Announcements

Introducing Claude Sonnet 4.5

Sep 29, 2025 • 5 min read

Claude Sonnet 4.5 is the best coding model in the world. It's the strongest model for building complex agents. It's the best model at using computers. And it shows substantial gains in reasoning and math.

Code is everywhere. It runs every application, spreadsheet, and software tool you use. Being able to use those tools and reason through hard problems is how modern work gets done.

Claude Sonnet 4.5 makes this possible. We're releasing it along with a set of major upgrades to our products. In Claude Code, we've added checkpoints—one of our most requested features—that save your progress and allow you to roll back instantly to a previous state. We've refreshed the terminal interface and shipped a native VS Code extension. We've added a new context editing

feature and memory tool to the Claude API that lets agents run even longer and handle even greater complexity. In the Claude apps, we've brought code execution and file creation (spreadsheets, slides, and documents) directly into the conversation. And we've made the Claude for Chrome extension available to Max users who joined the waitlist last month.

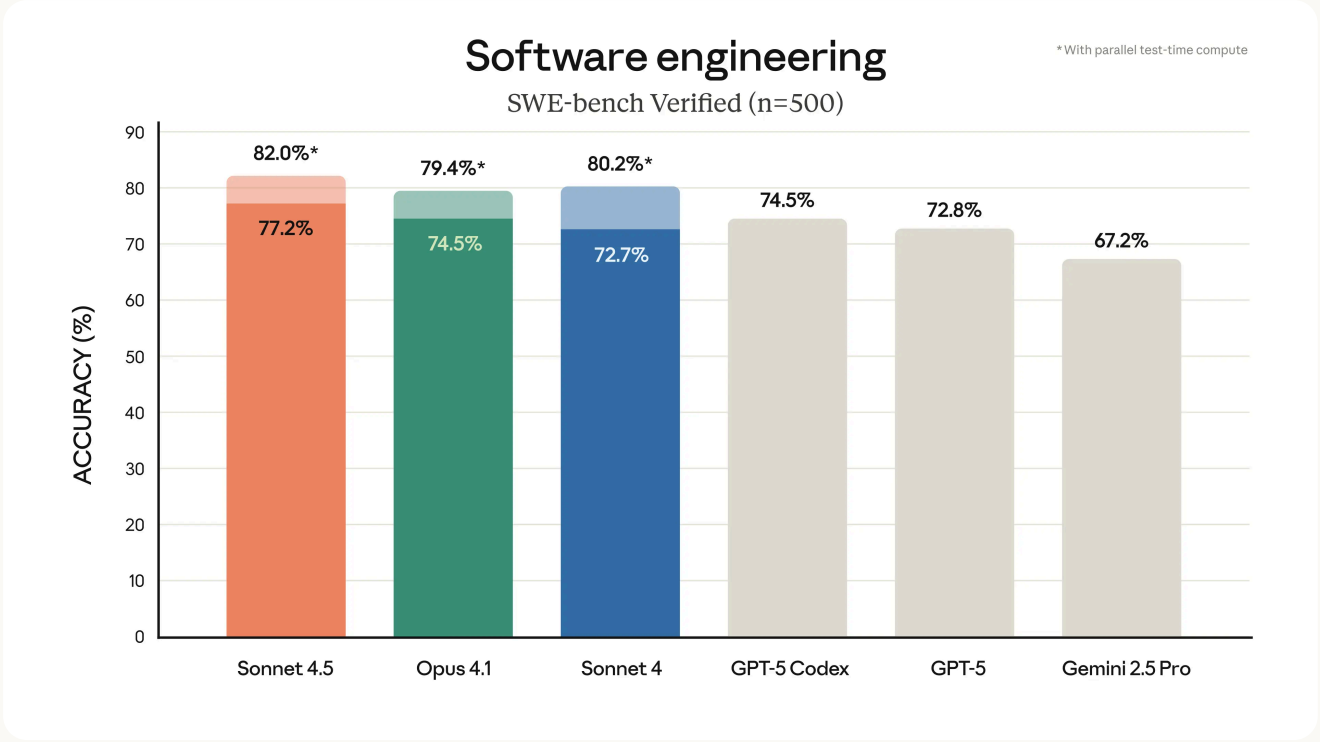
We're also giving developers the building blocks we use ourselves to make Claude Code. We're calling this the Claude Agent SDK. The infrastructure that powers our frontier products—and allows them to reach their full potential—is now yours to build with.

