

In the fast-paced race toward AI adoption, many enterprises encounter a perplexing phenomenon: AI pilots that shine brightly in demos but stumble during real-world deployment. Companies like Garrett Motion, PMI, and Pandora have all reported promising early results with tools such as Large Language Models (LLMs) and retrieval-augmented generation (RAG). Yet, despite stellar model performance metrics, these pilots often fail to drive meaningful bottom line impact. o9solutions.com What's going on behind the scenes?

The Bottleneck: Trust, Not Accuracy

When we talk about **AI adoption failure**, the issue is rarely the AI model's raw accuracy. Instead, the bottleneck is *planner trust* — the decision-makers who must rely on AI recommendations for high-stakes business decisions. These planners rarely accept "the model decided" as a sufficient explanation. This mistrust persists even when accuracy gains in demos appear significant.

Trust is multi-dimensional. It hinges on explainability, consistency, and alignment with existing business logic, not just numerical performance. Garrett Motion's experience with AI pilots illustrates this well. Although their predictive maintenance models built on LLMs performed admirably in trials, operations managers hesitated to act on recommendations without clear reasoning paths or audit trails.

What Would I Show in an Audit?

This question defines planner trust. Decision-makers want to see:

- Where the data came from
- How the AI arrived at a specific recommendation
- Which rules or business logic the model applied
- Consistency across similar inputs

Without satisfying this checklist, AI pilots risk becoming opaque black boxes—an Achilles' heel in highly regulated or risk-sensitive industries.

Explainability: The Achilles' Heel of Many AI Pilots

Explainability is no longer optional for enterprise AI; it's a prerequisite. PMI, a company with rigorous compliance standards, struggled to scale its AI-driven supply chain optimizations for this very reason. Although their LLM-powered RAG systems generated accurate insights, planners rejected outputs when explanations were vague or inconsistent.



Underlying this challenge are two common issues:

1. **Hallucinations:** LLMs sometimes produce fabricated facts with undue certainty, undermining credibility.
2. **Random Output Variance:** Given the same prompt, responses may vary, making it difficult to anticipate AI behavior in production.

Both issues erode the rigor enterprises expect and fuel skepticism about AI recommendations.

Hallucinations and Fabricated Certainty

Hallucinations aren't just minor nuisances—they pose real risks. Pandora, operating in a highly competitive media environment, ran pilots using RAG to generate personalized customer engagement plans. Their teams found that when hallucinations occurred, the AI confidently presented incorrect data, leading to costly missteps during review stages.

Fabricated certainty—the AI's unwavering tone even when output is incorrect—makes it difficult for planners to spot errors without deep domain expertise. This false confidence makes the AI appear unreliable or even dangerous.

Consistency and Alignment with Business Logic

Accuracy alone doesn't guarantee operational trust. Consistency matters immensely. Planners need predictable AI behavior to integrate recommendations into workflows seamlessly.

Alignment with existing business logic is another non-negotiable. If AI advice contradicts proven heuristics or known constraints, users reject it as nonsense. Many pilots fail because developers optimize for accuracy metrics detached from real-world decision rules.

Garrett Motion's pilot faded when the AI ignored maintenance schedules or replaced engineer know-how with black-box suggestions. The disconnect forced teams to double-check every recommendation, nullifying time-savings benefits and sowing doubt in the AI's value.

Why Accuracy Gains Are Not Enough

High accuracy in a controlled demo or test environment doesn't always translate into trust in the field. This gap explains why promising AI models don't guarantee bottom line impact.

- **Demo Conditions Are Artificial:** Metrics often come from curated datasets, not real-time, noisy enterprise data.
- **No Clear Decision Traceability:** Success is measured by output correctness alone, ignoring explainability.
- **Stakeholder Alignment Is Missing:** Users aren't involved early enough to calibrate AI outputs to business needs.
- **Changing Success Metrics:** After go-live, stakeholders often shift what "success" looks like, invalidating pilot results.

Without addressing these gaps head-on, AI projects move from promising pilots to stalled proofs of concept.

Lessons from the Field

Company	Use Case	AI Tools	Key Challenge	Outcome
Garrett Motion	Predictive Maintenance	LLMs, RAG	Lack of explainability, inconsistent outputs	Pilot stalled due to planner mistrust
PMI	Supply Chain Optimization	LLMs, RAG	Opaque reasoning and hallucinations	Rejected recommendations, no adoption
Pandora	Customer Engagement	RAG	Fabricated certainty led to costly errors	Recalibrated AI with human-in-the-loop review

How to Move Beyond Pilot Failure

To overcome AI adoption failure, enterprises need a multi-pronged strategy focused on trust-first principles:

1. **Visibility & Auditability:** Build systems that show decision traceability clearly. For every recommendation, show data sources, rule checks, and reasoning paths.
2. **Human-in-the-Loop:** Use human review as a core component—not bolt-on—to mitigate hallucinations and calibrate AI outputs.
3. **Align Metrics Early:** Co-design success criteria with stakeholders, making explainability and consistency first-class goals alongside accuracy.
4. **Control Randomness:** Stabilize model outputs for the same prompt using deterministic techniques or ensemble methods.
5. **Educate Users:** Train planners to understand AI limitations and nuances, building informed trust rather than blind faith.

Conclusion: Trust Is the True ROI of AI Pilots

AI pilots often fail not because the model looks bad, but because the enterprise lacks trust in the solution. Despite exciting advances exemplified by LLMs and RAG, companies like Garrett Motion, PMI, and Pandora reveal the complexities that lie beneath surface accuracy.

Achieving meaningful AI adoption requires explainability, consistency, and rigorous business-logic alignment. Without trust, metrics alone cannot translate into decisive action or bottom line impact. Addressing these root causes head-on transforms AI from a shiny demo into a trusted enterprise partner.

