

Proactive Agents Arrive: Claude Duet, ChatGPT Parallel Agents and the Rise of Hermes

Anthropic's background agent Claude Duet, OpenAI's parallel agents in GPT-6 Pro, and Nous Research's self-improving Hermes Agent show AI shifting from answering to working on its own.

Themes: Models & Frontier · Enterprise & Strategy

Note: most details below come from news roundups and secondary coverage; vendor pages could not be fetched this run, and unconfirmed points are marked.

1) Claude Duet: An Agent That Works While You Don't

Anthropic launched Claude Duet as a research preview on October 21, 2026. Reports say it runs on Claude Opus 5.5, watches connected apps such as email and calendar, sends a morning brief, and checks again every few hours. It is waitlist-only for US Pro and Max subscribers, on iOS, Android and the web, and background activity counts toward existing usage limits. Reports say it asks approval before sending messages, making payments, signing documents or deleting files; one source adds that trading and account transfers are blocked. Anthropic itself says it can misjudge what is urgent.

The concept, explained simply: A proactive agent does not wait for a prompt. It runs on a schedule, notices events in your apps, and proposes or takes actions, pausing for your approval on anything consequential.

Why it matters: Chatbots need you to remember to ask; proactive agents shift the work to monitoring and judgement. The approval gate is the key safety design, because a wrong guess about urgency costs more when no one is watching.

2) ChatGPT Parallel Agents in GPT-6 Pro

On October 16, OpenAI reportedly added parallel agents to GPT-6 Pro in ChatGPT. One complex request is split into subtasks that several agents run at once; users can follow progress in the chat and open each finished subtask to inspect it. It is available on Pro, Business and Enterprise plans, rolling out gradually. Coverage is secondary: OpenAI has not published concurrency limits, latency or quality numbers (unverified). One report mentions a default cap of 6 subagents in the API docs, which may not apply to ChatGPT.

The concept, explained simply: Parallel agents work like a manager handing different parts of a report to several analysts at the same time, then combining the results. It saves wall-clock time, but costs more tokens and needs the pieces to be independent.

Why it matters: Inspectable subtasks make multi-agent work auditable. Speed gains depend on tasks that split cleanly, and total usage can rise even as waiting time falls.

3) Kimi K3.1 Pro and the Open-Weight Race

An aggregator lists Moonshot's Kimi K3.1 Pro as the newest release, with a 1M-token context window, available through Kimi's apps and API, and full weights promised by November 4. We could not open Moonshot's own blog, so treat details as unverified. It follows other open-weight agent models this month, including Nvidia's Nemotron 3.5.

The concept, explained simply: Open-weight means the trained model files are published so anyone can run and fine-tune them. A context window is how much text the model can consider at once; 1M tokens is roughly several long books.

Why it matters: Open-weight models near the frontier give companies a cheaper, private alternative and pressure closed-model pricing. A promise of weights is not the release itself, so wait for the files and independent benchmarks.

VIRAL / OPEN-SOURCE SPOTLIGHT

Hermes Agent (Nous Research)

An open-source agent released in February 2026 that creates reusable skills from experience, keeps persistent memory across sessions, and runs from a terminal or messaging apps. Reports put it between about 214,000 and

230,000 GitHub stars, with the higher figure the newest and least verified. TechCrunch-sourced coverage says Nous is finalizing a round of at least \$75 million at a \$1.5 billion valuation, not officially confirmed. It is taking off because an agent that improves with use feels different from one that starts fresh every chat, and it is free and self-hostable.

MARKET SIGNAL

Every major lab now ships agents that act without being asked or that run in parallel, while open projects like Hermes attract valuation-scale funding. Competition is moving from model benchmarks to agent products: permissions, memory, supervision and cost per completed task.

PRACTICAL TAKEAWAYS

1) Define your approval rules first.

Before trying a proactive agent, list which actions it may take alone and which need your sign-off.

2) Split tasks to use parallel agents.

Break work into independent pieces with clear outputs so parallel agents save time instead of colliding.

3) Track tokens, not just minutes.

Parallel and background agents shorten waiting but can multiply usage, so measure cost per finished task.

4) Treat promised weights as pending.

Evaluate open models only after the files and independent benchmarks are public.

SELF-TEST

Q1. What makes an agent proactive?

A1. It runs in the background, monitors apps and acts or alerts without a prompt, asking approval for consequential steps.

Q2. What do parallel agents do?

A2. They split one task into subtasks run at the same time and then combine the results.

Q3. What does open-weight mean?

A3. The trained model files are published so others can run and adapt them.

Q4. Why did Hermes Agent go viral?

A4. It builds skills and keeps memory so it improves with use, and it is open source.

TOMORROW

What We're Watching: Anthropic's own documentation for Claude Duet and any waitlist expansion. What We're Watching: OpenAI's release notes on parallel-agent limits and pricing. What We're Watching: whether Kimi K3.1 Pro weights appear on schedule.

TOPICS COVERED SO FAR (RECAP)

Recent days covered AI-found vulnerabilities and patching speed, Claude Opus 5.5 pricing, Nvidia's Vera Rubin POD, OpenAI oversight incidents, Nemotron 3.5, GPT-6.1 Astra, voice agents and Gemini 4 Argon. Today added Claude Duet's proactive agent model, ChatGPT parallel agents, Kimi K3.1 Pro and the open-weight race, and the Hermes Agent.

Sources: MacRumors on Claude Duet (macrumors.com/2026/10/21/anthropic-claude-duet/), testingcatalog.com/claude-duet, iphoneincanada.ca, kingy.ai and mind-verse.de on GPT-6 Pro parallel agents, llm-stats.com/ai-news, TechCrunch on Nvidia Nemotron 3.5 (techcrunch.com/2026/10/19/), CO/AI, Techzine and AI Weekly on Hermes Agent.