This is the most aligned frontier model we've ever released, showing large improvements across several areas of alignment compared to previous Claude models.

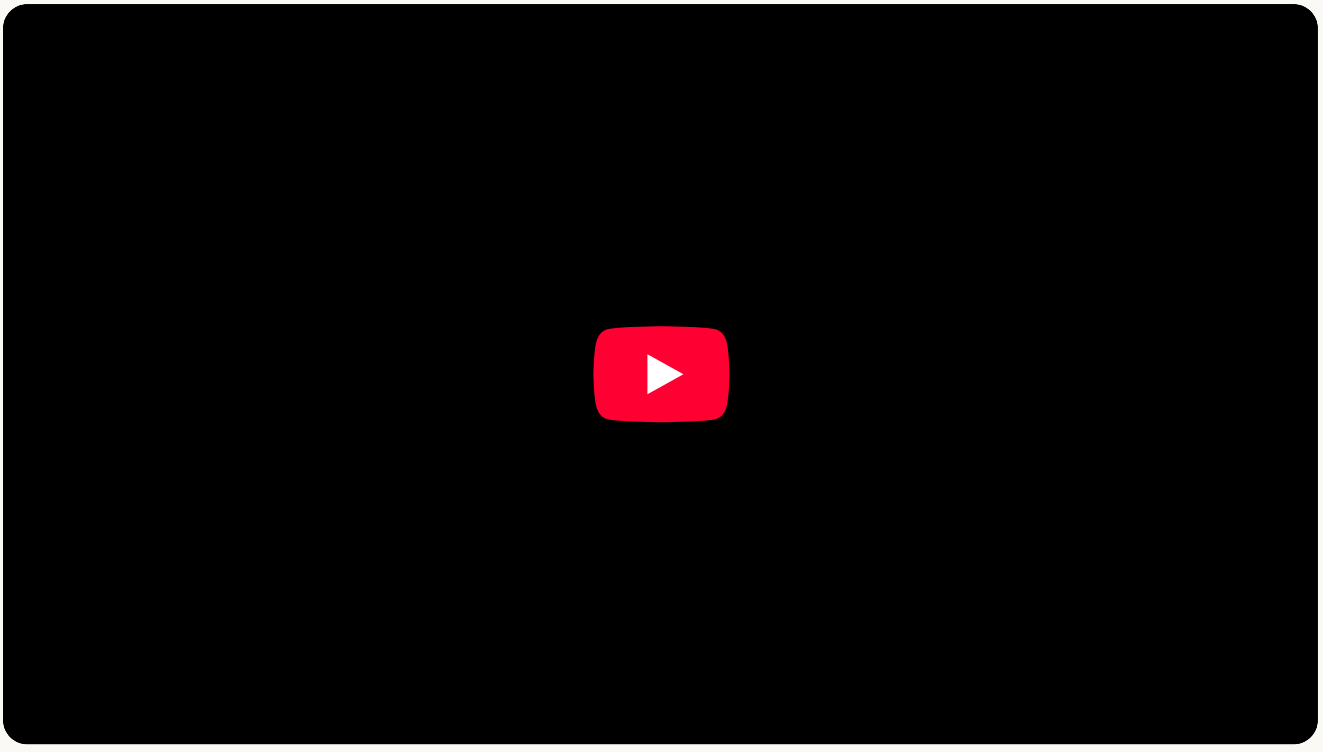
Claude Sonnet 4.5 is available everywhere today. If you're a developer, simply use `claude-sonnet-4-5` via the Claude API. Pricing remains the same as Claude Sonnet 4, at $3/$15 per million tokens.

Frontier intelligence

Claude Sonnet 4.5 is state-of-the-art on the SWE-bench Verified evaluation, which measures real-world software coding abilities. Practically speaking, we've observed it maintaining focus for more than 30 hours on complex, multi-step tasks.



