neu•eon

Strategic Insights for Technology Leaders

From Prototype to Production: Scaling AI Systems in the Enterprise

Prototype

Innovation begins with focused experimentation and proof of concept

Pilot

Controlled deployment validates feasibility and surfaces integration challenges

Production

Enterprise-scale systems deliver measurable business value and competitive advantage

---

Scott Weiner
Chief Technology Officer & AI Lead
NeuEon, Inc.
sweiner@neueon.com

www.neueon.com

www.linkedin.com/in/sweiner

> *"High-performing technology organizations are 2.5x more likely to meet or exceed their organizational performance goals."*
>
> — DORA State of DevOps Report

Enterprise AI Reality: The Amplification Effect

Why 42% of Companies Abandoned AI Initiatives in 2025

42%

of companies abandoned most AI initiatives in 2025 (up from 17% in 2024)

Source: S&P Global Market Intelligence (1,006 enterprises)

46%

of POCs scrapped before reaching production

Top 3 Organizational Barriers

43%

Data Quality & Readiness

Organizations aren't ready for AI—data infrastructure and quality remain the biggest blocker

43%

Technical Maturity

Processes and infrastructure need significant maturity before AI can scale effectively

35%

Skills Shortage

Training gaps and talent shortages limit organizational capability to implement and maintain AI

Source: Informatica CDO Insights 2025 (600 CDOs)

> **"AI doesn't fix a team. It amplifies what's already there."**
>
> **— Google Cloud DORA 2025 Report**

Weak Foundations

Problems Amplified at Scale

- Weak testing → More bugs in production
- Poor data quality → Worse AI accuracy
- Unclear requirements → Wrong outputs faster
- Skills gaps → Complexity overwhelm

Strong Foundations

Success Accelerated at Scale

- Strong testing → Better coverage automatically
- Clean data → Higher model accuracy
- Clear requirements → Precise outputs
- Strong skills → Capability multiplied

Research Sources: S&P Global Market Intelligence 2025, Informatica CDO Insights 2025, MIT/MLQ.ai 2025, DORA 2025

Four Components of AI at Scale

The Technical Reality: Easy to Prototype, Hard to Scale

RAG Evolution

Hallucination Management

PILOT: Basic chunking, Simple vector search, Fast responses

PRODUCTION: Graph RAG/reranking, Latency explosion, Accuracy degradation

DEMO: Live RAG pipeline ✓

Model Cost

Variance & Management

PILOT: One model, Fixed costs, Predictable

PRODUCTION: 20x cost variance, Model drift, Dynamic routing needed

DEMO: Cost comparison ✓

AI Governance

Security & Compliance

PILOT: Trust the model, Manual review, No formal policies

PRODUCTION: Prompt injection risks, PII exposure, 45% need human approval

DEMO: Security checks ✓

Multi-Agent

Coordination & Communication

PILOT: Single agent, Simple workflow, Easy to debug

PRODUCTION: Agent orchestration, A2A communication, Exponential complexity

DEMO: Agent swarm ✓

THE PATTERN: EASY TO PROTOTYPE, HARD TO SCALE - Each component works beautifully in pilots. Each becomes complex at production scale. Each requires organizational readiness, not just technology.

Component 1: RAG Evolution & Scaling Challenges

From Basic Chunking to Agentic RAG—and What Breaks at Scale

BASIC RAG (Simple)	SPARSE RETRIEVAL (Better)	RERANKING (Slower)
CONTEXTUAL EMBEDDINGS	KNOWLEDGE GRAPHS	AGENTIC RAG (Complex)

Organizations get sophisticated to improve accuracy

What Breaks at Scale

1. ACCURACY DEGRADATION

- 12% precision drop at 100,000-page scale
- Wrong chunks retrieved = confident wrong answers
- Sophistication helps but doesn't eliminate this

Source: ChromaDB/Weaviate Research 2024-2025

2. LATENCY EXPLOSION

- Pilot: <1 second response (beautiful UX)
- Production: 8-15 seconds (user frustration)
- Each sophistication layer adds overhead

Question: What's your latency tolerance?

