The Agent Harness

Introduction

NEW IN 2026

Harness engineering emerged through 2026 as a distinct discipline, formalized in surveys covering 100+ papers and 23 production systems. The core claim: at fixed model capability, agent-computer interface design materially changes benchmark outcomes.

The Equation:

Agent = Model + Harness

The harness treats the LLM as a frozen reasoning utility and moves the responsibility for safety, execution accuracy, multi-step orchestration, and adaptive memory into the surrounding infrastructure. A production harness is generally described as a three-layer system:

Layer	Responsibility
Information Layer	What data the agent can observe; what tools it may invoke at a given moment; vector storage, memory compilation, tool registries.
Control Layer	Planner / generator / evaluator roles; structured handoff artefacts; verification gates; state machines governing step sequencing.
Governance Layer	Permission boundaries, policy checks, approval gates, and the audit trail that makes autonomy observable.

Harness Components

Six Harness Components

- State and persistence — durable memory and workflow checkpoints across long-running runs.
- Security and governance — permissioning, sandboxing, credential scoping.
- Orchestration and tool routing — how planner/executor/verifier roles hand off work.
- Context assembly — the same responsibility described in the Context Engineering layer below.
- Observability — traceable artefacts across the full agent lifecycle (AgentOps).
- Recovery and self-healing — retries, re-planning, and graceful degradation when a step fails.

Harness vs. Framework

A framework (LlamaIndex, LangGraph) supplies reusable primitives. A harness is the specific, opinionated configuration of those primitives — plus the repo-local instructions, feedback loops, and verification gates — that a team assembles for one production system. Two teams using the same framework can have very different harnesses, and that difference is now measurable: Harness-Bench reports swings of over 20 points in task success when the harness changes but the underlying model pool does not.