syllotips

5-minute lightning insight

When AI fails: who owns AI reliability?

Recent scientific and industry approaches to continuous improvement for AI agents

Giorgio Barnabò, PhD — CEO and co-founder, Syllotips

Stories from the industry

Two production AI support systems show the same lesson: reliability is not a launch milestone, it is an operating loop.

RAG chatbot

DoorDash

Summarize issue → Retrieve knowledge → Guardrail response

Main risk: confident answers that drift from the knowledge base.

Fix: guardrails, LLM Judge, human review, KB + retrieval improvements.

90% fewer hallucinations

99% fewer severe compliance issues

Planning agent

Lyft

Detect intent → Route to subagent → Call tools safely

Main risk: wrong routing, weak prompts, or brittle tool use at scale.

Fix: structured prompts, tracing, dashboards, safety checks, LLM-as-a-Judge.

~2 weeks to launch new agents

100% automated evaluation coverage

From case studies to recurring failure modes

What generalizes?

RAG chatbot failure points

- Knowledge gaps
- Knowledge drift
- Retrieval failure
- Weak citations
- Content dumping

Planning agent failure points

- Wrong intent
- Bad planning
- Poor tool calls
- Unclear tool semantics
- State failures
- Guardrail gaps

Sources: DoorDash engineering blog, 2024; LangChain / Lyft case study, 2026.

Symptoms and consequences of weak continuous improvement

When reliability loops are missing, technical issues quickly become business problems.

RAG chatbot

RAG without continuous improvement

Operational signals

Symptoms

- Hallucinations
- "I don't know" answers
- Incomplete / partial answers
- No business people involved

Leads to

Consequences

- Low satisfaction & usage
- Slow time-to-value
- Missed opportunity
- Reputational damage
- Low maintainability

If content quality and retrieval quality do not improve continuously, the chatbot becomes less trusted and less useful.

Planning agent

Agents without continuous improvement

Business impact

Symptoms

- Many failed runs
- Slow execution and infinite loop
- Disproportionate token spending
- No business people involved

Leads to

Consequences

- Low satisfaction & usage
- Token waste
- Increased security risks
- Slow time-to-value
- Low maintainability

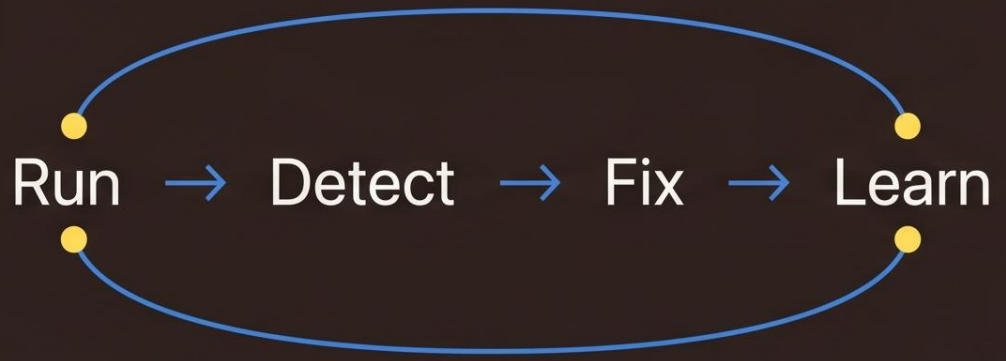
If prompts, routing, tools, and guardrails are not continuously improved, agent systems become expensive, brittle, and risky.

What both patterns share

- Technical failures become trust failures
- No business feedback loop = slower improvement
- Continuous improvement is a product capability, not an afterthought

Research points to one pattern

AI agents do not become reliable at launch. They become reliable through measured, repeated improvement.



Benchmark reality

Scientific result

τ-bench: A benchmark for tool-agent-user interaction in real-world domains

ICLR 2025

pass^8 < 25%

Even strong agents struggle to behave consistently across repeated real-world interactions.

Failure diagnosis

Scientific result

Where LLM agents fail and how they can learn from failures

Under review, ICLR 2026

+26% task success

Fixing the root cause of a failure matters more than patching every visible symptom.

Evolving context

Agentic context engineering: evolving contexts for self-improving language models

ICLR 2026

+10.6% on agents

Agents improve when their instructions and memory evolve into structured playbooks.

Evolving skills

SkillOpt: executive strategy for self-evolving agent skills

arXiv, 2026

Best or tied on 52 / 52 tests

Reusable skills can be trained and validated like a practical improvement layer.

Memento-Skills: Let Agents Design Agents — arXiv, 2026 — skills improve through reflective learning.

From industry cases to research benchmarks, the message is the same: AI reliability is an operating capability.