

Anthropic releases Claude Sonnet 5, built for autonomous agent work

Source copy kept by wheresthe.ai for the story "Anthropic releases Claude Sonnet 5, built for autonomous agent work" (<https://wheresthe.ai/d/anthropic-sonnet-5-release-2026-06/>).

Source: <https://www.anthropic.com/news/claude-sonnet-5>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 9f398fb5986f643200d8cc21d4dc03d61920c3585ac32e697b0b3c45d6ae9f68

Documents the source offers: <https://www-cdn.anthropic.com/9e6a1044980d8c4ed85669faf9c2a8342e2e9f1e/Claude%20Sonnet%205%20System%20Card.pdf>

Product Introducing Claude Sonnet 5 Jun 30, 2026

Claude Sonnet 5 is built to be the most agentic Sonnet model yet. It can make plans, use tools like browsers and terminals, and run autonomously at a level that, just a few months ago, required larger and more expensive models. For many developers, the agentic AI era began with Sonnet-class models: Claude Sonnet 3.5, 3.6, and 3.7 were the first models that showed impressive skills in coding and tool use. More recently, though, the clearest gains in agentic capabilities have been in our Opus-class models. Sonnet 5 narrows the gap: its performance is close to that of Opus 4.8, but at lower prices. It's a substantial improvement over its predecessor, Sonnet 4.6, on important aspects of agentic performance like reasoning, tool use, coding, and knowledge work: Scores for Sonnet 5 on a variety of evaluations compared to those of Sonnet 4.6 and Opus 4.8 (a more generally capable model, for reference). The Claude Sonnet 5 System Card reports a broader set of evaluations in detail. Our safety assessments found that Sonnet 5 shows an overall lower rate of undesirable behaviors than Sonnet 4.6, and is generally safer to use in agentic contexts. Evaluations also show that it has a much lower ability to perform cybersecurity tasks than our current Opus models. From today, Claude Sonnet 5 is available across all plans: it is the default model for Free and Pro plans, and is available to Max, Team, and Enterprise users. It is priced at \$2 per million input tokens and \$10 per million output tokens. Developers can use claude-sonnet-5 via the Claude API . Working with Claude Sonnet 5 The charts below compare the performance of Sonnet 5 with Sonnet 4.6 and Opus 4.8 at different effort levels on the agentic search evaluation BrowseComp and the computer use evaluation OSWorld-Verified . Sonnet 5 (orange line) is a strict improvement over Sonnet 4.6 (gray line) and covers a much wider range of cost-performance options than Opus 4.8 (yellow line). It provides substantially improved cost efficiency at medium effort; its higher-effort performance can match Opus 4.8 on some tasks. Between Sonnet 5 and Opus 4.8, users can adjust the effort level to find the right balance of cost and performance. Agentic search Agentic computer use Cost-performance curves at different effort levels. The previous best Sonnet model (Sonnet 4.6) fell well short of Opus 4.8. Sonnet 5 offers a wider range of cost-performance options than Sonnet 4.6, and in some cases matches Opus 4.8's capability levels. The charts show Sonnet 5 priced at \$3 per million input tokens and \$15 per million output tokens (existing standard pricing). Sonnet 5's introductory pricing of \$2/MTok input and \$10/MTok output has since been made permanent, so its actual cost is lower than shown here. Opus 4.8 is priced at \$5/MTok input and \$25/MTok output. xhigh = extra high effort level. Cost-performance curves at different effort levels. The previous best Sonnet model (Sonnet 4.6) fell well short of Opus 4.8. Sonnet 5 offers a wider range of cost-performance options than Sonnet 4.6, and in some cases matches Opus 4.8's capability levels. The charts show Sonnet 5 priced at \$3 per million input tokens and \$15 per million output tokens (existing standard pricing). Sonnet 5's introductory pricing of \$2/MTok input and \$10/MTok output has since been made permanent, so its actual cost is lower than shown here. Opus 4.8 is priced at \$5/MTok input and \$25/MTok output. xhigh = extra high effort level.

Feedback from our early access partners has been consistent: Sonnet 5 is much more agentic than its predecessors. Testers described how it finishes complex tasks where previous Sonnet models would stop short, how it checks its own output without explicitly being asked, and how it does all this agentic work at an attractive price point:

Claude Sonnet 5 gives our agents a strong execution layer for multi-step software engineering work. It handles sustained coding, tool use, and debugging well across messy technical contexts, and has been especially useful for workflows where follow-through and technical grounding matter.

