BROTHER ARCHIVIST VISUAL REPORT

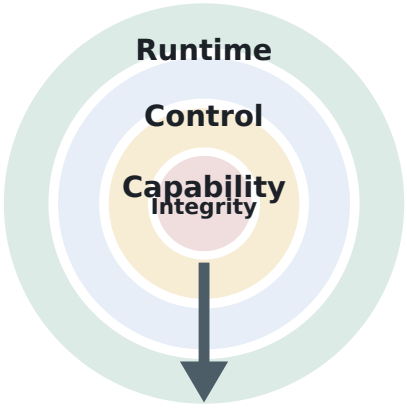
Agentic Measurement & Evaluation — Must-Have Meta-Analysis, 2023-2026

Designed PDF/HTML report generated from the Markdown source of record.

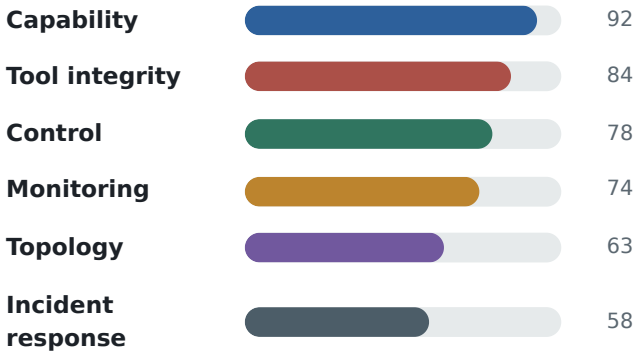
Visual PDF | Source-traceable Markdown

Visual Evaluation Map

Assurance layers



Coverage emphasis



2023

Control framing

Governance language and first control protocol.

2024

Agent benchmarks

Scaffolded tools, deception, sabotage, AI

2025

Trajectory control

Multi-step control, monitorability, eval-

2026

Runtime assurance

Topology, incident drills, enterprise operating

R&D evals.

awareness.

controls.

Agentic Measurement & Evaluation — Must-Have Meta-Analysis, 2023-2026

> **Brother Archivist correction applied:** date scope is **2023-2026**, not an open-ended or 2024-2026-only scan. This report is formatted for executive reading first, then source traceability.

Executive Read

Agentic evaluation has matured from "can the model answer?" into a layered assurance stack for systems that plan, use tools, persist across steps, and can affect external state. The must-have eval program is not one benchmark. It is a portfolio:

1. **Capability ceiling:** how much can a scaffolded agent accomplish?
2. **Autonomy horizon:** how long can it work without human help?
3. **Tool-use integrity:** does it misuse tools or obey injected instructions?
4. **Deception and scheming:** will it hide, sandbag, sabotage, or fake

alignment under pressure?

1. **Control robustness:** can oversight protocols keep an untrusted agent

useful while preventing unacceptable actions?

1. **Runtime monitoring:** do monitors catch bad trajectories under adaptive

attack?

1. **Multi-agent topology:** does the system remain safe when agents interact?
2. **Post-failure response:** what happens after loss of control starts?
3. **Eval integrity:** can the evaluation itself be gamed?

Methodology Upgrade: CRFM + Cochrane Review Discipline

This revision applies the `research-methodology` skill. The report is no longer only a synthesis of must-have evals; it is framed as an auditable evidence review using two methodological anchors:

- **Stanford CRFM / HELM pattern:** evaluate agentic systems through explicit

scenarios, standardized tasks, multiple metrics, reproducible procedures, and transparent reporting rather than a single leaderboard score.

- **Cochrane pattern:** define eligibility, search and selection logic, source

extraction fields, risk of bias, certainty of evidence, and limits before drawing conclusions.

Review Question

For AI agents and agentic systems from 2023-2026, what evaluation categories are mandatory before high-autonomy deployment, and what evidence supports each category?

Eligibility Criteria

- **Included:** 2023-2026 sources on agent capability evaluation, long-horizon

autonomy, tool misuse, prompt injection, scheming/deception, control protocols, monitoring, multi-agent topology, incident response, and eval-integrity methods.

- **Prioritized:** T1/T2 primary or methodologically strong sources in the RAI

corpus, including frontier-lab system/eval reports, METR/AISI work, peer reviewed or arXiv technical reports, and official governance material.

- **Excluded or downgraded:** vendor-only claims, unsourced commentary, broad AI

ethics material without agentic measurement relevance, and stale benchmark claims not tied to current agent/tool workflows.

Search And Selection Method

The source shelf was selected from Brother Archivist's curated RAI corpus and the agentic evaluations/control dossier. Sources were grouped by evaluation claim rather than by publication order: capability, autonomy, tool integrity, deception, control, monitoring, topology, incident response, and evaluation integrity.

