

Your Demo Is Crushing It. Your Production System Isn't.

The gap between "works in a demo" and "works at scale" is an engineering discipline, not a mystery.

SWIPE →

Six Disciplines That Separate Demo From Production

01 Context management and memory architecture

02 Evaluation beyond "did it get the right answer"

03 Observability, cost, and versioned rollout

04 Governance and human accountability

91%

Lower Latency From Context Tiering Alone

The MemO system paired that with a 90% reduction in token cost and 26% higher correctness, all from intelligently managing what the model actually sees.

Source: MemO production benchmarks

87%

Potential Cost Reduction From Model Cascading

Route "what's our office address" to a small model. Escalate "analyse this contract" to a large one. Right-sized resources for each task.

Source: model cascading cost analyses

74%

Of production AI agents still rely on human evaluation for edge cases. Automated metrics can't catch everything — especially sensitive decisions.

Source: production AI agent evaluation surveys



It's not the AI model's fault. It's the operational architecture around it — or the lack thereof.

Terence Kok

Enterprise AI Strategist

Cooking Dinner vs. Running a Restaurant

PROTOTYPE AI

You know what everyone likes

Adjust on the fly

Fix problems as they appear

PRODUCTION AI

Systems, process, quality control

Handles thousands of unknown cases

Doesn't fall apart under load

Great Demos That **Never Ship.**

Many organisations get stuck between an impressive pilot and an operational system: siloed data that doesn't talk to itself, unclear governance, and organisational inertia across departments. Breaking through takes data integration infrastructure, clear accountability for outcomes, and cross-functional teams spanning engineering, operations, policy, and domain experts — not a better demo.

How Production Systems Remember

Ephemeral (task scratch space) → Session (this conversation) → Persistent (long-term, governed)

The full production-readiness checklist:

terencekok.com/blog/why-your-ai-demo-works-great-but-production-system

Build for Demos, or Build for Production. Don't Confuse the Two.

Disciplined engineering achieves 90%+ reductions in latency and cost while improving quality. It's not magic — it's systematic practice from day one.



Terence Kok · Enterprise AI Strategist

SAVE · COMMENT · FOLLOW