WHY AI PROJECTS FAIL

The POC-to-Production Gap

87% of AI projects never make it to production. The demo worked. The pilot impressed stakeholders. Then nothing. Here's why - and how to fix it.

The Pattern

Teams build impressive demos in weeks. Business cases get approved. Then months pass. Budgets drain. The POC that worked perfectly fails at scale. The project quietly dies in "production planning" forever.

87%	6-12x	$2.4M	5
Never reach production	Prod cost vs POC	Avg wasted investment	Hidden gaps

01 - THE FIVE GAPS

What kills AI projects between demo and deployment

POCs work because they're simple. Production breaks because it isn't. Five gaps consistently predict failure:

Gap 1: Cost

POC costs $5K. Production costs $500K. The math that worked with 100 test queries breaks at 10M real queries. Nobody modeled the inference economics at scale.

Gap 2: Latency

Demo took 3 seconds - "acceptable for a demo." Production needs 200ms or users leave. The architecture wasn't designed for real-time. Now it needs a rewrite.

Gap 3: Reliability

Hallucinations were "interesting edge cases" in testing. In production they're liability. The model worked 95% of the time. But 5% failure at scale means hundreds of incidents daily.

Gap 4: Observability

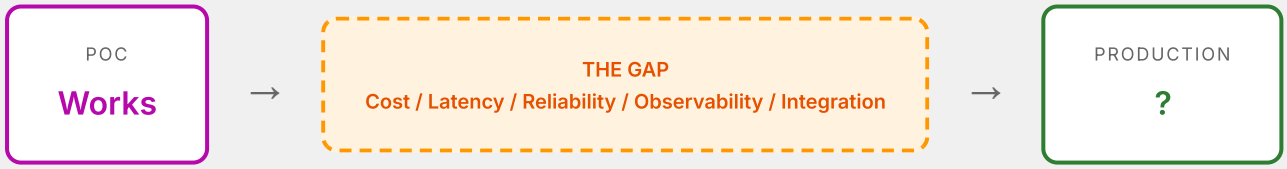
POC had no monitoring - it was a demo. Production needs real-time visibility. When things break at 2am, nobody knows until customers complain.

Gap 5: Integration

POCs run in isolation. Production connects to everything: auth systems, data pipelines, legacy databases, compliance workflows. Integration is 60% of production effort - and 0% of POC effort.

The most common gap. Teams underestimate integration by 4-8x.

The POC-to-Production Gap



02 - ANTI-PATTERNS

What teams do wrong

These patterns seem reasonable during POC but create massive technical debt:

1. **Monolithic Prompts**

   One giant prompt that does everything. Works for demos. In production: expensive, slow, unpredictable. Change one thing, break everything else.

2. **Retrieval Firehose**

   Stuff maximum context into every query. 10K tokens per request. Works in POC with 100 queries. At scale: $50K/month just in context tokens.

3. **Retry Spirals**

   Failure? Just retry. Keep retrying. In production: one bad input triggers infinite retries. Costs explode. Rate limits hit. Everything fails.

4. **Context Amnesia**

   Every request starts fresh. No memory across sessions. Users repeat themselves. Efficiency tanks. The "smart" AI feels dumb.

5. **One Model Fits All**

   GPT-4 for everything. Even simple classification. Even when rules would work. Expensive, slow, and often worse than simpler solutions.

The Root Cause

POC success metrics are wrong. Teams optimize for "did it work?" instead of "will it work at scale, reliably, within budget?" POCs should be production rehearsals, not demos.

03 - BRIDGING THE GAP

How to succeed where 87% fail

Teams that reach production do things differently from the start:

Cost-First Architecture

Model cost per query from day one. Use Trust Cascade to route simple requests to cheap solutions. Only use expensive models when needed. Budget visibility before POC approval.

Production Latency in POC

Set production latency targets before building. Design architecture to meet them. If demo can't hit targets, production won't either. Fail fast.

Reliability from Day One

Add monitoring in POC. Track hallucinations. Measure confidence. Build guardrails before deployment. Production reliability is designed in, not bolted on.

Integration Rehearsal

POC connects to real systems (or realistic mocks). Authentication, data pipelines, compliance hooks. Discover integration problems before production planning.

The Rotascale Approach

Every Rotascale engagement starts with production requirements, not demo goals. We model costs, latency, and reliability before building. Governance infrastructure goes in first, not last.

Our clients don't have POCs that work and production that doesn't. They have POCs that prove production will work.

Stop building demos that never ship

Get a production readiness assessment for your AI initiatives.

Get Assessment

View Services