Proactive Agents Arrive: Claude Duet, an Open-Weight Giant and Consent Gates

Anthropic's Claude Duet research preview joins Meta Muse and OpenAI dots in the race for agents that act without being asked, Moonshot's Kimi K3.1 Pro brings a 2.8-trillion-parameter mixture-of-experts model to the open-weight world, and both stories turn on how to keep powerful agents safe.

Themes: Models & Frontier · Governance & Safety

Note: web page fetches were blocked today, so this lesson relies on search-result summaries from news and aggregator sites. Details such as pricing, model internals and availability are secondary reports and unverified; check the original announcements.

1) Claude Duet: An Agent That Messages You First

Anthropic opened a research preview of Claude Duet on October 21, 2026. It is a separate app that connects to your work accounts such as email and calendar, keeps working on assigned tasks while you are away, and runs on its own virtual machine so your laptop or phone need not stay on. Reports say it checks connected accounts every few hours rather than continuously, and that it asks permission before sending messages, making payments, signing documents or deleting anything. Access starts as a US waitlist for Claude Pro and Max subscribers. Some coverage links it to Claude Opus 5.5, but the underlying model is not confirmed.

The concept, explained simply: A proactive agent flips the usual chatbot pattern. Instead of waiting for your question, it watches your inbox and calendar, notices what needs doing, drafts or prepares the work, and then taps you on the shoulder. Think of the difference between a library you must visit and an assistant who leaves a briefing on your desk each morning.

Why it matters: Rivals are moving the same way: Meta announced Muse on September 8 as a personal agent for errands and shopping, and OpenAI unveiled dots on September 29 as persistent project agents, each with a cloud computer. Duet targets office work. As agents start acting between your prompts, the questions shift from 'how smart is it?' to 'how much can it do unsupervised, and who is accountable?'

2) Kimi K3.1 Pro: A Trillion-Parameter Model That Only Wakes a Few Experts

Moonshot AI released Kimi K3.1 Pro through its API on October 23, 2026, with a 1-million-token context window, and said the full weights would follow by November 4. Reports describe a mixture-of-experts design with about 2.8 trillion total and 104 billion active parameters, priced at $2 per million input tokens and $10 per million output tokens, roughly half the cost of Claude Opus 5.5 by one account. I could not confirm that the weights are actually live or what licence applies.

The concept, explained simply: A mixture-of-experts (MoE) model is like a hospital with many specialist doctors. A router sends each piece of text only to the few specialists best suited to it, so most of the model sits idle on any one step. 'Total parameters' counts every specialist; 'active parameters' counts only those working on the current token.

Why it matters: MoE is how labs get frontier-level knowledge at a fraction of the compute per token, which is why a 2.8T model can be priced like a much smaller one. Open weights let companies host the model themselves, keeping data in-house and avoiding vendor lock-in, though self-hosting a model this size needs serious hardware.

3) Guardrails by Design: Consent Gates and Filtered Training Data

Both stories above carry a safety design choice. Duet pauses for human approval before irreversible actions. Moonshot says it removed cyberattack-related material from training before the open release to limit what freely distributed weights can do. Meta's Muse, by contrast, drew privacy complaints in coverage, including claims it tried to read private messages without permission.

The concept, explained simply: A consent gate is a deliberate stop sign: the agent may prepare anything, but a human must click yes before money moves, a document is signed or data is deleted. Training-data filtering works earlier in the pipeline: if dangerous know-how is never learned, it cannot be extracted later, which matters most for open weights that cannot be recalled once published.

Why it matters: Open weights cannot be patched or switched off after release, so safety must be built in beforehand. Hosted agents can be gated at the action layer. Expect buyers to ask both questions: what can this agent do without asking, and what did the model learn in the first place? Filtering effectiveness is vendor-claimed and still needs independent testing.

VIRAL / OPEN-SOURCE SPOTLIGHT

Claude Duet (Anthropic): A proactive 'chief of staff' agent that triages your inbox and calendar, prepares replies and tasks, and messages you when something needs a decision, then waits for approval before anything irreversible. It is trending because it is the third big proactive-agent launch in six weeks, and because the pitch is relatable: work continues while you step away. Limits to know: US-only waitlist, checks accounts every few hours, and Anthropic itself says personal-errand skills like travel booking are not yet strong. I could not verify user numbers. Try it, once you have access, by giving it one narrow job such as 'draft replies to scheduling emails' and reviewing every approval request.

MARKET SIGNAL

The three big labs now each ship an always-on agent product, and competition is moving from chat quality to autonomy, trust and price. Open-weight models like Kimi K3.1 Pro at about half the price of a US flagship keep pressure on API margins. Expect differentiation on approval controls, privacy and cost per completed task rather than raw benchmarks. All figures are unverified secondary reports.

PRACTICAL TAKEAWAYS

1) Start proactive agents with read-only access.
Grant an agent inbox and calendar viewing first and widen permissions only after you trust its judgment.

2) Demand consent gates on irreversible actions.
Pick agents that stop for approval before payments, signatures, sending and deletion, and read each request instead of auto-approving.

3) Read 'active' not just 'total' parameters.
For MoE models, active parameters and per-token price predict cost and speed better than the headline size.

4) Price-test open-weight alternatives.
Run your own task set against a cheaper open model and compare cost per completed task before committing to one vendor.

5) Verify launch claims at the source.
Aggregator coverage conflicts on dates and specs, so confirm availability and pricing on the vendor's official page.

SELF-TEST

Q1. What makes an agent 'proactive'?

A1. It acts on your connected accounts and contacts you without waiting for a prompt, then asks approval for irreversible steps.

Q2. What is the difference between total and active parameters in a mixture-of-experts model?

A2. Total counts all experts; active counts only the experts used for the current token, which drives compute cost.

Q3. What is a consent gate?

A3. A required human approval before an agent takes an irreversible action like paying, signing or deleting.

Q4. Why is filtering training data especially important for open weights?

A4. Released weights cannot be recalled or patched, so risky knowledge must be kept out before release.

TOMORROW

What We're Watching: whether Kimi K3.1 Pro's open weights and licence are actually published. What We're Watching: how Duet's waitlist expands beyond the US and to Team plans. What We're Watching: independent tests of consent-gate reliability in proactive agents. What We're Watching: primary-source confirmation of today's secondary reports.

TOPICS COVERED SO FAR (RECAP)

Recent days covered the cheap fast model tier (GPT-6.1 Terra), the Agent Client Protocol in Android Studio, compute commitments in Anthropic's amended S-1, always-on Dots agents, staged model release and agent monitoring, agent swarms, Chrome DevTools MCP, orbital TPUs and the Hermes Agent surge. Today added proactive agents (Claude Duet, Meta Muse, OpenAI dots), mixture-of-experts open-weight models (Kimi K3.1 Pro), and consent gates plus training-data filtering as safety design.

Sources: letsdatascience.com, vktr.com and tech-insider.org (Claude Duet); axios.com, testingcatalog.com and techrepublic.com (Meta Muse); analyticsinsight.net (Duet vs Muse vs dots comparison); vals.ai, codersera.com and fourweekmba.com (Kimi K3.1 Pro). Figures are secondary reports and unverified.