3. COST & COMPLEXITY

- Contextual embeddings = LLM call per chunk (expensive)
- Knowledge graphs need constant maintenance
- Agentic RAG has unpredictable retrieval costs

More sophistication ≠ lower costs

Decision Framework

Questions for CIOs:

- What accuracy degradation is acceptable?
- What's your latency tolerance?
- Build vs managed? (Pinecone, Weaviate, pgvector, Databricks)
- What's your sophistication budget?

Live Demo: Watch RAG pipeline in action

Component 2: Model Variance & Cost Management

The 20x Problem

Model Pricing Table (October 2025)

Model	Cost per 1M	vs GPT-3.5 Baseline
GPT-3.5 Turbo	~$1.00	BASELINE (1x)
GPT-4	$30.00 (input)	🔴 30x MORE EXPENSIVE
Claude 4 Opus	$15.00 (input)	🟠 15x MORE EXPENSIVE
OpenAI o1	$15-60 (varies)	🟠 15-60x MORE

20x cost multiplier

between cheapest and most expensive models

THE PILOT

- ✓ 10 users in testing
- ✓ "Let's use the best model!"
- ✓ Cost: ~$50/month (negligible)
- ✓ Leadership excited

WHAT BREAKS AT SCALE

1. MODEL DRIFT

- Providers update models without notice
- Response patterns change over time
- What worked in testing breaks in production
- No version control for external models

2. COST EXPLOSION

- 10 users: $50/month (negligible)
- 10,000 users: $50K - $500K/month
- No cost controls = budget surprises
- Six-figure bills organizations didn't anticipate

3. INCONSISTENT RESPONSES

- Same prompt, different models = different answers
- Quality variance across model tiers
- Compliance challenges with inconsistent outputs
- User confusion: "Why different answers?"

Cost Control Strategies

- Intelligent routing (cheap for simple, expensive for complex)
- Rate limiting per user/department
- Budget alerts and quotas
- Usage analytics and optimization

Live Demo: Model cost comparison + real-time dashboard

AI Governance, Security & Compliance

Governance Gap

83%

using AI systems

31%

have comprehensive AI governance policies

52 PERCENTAGE POINT GAP

Source: ISACA 2024-2025

The Paradox

54%

say governance is a priority

38%

say governance is the biggest barrier

The Paradox: Everyone knows governance matters, but it's seen as blocking progress

⚠️ **OWASP Top 10 for LLMs - #1 Vulnerability**: PROMPT INJECTION

Direct injection: Malicious user inputs

Indirect injection: RAG poisoning through documents

System prompt leakage revealing business logic

SHADOW AI IS INEVITABLE

75%

of employees will use unsanctioned AI by 2027 (up from 41% in 2022)

Your employees are already using:

ChatGPT • Claude • Gemini • GitHub Copilot

Copying corporate data into external services

No governance • No control • No audit trail

Source: Gartner Research

Five Guardrails Framework

SECURITY RAILS

- Block prompt injection
- Data exfiltration
- Unauthorized access

SAFETY RAILS

- Prevent harmful outputs
- Toxic content
- Dangerous advice

QUALITY RAILS

- Accuracy thresholds
- Validation checks
- Confidence scores

COMPLIANCE RAILS

- GDPR, HIPAA, SOX
- PII detection
- Audit logging

TOPICAL RAILS

- Keep AI on approved domains
- Prevent scope creep

Live Demo: Prompt injection detection & PII flagging in action

THE QUESTION ISN'T WHETHER EMPLOYEES USE AI. THE QUESTION IS: DO YOU HAVE VISIBILITY & CONTROL?

Component 3: Agentic AI with Tools

Autonomous Operations & the 45% Oversight Requirement

TRADITIONAL AI

Capability:

"Answers questions"

Scope: Read-only, passive

Risk: Low

AGENTIC AI WITH TOOLS

Capabilities:

Query databases

Send emails

Modify records

Execute transactions

Chain multiple actions

Iterative reasoning

Scope: Write-access, active

Risk: Variable (LOW → CRITICAL)

