

Anthropic releases Claude Opus 4 and Sonnet 4 and makes Claude Code generally available

Source copy kept by wheresthe.ai for the story "Anthropic releases Claude Opus 4 and Sonnet 4 and makes Claude Code generally available" (<https://wheresthe.ai/d/anthropic-claude-4-release-2025-05/>).

Source: <https://www.anthropic.com/news/claude-4>

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

SHA-256 of the saved text: b4ff748a896bf36e843fb42d0218ec2ce3c923c39779b7b6e7ffd823fbb030cd

Announcements Introducing Claude 4 May 22, 2025

Today, we're introducing the next generation of Claude models: Claude Opus 4 and Claude Sonnet 4, setting new standards for coding, advanced reasoning, and AI agents. Claude Opus 4 is the world's best coding model, with sustained performance on complex, long-running tasks and agent workflows. Claude Sonnet 4 is a significant upgrade to Claude Sonnet 3.7, delivering superior coding and reasoning while responding more precisely to your instructions.

Alongside the models, we're also announcing: Extended thinking with tool use (beta) : Both models can use tools—like web search—during extended thinking, allowing Claude to alternate between reasoning and tool use to improve responses. New model capabilities : Both models can use tools in parallel, follow instructions more precisely, and—when given access to local files by developers—demonstrate significantly improved memory capabilities, extracting and saving key facts to maintain continuity and build tacit knowledge over time. Claude Code is now generally available : After receiving extensive positive feedback during our research preview, we're expanding how developers can collaborate with Claude. Claude Code now supports background tasks via GitHub Actions and native integrations with VS Code and JetBrains, displaying edits directly in your files for seamless pair programming. New API capabilities: We're releasing four new capabilities on our API that enable developers to build more powerful AI agents: the code execution tool, MCP connector, Files API, and the ability to cache prompts for up to one hour. Claude Opus 4 and Sonnet 4 are hybrid models offering two modes: near-instant responses and extended thinking for deeper reasoning. The Pro, Max, Team, and Enterprise Claude plans include both models and extended thinking, with Sonnet 4 also available to free users. Both models are available on our API, Amazon Bedrock, and Google Cloud's Vertex AI. Pricing remains consistent with previous Opus and Sonnet models: Opus 4 at \$15/\$75 per million tokens (input/output) and Sonnet 4 at \$3/\$15. Claude 4 Claude Opus 4 is our most powerful model yet and the best coding model in the world, leading on SWE-bench (72.5%) and Terminal-bench (43.2%). It delivers sustained performance on long-running tasks that require focused effort and thousands of steps, with the ability to work continuously for several hours—dramatically outperforming all Sonnet models and significantly expanding what AI agents can accomplish. Claude Opus 4 excels at coding and complex problem-solving, powering frontier agent products. Cursor calls it state-of-the-art for coding and a leap forward in complex codebase understanding. Replit reports improved precision and dramatic advancements for complex changes across multiple files. Block calls it the first model to boost code quality during editing and debugging in its agent, codename goose, while maintaining full performance and reliability. Rakuten validated its capabilities with a demanding open-source refactor running independently for 7 hours with sustained performance. Cognition notes Opus 4 excels at solving complex challenges that other models can't, successfully handling critical actions that previous models have missed. Claude Sonnet 4 significantly improves on Sonnet 3.7's industry-leading capabilities, excelling in coding with a state-of-the-art 72.7% on SWE-bench. The model balances performance and efficiency for internal and external use cases, with enhanced steerability for greater control over implementations. While not matching Opus 4 in most domains, it delivers an optimal mix of capability and practicality. GitHub says Claude Sonnet 4 soars in agentic scenarios and will introduce it as the model powering the new coding agent in GitHub Copilot. Manus highlights its improvements in following complex instructions, clear reasoning, and aesthetic outputs. iGent reports Sonnet 4 excels at autonomous multi-feature app development, as well as substantially improved problem-solving and codebase navigation—reducing navigation errors from 20% to near zero. Sourcegraph says the model shows promise as a substantial leap in software development—staying on track longer, understanding problems more deeply, and providing more elegant code quality. Augment Code reports higher success rates, more surgical code edits, and more careful work through complex tasks, making it the top choice for their primary model. These models advance our customers' AI strategies across the board: Opus 4 pushes boundaries in coding, research, writing, and scientific discovery, while Sonnet 4 brings frontier

performance to everyday use cases as an instant upgrade from Sonnet 3.7. Claude 4 models lead on SWE-bench Verified, a benchmark for performance on real software engineering tasks. See appendix for more on methodology.