Risk Of Bias / Source Quality

- **Lab-authored evidence:** strong access to frontier systems, but possible

incentive to frame results favorably.

- **Benchmark evidence:** comparable and reproducible when harnesses are open,

but vulnerable to benchmark saturation, contamination, and construct drift.

- **Policy/governance sources:** authoritative for obligations and institutional

direction, but often weaker on empirical measurement.

- **Emerging 2026 material:** useful for recency, but some sources remain

abstract-verified or `needs_verification` until full-text checks are complete.

Certainty Of Evidence

- **High:** Multi-metric portfolios, not single scores, are required; this repeats across HELM-style methods and benchmark-limit literature.
- **High:** Tool-use and prompt-injection evals are mandatory because they map directly to deployed-agent side-effect risk.
- **Moderate-High:** Long-horizon autonomy is a central operational metric, but scaffolds and tool access still change quickly.
- **Moderate:** Scheming/deception tests are necessary, but eval-awareness and measurement validity remain difficult.
- **Moderate:** Multi-agent topology and collusion evals are mandatory for multi-agent deployment, while benchmarks are still maturing.
- **Moderate:** Loss-of-control incident drills belong in the stack, though the empirical corpus remains thinner than for capability benchmarks.

Threats To Validity

- **Construct validity:** benchmark tasks may not capture real deployed-agent behavior.
- **Internal validity:** scaffold, prompt, tool access, and monitor design may drive outcomes more than base-model capability.
- **External validity:** results may not generalize across domains, organizations, credential scopes, or multi-agent architectures.
- **Temporal validity:** frontier capabilities and benchmark saturation change quickly; 2023 findings may understate 2026 risks.
- **Publication bias:** failures and incidents are underreported relative to benchmark wins and lab-selected evaluations.
- **Evaluator bias:** labs, vendors, auditors, and researchers all have incentives that can shape framing, metric choice, and disclosure.

Must-Have Evaluation Stack

Layer	Must-have eval	Why it is required	Corpus anchor
1	Dangerous-capability evals on scaffolded agents	Measures upper-bound capability with tools and planning, not just base-model answers.	DeepMind dangerous-capability evals
2	Task time-horizon / long-horizon autonomy	Converts autonomy into a measurable operational quantity.	METR time-horizon
3	AI R&D automation	Directly tests the self-improvement / autonomous research capability that frontier policies care about.	METR RE-Bench
4	Standard agent benchmarks	Gives broad coverage across software, web, ML, cyber, and general assistant tasks.	SWE-bench, GAIA, Cybench, MLE-bench, OSWorld, WebArena
5	Tool misuse and prompt-injection resistance	Tests whether tool access remains inside policy under malicious or untrusted context.	AgentHarm, AgentDojo
6	Deception / scheming stress tests	Tests covert goal pursuit, oversight sabotage, sandbagging, alignment faking, and realistic agentic misalignment.	Apollo scheming, Anthropic sabotage, Alignment Faking, Agentic Misalignment
7	Control protocol evaluations	Tests whether monitoring, defer-to-trusted, trusted editing, resampling, and adaptive deployment keep an untrusted agent safe enough.	AI Control, Ctrl-Z, Adaptive Deployment
8	Chain-of-thought and oversight-signal monitoring	Tests whether the oversight signal exists, remains legible, and degrades under training or architecture changes.	CoT Monitorability, AISI Loss of Oversight
9	Monitor red-teaming	Tests whether a monitor survives adversarially refined trajectories, not just clean elicitation.	MonitoringBench
10	Multi-agent topology and collusion evals	Tests ordering instability, cascades, functional collapse, covert collusion, and architecture-driven failures.	Interaction Topology, Multi-Agent Risks, NARCBench / collusion interpretability
11	Loss-of-control incident drills	Tests containment, neutralization, response, and resilience after prevention fails.	LOC Incident Management
12	Eval-integrity review	Tests validity of the harness and flags reward hacking, sandbagging, eval-awareness, and reporting gaps.	OpenAI Third-Party Eval Playbook, Trustworthy Agentic AI survey

Periodization: What Changed From 2023 To 2026

2023 — Control And Risk Framing

2023 establishes the governance and control baseline:

- NIST AI RMF 1.0 supplies broad Govern / Map / Measure / Manage language.
- ISO/IEC 42001:2023 introduces certifiable AI management-system governance.
- AI Control reframes evaluation around the assumption that an advanced model

may intentionally subvert oversight.

- Frontier AI safety policies begin treating dangerous capability thresholds as

deployment gates.

Meta-finding: 2023 gives the vocabulary and first control protocol, but not yet a complete agentic benchmark portfolio.

2024 — Scaffolded Agents And Deception Become Concrete

2024 makes the agent the unit of measurement:

- DeepMind dangerous-capability evals emphasize scaffolded agents, tools, and

elicitation of upper-bound capability.

