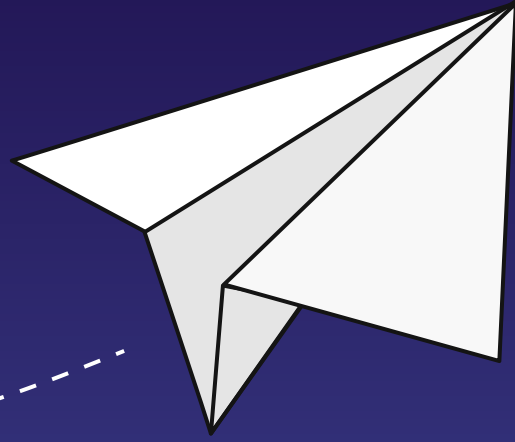




# From POC To Production: Why Enterprise AI Struggles To Scale

It's a critical emerging problem. But why, at the point of production does your AI Proof of Concept (POC) struggle to scale? The answer is simple:

**AI capability is no longer the bottleneck...  
production discipline is.**

# The Reality Leaders Are Facing

## **Data readiness**

Real-world data is unpredictable and creates data drift outside of the controlled pilot environment that decreases model accuracy and hampers performance.

## **Governance & security**

The evaluation frameworks, human fallbacks and guardrails that are often absent during POC become essential in production to trust and monitor model behaviour at scale.

## **Cost visibility**

AI costs compound quickly and unpredictably in production, challenging technical decisions and making the necessary unit economics difficult to achieve.

## **Ownership**

Organisational ambiguity over who 'owns' the AI is common and creates a fragmentation of accountability that can delay critical decision-making and stall scaling for months.

## **Adoption**

A lack of AI maturity creates a 'funding paradox' where stakeholders demand innovation but struggle to grasp the centralised expertise and complex costing models required for success.

While many organisations might be able to build a credible AI POC in as little as 4-8 weeks, insight from senior tech leaders shows that far fewer are making the leap to productionising this work in a way that scales. And of course, because moving an AI project from proof-of-concept to production is not just more of the same but a fundamental shift into a different (and high-stakes) operating environment.

POCs are designed to optimise for speed and discovery within a controlled sandbox; a production system must prioritise reliability, security and rigorous governance. The bottleneck then is not the model's capability, but failures in the surrounding infrastructure aligned to one or more of five key 'production gates'.

# The Five Production Gates

1

## Measurable Production Outcomes

POCs prove possibility; production demands accountability.

- Identify a clear owner for measurable success
- Define specific business KPIs beyond technical feasibility
- Establish clear success thresholds to trigger expansion

2

## Real-World Integration & Performance

Latency compounds and models fail.

- Build architecture resilient to API limits and latency
- Design for real workflows... and real restraints
- Manage risk around model providers changing behaviours or terms

3

## Evaluation, Safety & Control

Because manual testing simply won't scale.

- Implement automated real-time guardrails
- Monitor proactively for data and behavior drift
- Re-test outputs when models or prompts change

4

## Cost, Value & Model Strategy

Value must be optimised, but relative to cost.

- Shift from abstract POC costs to ROI-focused scaling
- Select "good enough" models to balance performance and price
- Continuously optimise strategy as the model landscape evolves

5

## Operating Model & Ownership

Without explicit ownership, scaling slows.

- Resolve ambiguity over who owns production AI
- Formalise governance for monitoring and model upgrades
- Define response protocols for when live systems fail

“The sheer technical and organisational complexity of scaling AI can lead to unrealistic expectations or even paralysis from boards and senior stakeholders, with engineering leads often struggling to explain why a successful POC cannot be turned on immediately. CTOs must reframe AI from a ‘science project’ to a hardened piece of software that requires its own lifecycle management and can only be deployed accordingly.”



**Andy**  
CEO, Deazy

# What We're Seeing in Practice

Time and again we hear leaders referencing the same five challenges when it comes to scaling AI POCs, in fact it's striking how predictable the blockers are. The answer? Robust systems, rather than better prompts, and remembering these four things:

## Lesson 1

### Build the business case... and fast

Ambiguity about who owns the live system is a key blocker in scaling AI POCs (or even getting them off the ground). For success, CTOs must be prepared to quickly move from implicit governance to explicit accountability, establishing clear ownership, business outcomes and cost frameworks alongside robust incident response protocols and an ongoing plan for cost-to-value optimisation.

## Lesson 2

### Think systems, not capability

Moving from POC to production changes the nature of the system, not just the scale. So where POC focuses on speed and demonstration of value, production demands a fundamental shift in architecture and operating model. CTOs must treat production as an entirely different operating system: one that prioritises trust, governance and cost control instead of experimental potential.

## Lesson 3

### Move beyond manual testing

Manual testing fails to scale once AI hits the real world. Production environments require automated evaluation frameworks and guardrails that monitor for data drift, safety breaches and quality degradation - plus the necessary observability to trace problems when they occur. Without this, the risk of 'garbage output' increases as the system expands, eroding trust and damaging reputation.

## Lesson 4

### And remember...

While a POC might use the most powerful model available regardless of price, a production environment requires a constant trade-off between performance and cost. We recommend against prompt model coupling - model providers frequently update their systems so one prompt and one model does not mean the same prompt and same model will continue to work.

"The ease of building an AI POC is deceptive. For example, not only is manually monitoring outputs impossible at scale, but the risks of hallucinations or data drift are significant enough to threaten compliance and - particularly with agentic AI - can exponentially escalate costs out of control.

At Aveni we're finding that, even five years in, the last 20% of the work takes 80% of the time. So be prepared. Have evaluations to test your outputs and use a large enough data set to assure confidence in making changes without introducing risk. It takes significant time and effort but is essential to pass the five gates and successfully bridge this POC to production gap."



**Marko**  
CTO, Aveni



Deazy enables ambitious organisations to explore and harness AI to drive digital product innovation and operational efficiency, applying our award-winning AI and software delivery expertise to solve complex challenges, accelerate innovation and build resilient digital platforms that scale.

With a uniquely flexible delivery model, we provide rapid access to a diverse pool of 6,000+ experienced nearshore AI, software, and data professionals, managed by highly-experienced and multidisciplinary in-house product and delivery experts who provide the support and resources to guarantee success.

**For support with your AI development challenges, or if you'd like to explore where you are across the five production gates, drop us a line at [andy@deazy.com](mailto:andy@deazy.com).**

**We'd love to carry on the conversation.**

