

Claude Duet: Anthropic's Information Firewall for Agents

TechRounder PDF Edition

Live article: <https://www.techrounder.com/ai/claude-duet-proactive-agent-race/>

By Vipin PG | Published October 22, 2026 | Updated October 22, 2026 | Format: Deep Dive | 8 min read

In brief

Duet is the fourth always-on agent in ten weeks, not the third in seven. Its payment tokens and fixed prohibitions are structural; its approval gate inherits a 17% miss rate from Claude Code, not from Duet.

Anthropic's proactive agent borrows its approval gate from Claude Code, and the published miss rate on that gate is 17%. The stronger parts of Duet's safety design come from what it cannot reach.

Claude Duet opened as a waitlisted research preview on 21 October 2026, for Pro and Max subscribers with a US Claude account, on iOS, Android and claude.ai. It runs on Opus 5.5 in a cloud computer with its own browser, sends a morning brief, checks connected accounts at fixed intervals, and hands finished work back as a document, sheet, deck or dashboard. There is no extra charge, and it draws on the same five-hour and weekly limits as chat.

Most launch coverage described the product. This piece looks at what Anthropic built around it: an information-flow design made of data classes, brokered credentials and classifier-judged approvals. Against Anthropic's own published numbers, that design is stronger than the demo suggests in one place and weaker in another.

What Duet is

Three behaviours define the product. A morning brief pulls together what happened across connected apps. All-day monitoring messages you only when something needs a decision. Task handoff means you describe a job and get an artifact back while the work continues after you close the app or shut the laptop.

The interface follows that shape: one main thread plus optional topic channels that share memory, a Tasks tab for work in flight, and a Make tab where finished artifacts land. Anthropic says Duet reaches more than 6,000 connectors and plugins through the Claude Marketplace. The launch demo leaned on Gmail, Slack, QuickBooks, Stripe, Gusto, DocuSign, HubSpot, Canva and Shopify.

Anthropic built this for "professionals and small business owners with more on their plates than they can track," not for developers. That is a different buyer from the one Claude Code serves, and it explains several of the limits below.

Four always-on agents in ten weeks

Early coverage called Duet the third always-on agent in seven weeks. It is the fourth. xAI shipped Grok Bot on 11 August 2026, with always-on agents that have their own computer and return only when something needs approval. Meta shipped Muse on 8 September, a personal agent on a dedicated secure virtual machine that asks before sending email or making a purchase. OpenAI shipped dots on 29 September, with cloud computers and a browser, reachable through ChatGPT, Slack and Teams.

Duet arrived 71 days after Grok Bot and 22 days after dots. The "three in seven weeks" line survives only if you leave xAI out.

Four labs converged on the same shape within ten weeks: a cloud computer, a chat interface and an approval gate. That looks like a shared design consensus, so the differences in approval design are worth reading closely.

The information firewall

The announcement describes a dedicated classifier and safety system prompt that checks each message and guards against leaks of sensitive information. It defines two tiers.

Requests to share credit card numbers and government identification are rejected outright. Health information and personal finances are released only on explicit instruction. That is a data-class model, not a permission prompt, and it targets a real failure. An agent that reads email and sends messages constantly decides where facts go, and the usual problem is a helpful action that forwards the wrong detail to the wrong recipient.

Two things are missing. Anthropic does not say whether the classifier inspects outbound messages only or also the context it retrieves, and it publishes no false-positive or false-negative rate for this control. The announcement also describes the work as ongoing, promising stronger protections "in the coming days." Anyone writing this week is describing version one of a control the vendor is still tuning.

The approval gate is a classifier with a 17% miss rate

Duet's approvals run through the auto-mode classifier, which Anthropic describes as the same safeguard that protects Claude and Claude Code from taking steps you did not ask for. The engineering write-up from March 2026 explains how it works.

A server-side prompt-injection probe screens tool output before it enters context. A transcript classifier then judges each proposed action against block rules, an environment trust boundary and user intent. A fast single-token stage filters the easy cases, and only flagged actions get chain-of-thought review.

The numbers are unusually specific for a vendor. On 10,000 real tool calls the deployed pipeline produced a 0.4% false-positive rate. On 52 real over-eager actions it missed 17%. On 1,000 synthetic exfiltration attempts it missed 5.7%.

The 17% comes from cases where the classifier sees the action and wrongly concludes that an earlier consent signal covers its larger blast radius, so the judging model generalises consent.

Two design choices shape the failure modes. Assistant prose and tool results are stripped before classification, so the agent cannot talk its way past the gate. The same stripping means the classifier cannot tell whether a parameter came from your query or was fabricated. The backstop is escalation: three consecutive denials or twenty denials in total hands the decision to a human. That layer, not the classifier, catches a run of misses.

Anthropic publishes no Duet-specific evaluation, so applying 17% to Duet is an inference from a sibling deployment, not a measurement of this product. It is the best public number available.

Credentials Duet never sees

Duet follows the credential design used for Claude Managed Agents, so it never directly sees your passwords or sign-in tokens. The Help Center documentation is more concrete than the launch post: when a site asks for a password, you type it into a separate form on your device, and it goes straight to the browser, designed to stay out of your chat with Duet.

Payments are the stronger design. Purchases go through Link by Stripe or Shop Pay, each purchase uses a one-time card scoped to the approved amount, and your real card number is not shared. That is a payment-token architecture, not a prompt-level guard, so it holds even if the model is fully compromised.

