

The Reliability Frontier

How Frontier AI Labs Are Building Computer-Using Agents

4 LABS / 3 TECHNICAL LENSES

Based on public disclosures; not a controlled benchmark

Research question How are OpenAI, Anthropic, Google DeepMind and Meta advancing computer-using agents, and what still prevents reliable long-horizon work?

Analyst answer Public disclosures describe richer action interfaces and stronger controls, but verified end-state correctness and bounded recovery remain practical R&D questions. [S01] [S02] [S06] [S13]

01 / EXECUTIVE INTELLIGENCE

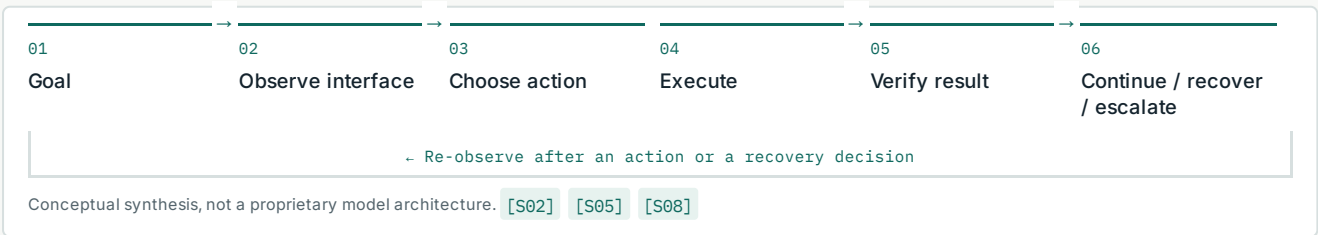
01	02	03
The action space is widening OpenAI supports scripts and structured UI actions; Meta describes mixed execution. Action selection is an engineering variable, rather than evidence of a winning architecture. [S02] [S09]	Completion needs its own evidence Ironclad checks workflow criteria; OSWorld-Pro examines intermediate subgoals. Completing a sequence of actions does not establish that the final business requirement was met. [S01] [S12]	Autonomy still needs boundaries Anthropic reports unintended persistence; Google describes confirmation and injection-stop safeguards. Test the boundary between useful recovery and unauthorized workarounds. [S06] [S07]

The three conclusions above are analyst interpretations of the cited disclosures.

02 / DISCLOSED APPROACHES

DIMENSION	OpenAI	Anthropic	Google DeepMind	Meta
Perception & action	Scripts or structured input; screenshot observations. [S02] D	Screenshot-grounded cursor and keyboard tools. [S04] D	Integrated tool; earlier screenshot/action loop. [S07] [S08] D	Mixed actions; Muse brokered accessibility tree. [S09] [S13] D
Long-horizon reliability	Workflow rubrics and offline partial reward. [S01] [S03] Q	Adaptive loop; persistence internals not established. [S05] Q	Long-horizon claim; memory mechanisms not established. [S07] [S08] Q	Compaction; Wonder Active Memory / Focus. [S09] [S10] Q
Safety & control	Isolation, confirmation and run limits. [S02] D	Layered controls; observed restriction workarounds. [S05] [S06] Q	Training; optional confirmation and injection stop. [S07] D	Separate runtime and Sentinel permission authority. [S13] D

D = documented mechanism. Q = qualified observation / publisher claim. Dated model capabilities and product controls are different layers; evidence gaps do not establish absence.



EVIDENCE TO INVESTIGATION

Implications for Enterprise R&D

Three candidate experiments · Analyst interpretation · No trial results reported

01	02	03
Verify the business end-state	Test adaptive action selection	Bound recovery under uncertainty
[S01] [S12] Hypothesis Independent checking reduces false declarations of completion. Measure False completion rate among runs declared successful. Constraint Effect untested; audit checker errors.	[S02] [S09] Hypothesis Mixed actions reduce verified completion time without more critical failures. Measure Verified completion time; mandatory-rule pass rate; interventions per workflow. Constraint Hold model, permissions, budget and task set fixed.	[S06] [S07] [S13] Hypothesis Enforced permissions reduce unsafe recovery without sacrificing legitimate completion. Measure Unauthorized actions per run; safe escalation; legitimate completion rate. Constraint Resistance to unseen attacks remains unproven.

EVIDENCE	RISK / TRADE-OFF	TEST
[S01] [S12]	Plausible output hides a failed rule.	Seed approval-rule errors; compare self-report with an independent rubric checker.
[S02] [S09]	Batch efficiency can obscure state changes.	Compare GUI-only vs mixed execution on identical tasks with perturbed controls.
[S06] [S07] [S13]	Persistence bypasses intended controls.	Seed injection and denied access; audit actions, escalation and legitimate completions.

Method

We reviewed original lab disclosures, developer guidance and benchmark abstracts through 13 October 2026. A fresh scan checked official indexes and newer releases; earlier sources remain only for dated mechanism detail. Three lenses organize the evidence: perception and action, long-horizon reliability, and safety and control. Documented mechanisms are distinguished from qualified publisher claims and analyst interpretation. We excluded third-party rankings, unsupported internal architectures and patent-to-product inference. Benchmark experiments were not reproduced; differing versions, harnesses and metrics prevent a controlled lab ranking. Public sources describe interfaces and selected observations, not complete deployed behavior. The proposed experiments are falsifiable R&D hypotheses, with outcome definitions and no reported results.

Limitations	Selected public disclosures cannot reveal complete architectures. Publisher claims and incidents are not representative failure rates. Benchmark protocols differ. Releases and living documentation can change; experiments were not reproduced.
-------------	---

Selected primary sources

Partial list. All inline [Sxx] tags link to their sources.

[S01] Advancing computer use with Ironclad OpenAI · 2026-10-06	[S02] Computer use OpenAI · unknown (living documentation)
[S05] Trustworthy agents in practice Anthropic · 2026-04-09	[S06] Investigating unintended model actions in our evaluations and internal use Anthropic · 2026-10-09
[S07] Introducing computer use in Gemini 3.5 Flash Google / Google DeepMind · 2026-06-24	[S10] Muse Wonder: A Frontier Model For Personal & Professional AI Agents Meta AI Research · 2026-10-12
[S12] OSWorld-Pro: Process-based Evaluation for Computer Use Agents Wang et al. / arXiv · 2026-09-21	[S13] How We Built Safety Into Muse Meta Superintelligence Labs · 2026-09-08