We handed Claude Sonnet 5 a two-part job—update Salesforce account tiers, send a launch announcement to

enterprise contacts—and it finished end to end. That used to stall halfway. For day-to-day automation, it's a no-brainer.

Claude Sonnet 5 gets more done with less. Same output quality, fewer steps to get there. It refuses unsafe requests cleanly and consistently, too. At Lovable, we're putting powerful tools in the hands of millions of builders. A model that knows when to say no is just as important as one that knows how to build.

We ran Claude Sonnet 5 against dozens of our most challenging real pull requests, and it carried each one through to a tested, verified result on its own — freeing our engineers to focus on the judgment, the decision, and the final sign-off.

I asked Claude Sonnet 5 to investigate a bug. Unprompted, it wrote a reproducing test, implemented the fix, then stashed it to confirm the bug came back without the change. All in a single pass.

With Claude Sonnet 5, agents stay on plan, follow our conventions, and ship clean multi-step changes, all at an efficient cost.

Claude Sonnet 5 is at its best on brownfield code—race conditions, hidden tests, the parts nobody wants to touch. It traces a failure to its actual root cause and ships a durable fix instead of patching the symptom.

Claude Sonnet 5 sits on the Pareto frontier for Eve's plaintiff-law tasks. We see the clearest gains in legal research and analysis, at a price-to-performance ratio that made the choice to migrate easy.

ClickHouse agents explore live data and produce insights on the fly, so time-to-insight matters when testing new models. Claude Sonnet 5 reasons in tighter steps and gets our users to answers noticeably faster. That speed is a difference our customers feel.

At Pace, our computer-use agents run insurance workflows—submission intake, FNOL, loss runs—on the systems our operations teams already use. Claude Sonnet 5 consistently takes the right action and does it quickly, which is what real insurance work demands.

For enterprise teams managing high-volume, complex workloads, Claude Sonnet 5 represents a genuine step forward — strong performance where it counts, with the speed and cost profile that makes scaling practical. On several complex tasks it exceeds the current frontier, while delivering fast responses at low costs. For enterprises running at scale, that's a real operational win.

Claude Sonnet 5 handled the full range of coding tasks we tested it on, while resolving more issues. It's a meaningful improvement to both quality and efficiency.

Claude Sonnet 5 is a strong agentic-coding model, delivering top-tier accuracy comparable to Opus-class models and a clear step-function improvement over Sonnet 4.6. It conducts thorough explorations and sustains focus noticeably longer on complex tasks.

01 / 13

Safety evaluations Our pre-deployment safety evaluations found that Sonnet 5 was overall an improvement on Sonnet 4.6. On agentic safety, the model is better at refusing malicious requests and resisting hijack attempts in prompt injection attacks. The model shows lower rates of hallucination and sycophancy than Sonnet 4.6. On our automated behavioral audit, which tests a wide range of misaligned behaviors such as cooperation with misuse and deception, Sonnet 5 scored lower (that is, safer) overall. However, it did show somewhat higher rates of misaligned behavior on this assessment compared to the more capable Opus 4.8 and Claude Mythos Preview. Rates of misaligned behavior on our automated behavioral audit, which tests for a very wide range of undesirable behaviors across many situations and contexts (see Section 6.4 of the Sonnet 5 System Card for a complete list and results for each specific behavior). Sonnet 5 shows an overall lower rate of misaligned behavior than Sonnet 4.6, though a higher rate than Mythos Preview and Opus 4.8. We did not deliberately train Sonnet 5 on cybersecurity tasks. It can perform some routine, non-harmful cyber tasks, but on evaluations testing potentially dangerous cyber skills, such as developing software exploits, it shows substantially poorer performance than models such as Opus 4.8 and Mythos 5. Scores from one evaluation, which tested models' ability to develop exploits for vulnerabilities in the Firefox browser, are shown in the chart below. Sonnet 5 was never able to develop a full working exploit, but it does show a slightly higher rate of partial success than Sonnet 4.6. This latter change is likely due to improvements in general intelligence rather than specific training. Scores measuring models' success at developing exploits for software vulnerabilities in Firefox 147 (this evaluation was developed in