Claude Sonnet 4.5 represents a significant leap forward on computer use. On OSWorld, a benchmark that tests AI models on real-world computer tasks, Sonnet 4.5 now leads at 61.4%. Just four months ago, Sonnet 4 held the lead at 42.2%. Our Claude for Chrome extension puts these upgraded capabilities to use. In the demo below, we show Claude working directly in a browser, navigating sites, filling spreadsheets, and completing tasks.



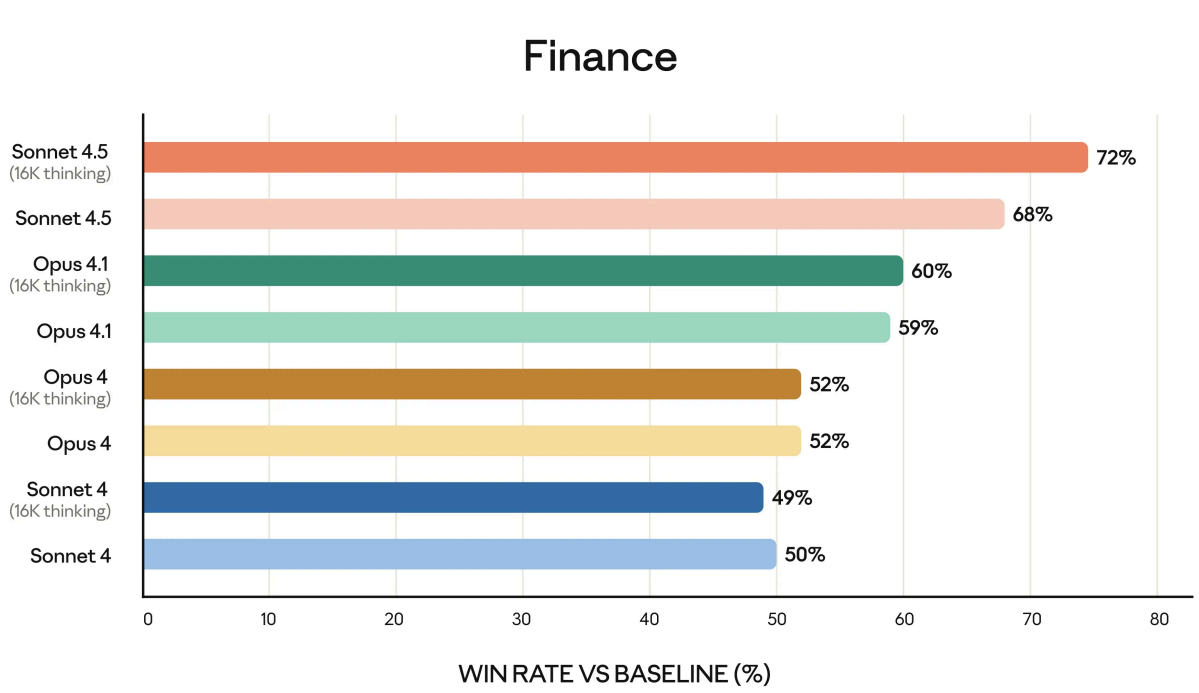
The model also shows improved capabilities on a broad range of evaluations including reasoning and math:

