

AGENT SIGNAL

16 October 2026 | Edition 03

Model updates

HydraFusion Max

Copilot combines multiple models into one response; Max is a higher-effort preview option. What's new: GitHub introduced Max, step-specific reasoning effort, conditional judging in Cascade and parallel workspace-aware critics in Critique. Why: Compare verified code quality, runtime and cost with a single-model approach before upgrading. [Source](#)

Gemini 4 Argon

A model built for long coding tasks, including migrations across a codebase. What's new: Google opened developer access on the paid Gemini API and in Antigravity for Google AI Ultra subscribers. Why: Try it on a multi-file migration with tests that can verify the result. The relevant question is whether it completes your repository task reliably. [Source](#)

Repositories to try

github/gh-aw | MIT

A CLI extension converts Markdown agent instructions to GitHub Actions workflows without needing inference to compile. The v0.92.0–v0.92.1 releases tighten model-routing and patch integrity, add an offline skill reader, and improve work-queue diagnostics. [Source](#)

AGENT SIGNAL

16 October 2026 | Edition 03

Agent ecosystem

AI Scan for pull requests

An AI security checker finds flaws in changed PHP, Shell, Terraform and other code outside CodeQL's coverage. Broadens the review but does not replace CodeQL, tests or manual checks. Metered Code Security AI Credits may apply; do not enable without a budget check. [Source](#)

Agentic autofix for security alerts

A coding agent investigates a CodeQL security alert, edits related files, reruns the scanner and prepares a draft fix PR. Checks a fix rather than merely suggesting one, but needs human review; not accessible via this personal free workflow. [Source](#)

Copilot Studio agent accounts and memory

Managed business-process agents use their own identities and remember process context across interactions. Identity and retained context could make long-running enterprise delegation auditable; not a free/open-source coding model. [Source](#)

Research watch

EvoProbe: evolving agent benchmarks | arXiv preprint — not peer-reviewed

Feedback from agent failures is used to regenerate more revealing executable benchmark tasks. Across two rounds, 66 validated cybersecurity tasks and eight models, updated tasks lowered average outcome 11.8% and consistent three-run success 40.0%, under the specified controls. Limit: Two rounds and one generated cybersecurity setting; not proof across production agents. [Source](#)

Tau-Emotion-Bench: emotionally difficult agent tasks | arXiv preprint — not peer-reviewed

An agent benchmark varies a simulated user's emotional response while holding the tool-based objective steady. Authors report that dissatisfied-user interactions expose reliability failures missed by calm, cooperative simulations. Limit: Simulated users in ten controlled domains may not generalize to real customers. [Source](#)

Today's experiment

Try github/gh-aw: offline skill and safe dry-run compile

Inspect gh-aw's offline skill and compile a small read-only agent workflow using --dry-run; never dispatch it. Check permissions and routing without paying for inference or running GitHub Actions. Tested: Official release and CLI docs reviewed 16 October; no commands, Actions or model experiment executed. [Source](#)