collaboration with Mozilla ; all vulnerabilities have been patched in Firefox 148). For each model, the left-hand bar shows how often the model (without safeguards) developed a working exploit; the right-hand bar shows how often the model had partial success. Neither of the Sonnet models could successfully develop a working exploit (both scored 0.0%); Sonnet 5 showed a slightly higher partial success rate than Sonnet 4.6. Both Sonnet models have substantially poorer cyber capabilities than Opus 4.8 and Mythos 5. For full details, see Section 3.2.4 of the Sonnet 5 System Card . Since Sonnet 5 is somewhat stronger than its predecessor on these tasks, we've launched it with cyber safeguards enabled by default. These safeguards—which detect and block dangerous cyber usage in real time—are the same as those present in Claude Opus 4.7 and 4.8 (because we judged that the overall level of cybersecurity risk from Sonnet 5 was low, the safeguards are less strict than those launched with Fable 5, which block a much wider range of cybersecurity tasks). 1 Our full assessment of Sonnet 5 across many safety and capability evaluations is reported in the Claude Sonnet 5 System Card . Availability and pricing Claude Sonnet 5 is available everywhere today at \$2 per million input tokens and \$10 per million output tokens 2 . We've increased rate limits across Chat, Cowork, Claude Code, and the Claude Platform 3 to accommodate the higher token usage of higher effort levels; users can select whichever level makes sense for their particular project.

Changelog Edit August 10, 2026: Sonnet 5's introductory pricing of \$2 per million input tokens and \$10 per million output tokens is now permanent. The standard pricing of \$3 input / \$15 output previously set to take effect September 1 no longer applies. The pricing references in this post have been updated accordingly.

Edit June 30, 2026: In the original version of this post, we included a cost-performance chart for the BrowseComp evaluation that was based on data from a simpler methodology that did not reflect the standard methodology we use for agentic search evaluations. This had the result of underestimating Sonnet 5's performance on the evaluation. We have now updated the chart so that it matches the methodology that we used and discussed in the Sonnet 5 system card (which used a 10M token budget with compaction and programmatic tool calling). We have also updated the surrounding text.

Footnotes 1 Sonnet 5 is part of our Cyber Verification Program , which is available today on the native Claude Platform, the Claude Platform on AWS, and Claude in Microsoft Foundry (hosted on Azure and Anthropic), and coming soon on Claude in Google Vertex. Organizations that are already enrolled in the Cyber Verification Program automatically have the same access on Sonnet 5, with no need to reapply. Overall, we recommend Claude Opus 4.8 for cybersecurity work that requires reduced guardrails. 2 Sonnet 5 is an upgrade to Sonnet 4.6, but it uses an updated tokenizer that changes how the model processes text to improve performance (this is similar to the tokenizer change we introduced with Claude Opus 4.7). The tradeoff is that the same input can map to more tokens: roughly 1.0–1.35× depending on the content type. 3 On April 26, 2026, we raised Sonnet and Haiku rate limits at every usage tier and simplified to three tiers (Start, Build, and Scale) on the native Claude Platform. You can view your tier and current limits in the Claude Console or read the documentation to learn more. Humanity's Last Exam: We updated the grader model for Humanity's Last Exam and have updated the Sonnet 4.6 score to 34.6% (no tools) and 46.8% (with tools). This is the reason the score differs from that reported in the Sonnet 4.6 launch blog . OSWorld-Verified: We made changes to how we run the OSWorld-Verified evaluation to more accurately reflect the model's performance in the real world, and have updated the Sonnet 4.6 score to 78.5%. This is the reason the score differs from that reported in the Sonnet 4.6 launch blog .

Related content

An update on independent evaluators To evaluate the safety of our systems, Anthropic is giving employee-like access to a team from METR, a nonprofit that evaluates frontier AI models, and we're in early discussions with several other organizations, including Hugging Face. Read more Preparing for the 2026 US midterms Earlier this year, we shared an update on our election safeguards. Here, we outline how we've continued to strengthen these safeguards ahead of this year's US midterms. Read more Anthropic opens Singapore office and signs MOU with the Government of Singapore Anthropic has opened an office in Singapore and signed an MOU with the country's government to advance safe, responsible AI adoption. Read more