.

Risks and Failure Modes

The lingering pilot. The most common failure is not a crash but a drift: a pilot that shows promise, never faces a decision gate, and consumes budget indefinitely. Prevention is structural, not motivational: no pilot without a pre-agreed metric, date, and decision rule.

The silent degrade. AI systems can decay without any code changing, as the world drifts away from their training data. Without monitoring for accuracy and input drift, the first sign of trouble is business damage, wrong answers at scale, discovered by customers. Any system without observability and a tested rollback is carrying this risk today.

The late veto. Projects that meet security, legal, or compliance for the first time at the end reliably lose months, and sometimes die there, after the engineering money is spent. In regulated sectors this is the single largest schedule risk. The mitigation is simply earlier: involve reviewers at architecture time.

The right model for the wrong metric. A churn model with 92% accuracy that achieves it by predicting "no churn" for nearly everyone is technically excellent and operationally useless. The risk is optimising what is easy to measure instead of what the business needs, and it is invisible until deployment. The defence is defining the business metric first and deriving the technical target from it.

The trust collapse. One high-profile wrong answer, an Air Canada moment, can end adoption overnight, especially where outputs face customers. Confidence is protected by designing human

oversight for high-stakes outputs, being honest with users about limitations, and giving them a fast way to flag errors, which also makes the system improve.

The cost surprise. Systems affordable at pilot volume can be ruinous at production volume. When the first oversized bill reaches the CFO without warning, programmes get cancelled regardless of merit. Cost modelling at realistic volume belongs in the approval decision, and cost monitoring belongs in the launch checklist.

The one-off trap. A successful deployment whose pipelines, governance, and lessons are not captured for reuse leaves the organisation exactly where it started, facing the next project from scratch. The risk is subtle because it looks like success. The defence is treating reusable infrastructure and documented lessons as deliverables of every project, not by-products.

And the risk of over-caution. The evidence cuts both ways: teams that wait for perfect conditions never ship, and a system kept in a sandbox forever generates no feedback and no value. The healthy posture is a small, well-instrumented first production release in a contained workflow, imperfect but real, improved on live evidence.

Recommendations

1. Refuse to start without the sentence. One page before any build: the measure, the number, the date, the named business owner, the baseline. If the sentence cannot be written, the project is not ready to exist. This is the cheapest risk control in this entire report.

2. Audit the data before you commit the budget. A two-to-four-week data readiness assessment, scoped to the specific use case, surfaces most fatal problems while they cost thousands rather than hundreds of thousands. Budget the data engineering honestly; it is most of the work.

3. Design for production from day one. Give every pilot a production path before it starts: how it would deploy, integrate, be monitored, and be rolled back. A pilot built as a throwaway will be thrown away, along with its budget.

4. Bring governance in at the architecture stage. Invite security, compliance, and legal to the design table, not the launch review. Build documentation, audit trails, and human-oversight paths as engineering deliverables. In regulated industries, this is the difference between six weeks and six months.

5. Name the permanent owner before launch. Someone must own the system's business outcome, data quality, performance, and incidents after the project team disbands, with the same seriousness as any core enterprise application. Launch without this and degradation is a matter of time.

6. Instrument everything, and rehearse the rollback. Monitoring for accuracy, drift, latency, cost, and usage from day one; alerts with named recipients; a rollback procedure tested before it is needed. Treat launch as the beginning of operations, not the end of the project.

7. Run the 90-day gate without sentiment. Scale, redirect, or stop, on evidence, against the original metric. Celebrate early stops publicly; they are the portfolio working as designed. The behavioural gap between the successful minority and the rest is largely this discipline.

8. Make the first deployment pay for the next ten. Capture the pipelines, governance templates, monitoring setup, and lessons as reusable assets, whether or not you formalise it as a centre of excellence. The goal is compounding capability: the tenth project should start with the accumulated advantage of the nine before it.

How SDTC Digital Thinks About This

SDTC Digital is a product engineering company, and this report's subject is, candidly, the reason our discipline exists. We have seen the pattern from both sides: the pilot that impressed everyone and then sat for a year, and the unglamorous production system that quietly changed how a business runs.

Our conviction, formed in delivery rather than theory, is that production readiness is not a phase at the end of an AI project. It is a property of how the project is run from the first week: the outcome written down before the code, the data audited before the model, the security review at the whiteboard rather than the launch meeting, the monitoring built before it is needed. When we take on AI work, the model is usually the smallest part of what we build. The larger part is the system around it, pipelines, integrations, oversight, observability, and the operating habits of the team that will own it, because that is where the evidence says success is decided.

We hold one more view that the research supports: the right first project is rarely the most ambitious one. It is the contained, measurable workflow where value can be proven in a quarter and the foundations laid can be reused by everything that follows. Production-grade delivery is not the cautious path. It is the fast one, measured in value rather than announcements.

Conclusion

The AI failure statistics look like an indictment of the technology. Read closely, they are almost the opposite: study after study finds the models work, and the failures concentrate in the human systems around them, unclear goals, unready data, absent owners, late governance, and pilots mistaken for products.

That is unusually good news, because every one of those causes is within a leadership team's control. No breakthrough is required to write down a success metric, audit the data first, design for production from day one, or stop a weak project at 90 days. The organisations in the successful minority are not smarter or richer. They do ordinary things in the right order, and they do them every time.

The demo was never the hard part. The last mile is, and it is won before the journey starts.

Sources referenced in this report include RAND Corporation's "The Root Causes of Failure for Artificial Intelligence Projects" (2024, based on 65 practitioner interviews); MIT Project NANDA's "The GenAI Divide: State of AI in Business 2025"; S&P Global Market Intelligence's Voice of the Enterprise survey (2025); Gartner forecasts and predictions on generative AI spending, AI-ready data, and agentic AI (2025); BCG's "Widening AI Value Gap" research (2025, 1,250+ companies); IBM's 2025 CEO study; McKinsey's State of AI research; the World Economic Forum's 2025 reporting on production AI outcomes; and practitioner analyses from ACL Digital, Pertama Partners, Coderio, Golabs, SR Analytics, Twendee, Exponential Tech, and HST Solutions. Figures cited from secondary sources, including the 60–80% data preparation estimate and vendor-reported case outcomes, should be verified against primary sources before publication.