There are two caveats. The 1Password integration is listed as coming soon, so the password-manager tie-in that commenters immediately treated as a bad idea is not live. Several actions are also prohibited even when you approve them: trading securities, moving money between accounts, and accessing someone else's account. Those fixed prohibitions bound the blast radius more tightly than any classifier.

Prompt injection, in Anthropic's own words

The Help Center says it directly: Duet reads content written by other people, such as email and web pages, and that content can contain instructions meant to mislead it. It adds that no defence is perfect, which is one reason consequential actions need approval.

The structural defence has two layers. The input probe warns the agent when tool output looks like hijack content, and the output classifier makes the final call on actions, so an injection has to beat both independently. Anthropic points to its browser-use research as the source of this work, and its Opus 5.5 page says the model ties Table 5.1 for the lowest prompt-injection success rate in a Gray Swan benchmark.

An agent that reads inbound email and can send outbound email is a two-way channel, so a successful injection is an exfiltration path, not just a bad answer. The structural mitigations are narrow: no phone calls, no texts to other people, email only with approval, and payment only through a scoped one-time card.

The catches, consolidated

Access is narrower than "available today." You need Pro or Max, a US account and an invitation, with waves and no guaranteed date. Team and Enterprise are excluded for now, and the waitlist is bound to the email you submit, which must match both your Claude account and your Apple or Google sign-in on the phone.

Usage is the sharpest practical catch. Duet draws on the same five-hour and weekly limits at the same rate as chats, background checks count against them, and Anthropic warns you may reach your limits sooner than with chat alone. Admission includes one full limit reset covering both windows, available until 30 November 2026.

The capability limits shape what you can delegate. There is no model choice, no watching or taking over the browser, and checks run at fixed intervals, so urgent items can slip between them. Duet cannot see all your projects or any local files, and its access to existing Claude work is read-only. Microsoft 365 is not available yet, Slack works with limits, and Link and Shop Pay exist inside Duet rather than in the connectors directory.

For anyone thinking about withdrawal: disconnecting an app stops future access, but memory of information from earlier conversations is kept. Anthropic also concedes that Duet may misread a thread or miscalibrate urgency, and that personal errands such as booking tickets are weaker than work tasks.

A setup and revocation order that limits the damage

Connect in ascending order of risk. Calendar and Drive go first, for context. Gmail comes next, if you want the morning brief. Slack goes last, because it is both read and write and is explicitly limited in Duet.

Money connectors go in only alongside the wallet path, not a stored card: Stripe for invoice reads, with Link or Shop Pay handling any purchase as a scoped one-time card. Keep health and finance out of the first week, and treat the explicit-instruction gate as a design claim to test.

Review the two surfaces that make behaviour legible: the Tasks tab for what needs input, and the provenance line where Duet names the email, file or spreadsheet behind a number. Watch for the first refusal. A classifier you have seen decline something tells you more than one described in a launch post.

Revoke in three places, because they are separate. Disconnect the app in Claude settings, turn off Search and reference chats under Settings, Memory to stop past-chat search, and remove the wallet connection inside Duet.

What is still unproven

Four questions remain open. Does the sensitive-information classifier inspect only outbound messages or also retrieved context, and what are its error rates? Nothing is published. Has anyone outside Anthropic tested Duet's security? The vendor cites red teams, an external penetration test and HackerOne coverage, but no independent report exists, and the bounty scope is Anthropic-wide rather than Duet-specific. What happens to price and limits after the preview, given the reset gift expires on 30 November? And will background runs get connector parity, starting with Microsoft 365?

One claim circulating in early coverage does not hold up: that scheduled runs cannot use custom connectors. The documented limits are Microsoft 365 unavailable, Slack limited, fixed-interval checks and read-only access to Claude work. The stronger version lacks a primary source and should be treated as unverified.

The Hacker News thread on launch, at 341 points and 303 comments when read, contained the sharpest scepticism: that these agents are popular without an obvious use case, and that they force relevance out of whatever context they can reach. A dots user in the same thread offered the fairest comparison evidence available, praising channel monitoring and finding coding poor. The category's value looks narrow and channel-shaped.

Duet's safety story is real but unevenly distributed. The payment token and the fixed prohibitions are structural, and they hold regardless of what the model decides. The approval gate is probabilistic, inherits a 17% miss rate from a different deployment, and has no published Duet evaluation behind it. The information firewall is the most interesting control and the least documented. Until Anthropic publishes its error rates, the honest reading is that Duet is safer than an agent with no gate, and less predictable than the phrase "checks with you first" implies.

References

1. x.ai - news / introducing-grok-bot - <https://x.ai/news/introducing-grok-bot>

2. about.fb.com - news / 2026 - <https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/>
3. openai.com - index / introducing-dots - <https://openai.com/index/introducing-dots/>
4. claude.com - resources / articles - <https://claude.com/resources/articles/claude-duet>
5. anthropic.com - engineering / claude-code-auto-mode - <https://www.anthropic.com/engineering/claude-code-auto-mode>
6. support.claude.com - en / articles - <https://support.claude.com/en/articles/17631495-get-started-with-claude-duet>
7. anthropic.com - research / prompt-injection-defenses - <https://www.anthropic.com/research/prompt-injection-defenses>
8. anthropic.com - claude-opus-5-5 - <https://www.anthropic.com/claude-opus-5-5>
9. hackerone.com - anthropic - <https://hackerone.com/anthropic>
10. news.ycombinator.com - item - <https://news.ycombinator.com/item?id=50186037>