Risk-Based Approval Framework

TIER 1 - LOW RISK

Decision: AUTO-APPROVE

Examples: Read documentation, Generate report, Search catalog

Action: Execute immediately, log interaction

TIER 2 - MEDIUM RISK

Decision: AUTO-APPROVE + MONITOR

Examples: Update preferences, Send notification, Create draft

Action: Execute with enhanced logging, pattern monitoring

TIER 3 - HIGH RISK

Decision: HUMAN APPROVAL REQUIRED

Examples: Delete data, Process refund >$1000, Grant permissions

Action: Email approval workflow, <2 hour SLA

⭐ **45% of decisions land here (Forrester)**

TIER 4 - CRITICAL RISK

Decision: ESCALATE + EXECUTIVE APPROVAL

Examples: Bulk deletion, Large transfers, System changes, External API calls

Action: Block until executive confirms, <15 min response needed

45% of enterprise AI decisions require human approval

— Forrester Research 2024

52% of enterprises have AI agents in production (Google Cloud 2025)
86% projected adoption by 2027 (Gartner)

When Agents Go Wrong

Tool Misuse & Privilege Escalation

Agent queries production database instead of test environment. Agent deletes files it should only read. Agent accesses data the current user shouldn't see. **One misconfigured permission equals security breach.**

Unpredictable Cost Spirals

You expect 5 API calls to complete a task. Agent makes 50. Runaway loops execute thousands of unnecessary operations because task definitions were ambiguous. Cost explosions happen overnight with no prediction mechanism.

Human Oversight Requirements

Industry research shows **45% of enterprise AI decisions require human approval**—especially in high-stakes scenarios. Medical diagnosis needs clinical oversight. Financial trading requires regulatory compliance review. Hiring decisions demand human review to address bias concerns.

Multi-Agent Coordination Failures

Agent A locks a resource, Agent B needs it—deadlock. Agents enter loops consuming resources indefinitely. One agent fails, entire dependent chain stops. Debugging cross-agent issues is exponentially more complex than single-system failures.

Source: Enterprise AI Implementation Research 2025

Component 4: Multi-Agent Communication

Distributed Intelligence & the Orchestration Challenge

SINGLE AGENT

Capability:

"One agent, one task"

Scope: Limited by single model's capabilities

Example: Answer customer question

✓ Simple, predictable

□ Can't handle complex multi-step workflows

MULTI-AGENT SYSTEM

Capability:

"Multiple agents, distributed intelligence"

Scope: Specialized agents collaborate

Example: Research → Analyze → Draft → Review → Execute

✓ Handles complex workflows, parallel processing

□ Coordination complexity, emergent behavior

Agent Roles & Communication Flow

ORCHESTRATOR AGENT

Role: Plans, routes, monitors

Decision: Analyzes request → delegates to specialists

RESEARCH AGENT

Role: Data gathering, RAG queries

WRITING AGENT

Role: Content generation, drafting

ANALYSIS AGENT

Role: Data analysis, patterns

Agent-to-Agent (A2A) Protocol: Standardized message format, shared state

Coordination Challenges

CHALLENGE 1: CONFLICTING DECISIONS

Problem: Agents with different priorities make incompatible choices

Example: Agent A deletes data that Agent B needs

Impact: System instability, data inconsistency

Frequency: 40% cite as primary blocker

CHALLENGE 2: EMERGENT COMPLEXITY

Problem: System behavior unpredictable from individual agents

Example: 3 agents work fine, adding 4th creates deadlock

Impact: Testing nightmare, production surprises

Frequency: 58% report as top concern

CHALLENGE 3: COMMUNICATION OVERHEAD

Problem: Message passing cost scales non-lin early ($O(n^2)$)

Example: 5 agents = manageable, 10 agents = latency explosion

Impact: Performance degradation, cost increase

Reality: 3-5 agents typical (coordination limit)

CHALLENGE 4: DISTRIBUTED STATE

Problem: No agent has complete picture

Example: Agent A thinks task complete, Agent B still waiting

Impact: Inconsistent decisions, coordination failures

Solution: Shared state management required

Orchestration Strategy

"Who decides? Who executes? Who monitors?"