Model improvements In addition to extended thinking with tool use, parallel tool execution, and memory improvements, we've significantly reduced behavior where the models use shortcuts or loopholes to complete tasks. Both models are 65% less likely to engage in this behavior than Sonnet 3.7 on agentic tasks that are particularly susceptible to shortcuts and loopholes. Claude Opus 4 also dramatically outperforms all previous models on memory capabilities. When developers build applications that provide Claude local file access, Opus 4 becomes skilled at creating and maintaining 'memory files' to store key information. This unlocks better long-term task awareness, coherence, and performance on agent tasks—like Opus 4 creating a 'Navigation Guide' while playing Pokémon. **Memory:** When given access to local files, Claude Opus 4 records key information to help improve its game play. The notes depicted above are real notes taken by Opus 4 while playing Pokémon. Finally, we've introduced thinking summaries for Claude 4 models that use a smaller model to condense lengthy thought processes. This summarization is only needed about 5% of the time—most thought processes are short enough to display in full. Users requiring raw chains of thought for advanced prompt engineering can contact sales about our new Developer Mode to retain full access. **Claude Code** Claude Code, now generally available, brings the power of Claude to more of your development workflow—in the terminal, your favorite IDEs, and running in the background with the Claude Code SDK. New beta extensions for VS Code and JetBrains integrate Claude Code directly into your IDE. Claude's proposed edits appear inline in your files, streamlining review and tracking within the familiar editor interface. Simply run Claude Code in your IDE terminal to install. Beyond the IDE, we're releasing an extensible Claude Code SDK, so you can build your own agents and applications using the same core agent as Claude Code. We're also releasing an example of what's possible with the SDK: Claude Code on GitHub, now in beta. Tag Claude Code on PRs to respond to reviewer feedback, fix CI errors, or modify code. To install, run `/install-github-app` from within Claude Code.

Getting started These models are a large step toward the virtual collaborator—maintaining full context, sustaining focus on longer projects, and driving transformational impact. They come with extensive testing and evaluation to minimize risk and maximize safety, including implementing measures for higher AI Safety Levels like ASL-3. We're excited to see what you'll create. Get started today on Claude , Claude Code , or the platform of your choice. As always, your feedback helps us improve.

Appendix

Performance benchmark data sources Open AI: o3 launch post , o3 system card , GPT-4.1 launch post , GPT-4.1 hosted evals Gemini: Gemini 2.5 Pro Preview model card Claude: Claude 3.7 Sonnet launch post **Performance benchmark reporting** Claude Opus 4 and Sonnet 4 are hybrid reasoning models. The benchmarks reported in this blog post show the highest scores achieved with or without extended thinking. We've noted below for each result whether extended thinking was used:

No extended thinking: SWE-bench Verified, Terminal-bench Extended thinking (up to 64K tokens): TAU-bench (no results w/o extended thinking reported) GPQA Diamond (w/o extended thinking: Opus 4 scores 74.9% and Sonnet 4 is 70.0%) MMMLU (w/o extended thinking: Opus 4 scores 87.4% and Sonnet 4 is 85.4%) MMMU (w/o extended thinking: Opus 4 scores 73.7% and Sonnet 4 is 72.6%) AIME (w/o extended thinking: Opus 4 scores 33.9% and Sonnet 4 is 33.1%)

TAU-bench methodology Scores were achieved with a prompt addendum to both the Airline and Retail Agent Policy instructing Claude to better leverage its reasoning abilities while using extended thinking with tool use. The model is encouraged to write down its thoughts as it solves the problem distinct from our usual thinking mode, during the multi-turn trajectories to best leverage its reasoning abilities. To accommodate the additional steps Claude incurs by utilizing more thinking, the maximum number of steps (counted by model completions) was increased from 30 to 100 (most trajectories completed under 30 steps with only one trajectory reaching above 50 steps). **SWE-bench methodology** For the Claude 4 family of models, we continue to use the same simple scaffold that equips the model with solely the two tools described in our prior releases here —a bash tool, and a file editing tool that operates via string replacements. We no longer include the third 'planning tool' used by Claude 3.7 Sonnet. On all Claude 4 models, we report scores out of the full 500 problems. Scores for OpenAI models are reported out of a 477 problem subset .

For our "high compute" numbers we adopt additional complexity and parallel test-time compute as follows: We

sample multiple parallel attempts. We discard patches that break the visible regression tests in the repository, similar to the rejection sampling approach adopted by Agentless (Xia et al. 2024) ; note no hidden test information is used. We then use an internal scoring model to select the best candidate from the remaining attempts. This results in a score of 79.4% and 80.2% for Opus 4 and Sonnet 4 respectively.

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](#)