

Executive Summary

Agentic RAG systems combine:

- Reasoning models
- Retrieval systems
- Tool use (increasingly via MCP)
- Memory (short-term, long-term, and now tiered/benchmarked)
- Workflow orchestration
- Governance controls
- A harness layer that governs all of the above

The goal is not maximum autonomy. The goal is reliable, auditable, policy-aligned autonomous execution. Modern production systems optimize for:

- Accuracy
- Cost efficiency
- Safety
- Explainability
- Human oversight
- Operational resilience

NEW IN 2026

The single biggest framing shift: the model is now understood as the smallest part of the system. Practitioners increasingly describe production agents with the equation

Agent = Model + Harness,

where the harness — not the model — determines reliability, and benchmarks such as Harness-Bench show that swapping the harness around an identical model pool can move task-success scores by over 20 points.

Revision #1

Created 2026-08-24 11:52:37 UTC by Aisha M. Nur

Updated 2026-08-24 11:53:00 UTC by Aisha M. Nur