ORCHESTRATION PATTERNS:

Hierarchical: Manager agent → specialist agents

Peer Collaboration: Equal agents negotiate

Consensus: Multiple agents vote on decisions

Feedback Loops: User evaluation shapes behavior

EMERGING STANDARDS:

Agent Protocol (A2A) - Open standard

MCP - Tool and agent communication

OpenAI Swarm - Experimental orchestration

LangGraph - Production-ready ⭐ (recommended)

CrewAI - Business automation platform

67%	40%	3-5
of enterprises exploring multi-agent systems	report coordination as primary blocker	agents typical (complexity limit)

Source: Gartner 2025, IONI AI, Multi-Agent Systems Research

Live Demo: Watch orchestrator delegate to specialist agents

See: Planning agent → Research agent → Writing agent → Execution agent

Agent-to-agent communication via simulated A2A protocol in n8n

Organizational Readiness

The Seven DORA Archetypes & AI Amplification

DORA Archetype Spectrum

LOW PERFORMANCE ← → HIGH PERFORMANCE

1. Foundational Challenges (10%)
"Struggling with basics"

2. Legacy Bottleneck (11%)
"Technical debt limits progress"

3. Constrained by Process (17%)
"Bureaucracy slows delivery"

4. High Impact, Low Cadence (7%)
"Quality work, infrequent"

5. Stable and Methodical (15%)
"Reliable, not fast"

6. Pragmatic Performers (20%)
"Speed + quality balanced"

7. Harmonious High Achievers (20%)
"Elite performance"

↓

AI Amplifies Problems

Examples: Weak testing → more bugs, Poor docs → unclear AI outputs

↑

AI Accelerates Success

Examples: Strong testing → better coverage, Clear docs → accurate AI

THE SEVEN DORA AI CAPABILITIES (Predictors of AI Success)

1. **Clear AI Stance → Governance (Component 3)**
   Reality: Only 31% have comprehensive governance (ISACA) 🔴
2. **Healthy Data Ecosystems → RAG foundation (Component 1)**
   Clean, structured, maintained data
3. **AI-Accessible Internal Data → RAG retrieval quality**
   APIs, documentation, searchable knowledge bases
4. **Strong Version Control → Model management (Component 2)**
   Model versioning, prompt versioning, deployment tracking
5. **Working in Small Batches → Agile + AI**
   Fast feedback loops, rapid iteration
6. **User-Centric Focus → Problem first, technology second**
   Not "AI for AI's sake"
7. **Quality Internal Platforms**
   90% have platforms, quality varies dramatically
   Platform quality determines AI success

WHERE DOES YOUR TEAM SIT ON THIS SPECTRUM?

- **Foundational challenges?**
  AI will amplify problems at scale.
- **Building capabilities?**
  AI adoption requires foundation first.
- **High performance?**
  AI will accelerate your existing success.

YOUR PILOT CAN'T TELL YOU WHICH YOU HAVE.

Pilots work in controlled environments with good practitioners. Scaling exposes organizational reality.

SCALING REQUIRES FOUR ELEMENTS

Technology Sophistication
RAG, model cost, agentic & multi-agent

Organizational Capabilities
DORA's seven AI capabilities enable delivery

Governance Frameworks
Built-in governance from day one

Cultural Readiness
Teams, processes, and leadership aligned

90%	80%	30%
using AI	productivity gains	don't trust AI-generated code

83% using AI, only 31% with policies

AI amplifies: weak foundations = problems, strong foundations = success

6 Critical Considerations for Enterprise AI Success

The 20% That Determines 80% of Outcomes

RAG Architecture & Optimization

Focus: Hallucination management, accuracy at scale, where it breaks

- Answer correctness rate
- Citation coverage percentage
- Retrieval precision at scale

Agentic AI & Orchestration

Focus: Tool permissions, circuit breakers, graceful failure handling

- Human-approval coverage percentage
- Rollback success rate
- Agent failure recovery time

Cost Management

Focus: Real-time visibility, model routing, budget control

- Cost per successful task
- Cache hit ratio
- Budget variance tracking

Governance & Policy

Focus: Operational enforcement, not theoretical compliance

- Policy conformance rate
- Audit artifacts generated per run
- Pre-deployment checkpoint completion

Prompt Injection Defense

Focus: #1 GenAI security threat (OWASP Top 10), runtime protection

- Blocked injection attempts
- Incident mean time to resolution (MTTR)
- Attack detection accuracy

Observability & Step-by-Step Verification

Focus: Monitor each workflow step, not just final outcomes

- Trace completeness percentage
- Step-level error detection rate
- Root cause identification speed

⚠️ ENTERPRISE REALITY WITHOUT THESE 6 CONSIDERATIONS:

- Agents fail, requiring rehiring of replaced staff
- $500K+ monthly budget surprises
- Security breaches from prompt injection attacks
- Inability to debug when systems fail
- Regulatory violations and compliance penalties
- User trust erosion from inconsistent outputs

Strategic Self-Assessment

Four Critical Questions for Technology Leaders

1 Where does your team sit in the DORA archetypes?

Be honest. Your pilot might work regardless, but AI will amplify your current reality when you scale. Foundational challenges? Building capabilities? High performance? Assessment drives strategy.