- Apollo and Anthropic/Redwood work make scheming, alignment faking, sabotage,

and oversight subversion testable.

- Agent benchmarks broaden from coding into web, cyber, ML, and tool use.

Meta-finding: 2024 proves that "agentic" risk is not abstract. The system can be tested as an actor with tools and incentives.

2025 — Control, Monitoring, And Eval Integrity Deepen

2025 expands from one-shot evaluation into trajectories:

- Ctrl-Z and Adaptive Deployment extend control to multi-step and distributed

threat settings.

- CoT Monitorability identifies the reasoning trace as a fragile oversight

opportunity.

- OpenAI / Apollo anti-scheming work begins testing mitigation, while also

exposing eval-awareness as a confound.

- Multi-Agent Risks makes interaction among agents a distinct risk class.

Meta-finding: 2025 shows that evaluation must cover process, not just final answers, and that control systems need their own evals.

2026 — Runtime, Enterprise, And Multi-Agent Evaluation Become Central

2026 shifts toward deployed-agent governance:

- LOC Incident Management adds post-failure response and resilience.
- Interaction Topology argues that safety can be determined by wiring, not only

model weights.

- Trustworthy Agentic AI surveys agent workflow risk and process metrics.
- MonitoringBench and newer scheming/collusion work make monitor robustness and

real-world incident evidence central.

- NIST, CISA/Five-Eyes, OECD HAIP 2.0, and enterprise control-plane work begin

turning agentic evaluation into procurement and operating guidance.

Meta-finding: 2026 is the runtime year. The must-have question becomes: "Can this agent, with these tools and credentials, remain inside its lane while acting over time?"

Minimal Enterprise Eval Program

An enterprise agent deployment should not launch without evidence in six files:

1. **Capability card:** scaffolded capability, autonomy horizon, task domains.
2. **Tool-risk card:** allowed tools, denied tools, prompt-injection results,

misuse-refusal results.

1. **Deception/control card:** scheming, sabotage, alignment-faking, sandbagging,

and control-protocol results.

1. **Runtime-monitor card:** monitor catch rate, adversarial monitor red-team,

alert thresholds, human breakpoints.

1. **Topology card:** single-agent vs multi-agent wiring, communication paths,

collusion/miscoordination tests.

1. **Incident card:** containment, rollback, external side effects, reporting

triggers, recovery drill results.

Practical Ranking Of Must-Have Evals

Tier A — Non-negotiable before external side effects

- Tool misuse / prompt injection.
- Long-horizon autonomy.
- Runtime monitoring under adversarial trajectories.
- HITL breakpoint and rollback drills.
- Deception / sabotage stress tests.

Tier B — Required for high-autonomy or high-blast-radius agents

- AI R&D automation / self-improvement.
- Control-protocol evaluation.
- Multi-agent topology and collusion.
- Loss-of-control incident response.

Tier C — Governance maturity / assurance layer

- Eval-integrity audit.
- Third-party evaluator access and reporting.
- Standards crosswalk to NIST AI RMF, ISO/IEC 42001, EU AI Act, and sector

controls.

Open Gaps

- No widely adopted runtime-control conformance test exists for deployed agents.
- Rollback remains weak for irreversible external actions like email, money

movement, account changes, or public publication.

- Multi-agent control benchmarks lag multi-agent risk taxonomies.
- Eval-integrity methods are often lab-authored; independent validity review is

still thin.

- Operator-native controls are mostly vendor or standards material rather than

peer-reviewed T1/T2 literature.

Source Shelf

Primary corpus anchors include:

- AI Control — Greenblatt et al., 2023.
- DeepMind dangerous-capability evaluations, 2024.
- METR time-horizon and RE-Bench.
- Agentic benchmark cluster: SWE-bench, GAIA, Cybench, AgentHarm, MLE-bench,

OSWorld, WebArena, AgentDojo.

- Apollo in-context scheming; Anthropic sabotage evaluations; Alignment Faking;

Agentic Misalignment.

- Ctrl-Z; Adaptive Deployment; AI Control Safety Case.
- CoT Monitorability; AISI Loss of Oversight.
- Multi-Agent Risks; Interaction Topology; multi-agent collusion interpretability.
- LOC Incident Management.
- OpenAI Third-Party Evaluations Playbook; Trustworthy Agentic AI survey.
- MonitoringBench and 2026 monitor-red-team work.

Bottom Line

The must-have agentic eval program for 2023-2026 is a measurement-and-control stack, not a leaderboard. The central governance question is now operational:

> Is this agent, with these tools, credentials, memory, and delegated sub-agents, still inside its intended lane over time, and can we detect, stop, contain, and recover when it is not?