	Claude Sonnet 4.5	Claude Opus 4.1	Claude Sonnet 4	GPT-5	Gemini 2.5 Pro
Agentic coding *SWE-bench Verified*	**77.2%** **82.0%** with parallel test-time compute	74.5% 79.4% with parallel test-time compute	72.7% 80.2% with parallel test-time compute	72.8% GPT-5 74.5% GPT-5-Codex	67.2%
Agentic terminal coding *Terminal-Bench*	**50.0%**	46.5%	36.4%	43.8%	25.3%
Agentic tool use *τ2-bench*	Retail **86.2%** Airline **70.0%** Telecom **98.0%**	Retail **86.8%** Airline 63.0% Telecom 71.5%	Retail 83.8% Airline 63.0% Telecom 49.6%	Retail 81.1% Airline 62.6% Telecom 96.7%	— — —
Computer use *OSWorld*	**61.4%**	44.4%	42.2%	—	—
High school math competition *AIME 2025*	100% (python) 87.0% (no tools)	78.0%	70.5%	99.6% (python) 94.6% (no tools)	88.0%
Graduate-level reasoning *GPQA Diamond*	83.4%	81.0%	76.1%	85.7%	**86.4%**
Multilingual Q&A *MMMLU*	89.1%	**89.5%**	86.5%	89.4%	—
Visual reasoning *MMMU (validation)*	77.8%	77.1%	74.4%	**84.2%**	82.0%
Financial analysis *Finance Agent*	**55.3%**	50.9%	44.5%	46.9%	29.4%

Claude Sonnet 4.5 is our most powerful model to date. See footnotes for methodology.

Experts in finance, law, medicine, and STEM found Sonnet 4.5 shows dramatically better domain-specific knowledge and reasoning compared to older models, including Opus 4.1.

Finance | Law | Medicine | STEM



The model's capabilities are also reflected in the experiences of early customers:

"

We're seeing state-of-the-art coding performance from Claude Sonnet 4.5, with significant improvements on longer horizon tasks. It reinforces why many developers using Cursor choose Claude for solving their most complex problems.

"

Claude Sonnet 4.5 amplifies GitHub Copilot's core strengths. Our initial evals show significant improvements in multi-step reasoning and code comprehension—enabling Copilot's agentic experiences to handle complex, codebase-spanning tasks better.

Michael Truell
CEO

Mario Rodriguez
Chief Product Officer

Our most aligned model yet

As well as being our most capable model, Claude Sonnet 4.5 is our most aligned frontier model yet. Claude's improved capabilities and our extensive safety training have allowed us to substantially improve the model's behavior, reducing concerning behaviors like sycophancy, deception, power-seeking, and the tendency to encourage delusional thinking. For the model's agentic and computer use capabilities, we've also made considerable progress on defending against prompt injection attacks, one of the most serious risks for users of these capabilities.

You can read a detailed set of safety and alignment evaluations, which for the first time includes tests using techniques from mechanistic interpretability, in the Claude Sonnet 4.5 system card.

Overall misaligned behavior scores from an automated behavioral auditor (lower is better). Misaligned behaviors include (but are not limited to) deception, sycophancy, power-seeking, encouragement of

delusions, and compliance with harmful system prompts. More details can be found in the Claude Sonnet 4.5 system card.

Claude Sonnet 4.5 is being released under our AI Safety Level 3 (ASL-3) protections, as per our framework that matches model capabilities with appropriate safeguards. These safeguards include filters called classifiers that aim to detect potentially dangerous inputs and outputs—in particular those related to chemical, biological, radiological, and nuclear (CBRN) weapons.

These classifiers might sometimes inadvertently flag normal content. We've made it easy for users to continue any interrupted conversations with Sonnet 4, a model that poses a lower CBRN risk. We've already made significant progress in reducing these false positives, reducing them by a factor of ten since we originally described them, and a factor of two since Claude Opus 4 was released in May. We're continuing to make progress in making the classifiers more discerning[1].

The Claude Agent SDK

We've spent more than six months shipping updates to Claude Code, so we know what it takes to build and design AI agents. We've solved hard problems: how agents should manage memory across long-running tasks, how to handle permission systems that balance autonomy with user control, and how to coordinate subagents working toward a shared goal.

Now we're making all of this available to you. The Claude Agent SDK is the same infrastructure that powers Claude Code, but it shows impressive benefits for a very wide variety of tasks, not just coding. As of today, you can use it to build your own agents.

We built Claude Code because the tool we wanted didn't exist yet. The Agent SDK gives you the same foundation to build something just as capable for whatever problem you're solving.

Bonus research preview

We're releasing a temporary research preview alongside Claude Sonnet 4.5, called "Imagine with Claude".

In this experiment, Claude generates software on the fly. No functionality is predetermined; no code is prewritten. What you see is Claude creating in real time, responding and adapting to your requests as you interact.