2 What foundational practices need strengthening before scaling AI?

Testing and quality assurance? Version control and deployment? Platform quality and reliability? Fast feedback loops? Fix foundations first. AI amplifies what's already there—good or bad.

3 Do you have visibility into AI tools employees are using?

Have you inventoried shadow AI? Mapped data flows to external services? Assessed risks of unsanctioned tools? Communicated governance policies? 75% will use unsanctioned AI by 2027—what's happening now?

4 Who owns AI governance in your organization?

CIO? CTO? CISO? Cross-functional council? Or honest answer—nobody clearly owns it? Governance without ownership fails. Who has authority and accountability for AI governance decisions?

neu•eon

Strategic Insights for Technology Leaders

From Prototype to Production: Scaling AI Systems in the Enterprise

Prototype

Innovation begins with focused experimentation and proof of concept

Pilot

Controlled deployment validates feasibility and surfaces integration challenges

Production

Enterprise-scale systems deliver measurable business value and competitive advantage

---

Scott Weiner
Chief Technology Officer & AI Lead
NeuEon, Inc.
sweiner@neueon.com

www.neueon.com

www.linkedin.com/in/sweiner

> *"High-performing technology organizations are 2.5x more likely to meet or exceed their organizational performance goals."*
>
> — DORA State of DevOps Report

Strategic Self-Assessment

Questions for Technology Leaders

FROM PROTOTYPE TO PRODUCTION

It's not just about technology. It's about organizational readiness.

Four Strategic Questions

1

WHERE DOES YOUR TEAM SIT IN THE DORA ARCHETYPES?

- ☐ Foundational Challenges / Legacy Bottleneck / Constrained by Process
- ☐ High Impact Low Cadence / Stable and Methodical
- ☐ Pragmatic Performers / Harmonious High Achievers

→ Be honest. AI will amplify your current reality.

2

WHAT FOUNDATIONAL PRACTICES NEED STRENGTHENING BEFORE AI?

- ☐ Testing and quality assurance
- ☐ Version control and deployment
- ☐ Platform quality and reliability
- ☐ Fast feedback loops

→ Fix foundations first. AI amplifies what's there.

3

DO YOU HAVE VISIBILITY INTO AI TOOLS EMPLOYEES ARE USING?

- ☐ Shadow AI inventory completed
- ☐ Data flow mapping to external AI services
- ☐ Risk assessment of unsanctioned tools
- ☐ Governance policies communicated and enforced

→ 75% will use unsanctioned AI by 2027. Do you know what's happening now?

4

WHO OWNS AI GOVERNANCE IN YOUR ORGANIZATION?

- ☐ CIO / CTO / CISO
- ☐ Cross-functional council
- ☐ Nobody (honest answer)

→ Governance without ownership fails. Who has authority and accountability?

KEY RESOURCES:

RESEARCH & FRAMEWORKS:

- DORA 2025 Report: cloud.google.com/dora
- OWASP LLM Top 10: genai.owasp.org/llm-top-10/
- NIST AI Risk Management Framework: nist.gov/itl/ai-risk-management-framework
- ISACA AI Governance: isaca.org
- EU AI Act: eur-lex.europa.eu (€35M penalties)
- ISO/IEC 42001: AI Management Systems

TECHNICAL REFERENCES:

- ChromaDB Research: research.trychroma.com
- Weaviate RAG Strategies: weaviate.io/blog
- Model Pricing: OpenAI, Anthropic (October 2025)
- OpenRouter Cost Comparison: openrouter.ai
- Model Context Protocol: modelcontextprotocol.io
- LangGraph (Agentic AI): langchain-ai.github.io/langgraph/

FROM PROTOTYPE TO PRODUCTION

It's not just about technology. It's about organizational readiness.

Scott Weiner

CTO & AI Lead, NeuEon

Available for follow-up discussions about your specific organizational context.

THANK YOU