It's a fun demonstration showing what Claude Sonnet 4.5 can do—a way to see what's possible when you combine a capable model with the right infrastructure.

"Imagine with Claude" is available to Max subscribers for the next five days. We encourage you to try it out on claude.ai/imagine.

Further information

We recommend upgrading to Claude Sonnet 4.5 for all uses. Whether you're using Claude through our apps, our API, or Claude Code, Sonnet 4.5 is a drop-in replacement that provides much improved performance for the same price. Claude Code updates are available to all users. Claude Developer Platform updates, including the Claude Agent SDK, are available to all developers. Code execution and file creation are available on all paid plans in the Claude apps.

For complete technical details and evaluation results, see our system card, model page, and documentation. For more information, explore our

engineering posts and research post on cybersecurity.

Footnotes

1: Customers in the cybersecurity and biological research industries can work with their account teams to join our allowlist in the meantime.

Methodology

- **SWE-bench Verified:** All Claude results were reported using a simple scaffold with two tools—bash and file editing via string replacements. We report 77.2%, which was averaged over 10 trials, no test-time compute, and 200K thinking budget on the full 500-problem SWE-bench Verified dataset.
  - The score reported uses a minor prompt addition: "You should use tools as much as possible, ideally more than 100 times. You should also implement your own tests first before attempting the problem."
  - A 1M context configuration achieves 78.2%, but we report the 200K result as our primary score as the 1M configuration was implicated in our recent inference issues.
  - 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 82.0% for Sonnet 4.5.
- **Terminal-Bench:** All scores reported use the default agent framework (Terminus 2), with XML parser, averaging multiple runs during different days to smooth the eval sensitivity to inference infrastructure.
- **τ2-bench:** Scores were achieved using extended thinking with tool use and a prompt addendum to the Airline and Telecom Agent Policy instructing Claude to better target its known failure modes when using the vanilla prompt. A prompt addendum was also added to the Telecom User prompt to avoid failure modes from the user ending the interaction incorrectly.
- **AIME:** Sonnet 4.5 score reported using sampling at temperature 1.0. The model used 64K reasoning tokens for the Python configuration.
- **OSWorld:** All scores reported use the official OSWorld-Verified framework with 100 max steps, averaged across 4 runs.
- **MMMLU:** All scores reported are the average of 5 runs over 14 non-English languages with extended thinking (up to 128K).
- **Finance Agent:** All scores reported were run and published by Vals AI on their public leaderboard. All Claude model results reported are with extended thinking (up to 64K) and Sonnet 4.5 is reported with interleaved thinking on.

- All OpenAI scores reported from their GPT-5 post, GPT-5 for developers post, GPT-5 system card (SWE-bench Verified reported using n=500), Terminal Bench leaderboard (using Terminus 2), and public Vals AI leaderboard. All Gemini scores reported from their model web page, Terminal Bench leaderboard (using Terminus 1), and public Vals AI leaderboard.

News

Anthropic officially opens Tokyo office, signs Memorandum of Cooperation with the Japan AI Safety Institute

Oct 29, 2025

News

Advancing Claude for Financial Services

Oct 27, 2025

News

Seoul becomes Anthropic's third office in Asia-Pacific as we continue our international growth

Oct 23, 2025

Claude and Slack

Claude in Excel

Max plan

Team plan

Enterprise plan

Download app

Pricing

Log in to Claude

Models

Opus

Sonnet

Haiku

Solutions

AI agents

Code modernization

Coding

Customer support

Education

Financial services

Government

Life sciences

Claude Developer Platform

Overview

Developer docs

Pricing

Amazon Bedrock

Google Cloud’s Vertex AI

Console login

Learn

Courses

Connectors

Customer stories

Engineering at Anthropic

Events

Powered by Claude

Service partners

Startups program

Company

Anthropic

Careers

Economic Futures

Research

News

Responsible Scaling Policy

Security and compliance

Transparency

Help and security

Availability

Status

Support center

Terms and policies

Privacy choices

Privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Usage policy

© 2025 Anthropic PBC