Announcements

Introducing the next generation of Claude

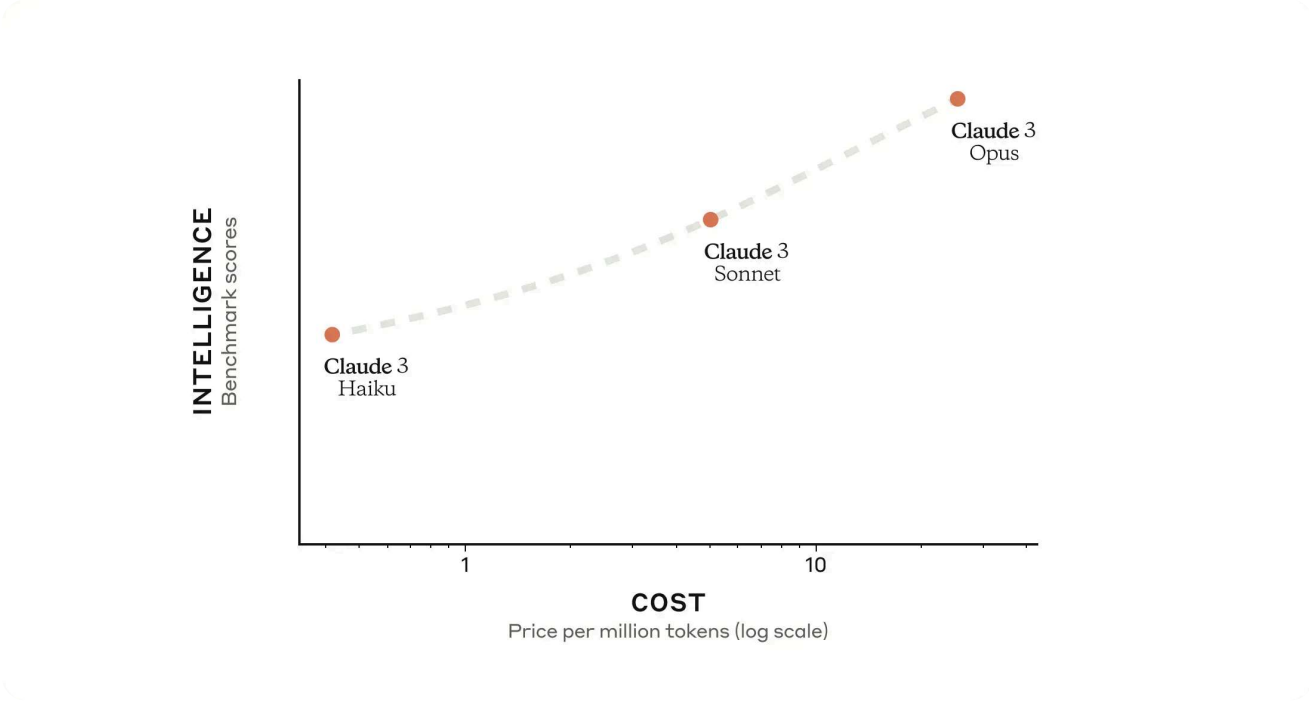
2024年3月4日 • 7 min read

Try Claude 3

Today, we're announcing the Claude 3 model family, which sets new industry benchmarks across a wide range of cognitive tasks. The family includes three state-of-the-art models in ascending order of capability: Claude 3 Haiku, Claude 3 Sonnet, and Claude 3 Opus. Each successive model offers increasingly powerful performance, allowing users to select the optimal balance of intelligence, speed, and cost for their specific application.

Opus and Sonnet are now available to use in claude.ai and the Claude API which is now generally available in 159 countries. Haiku will be available soon.

Claude 3 model family



A new standard for intelligence

Opus, our most intelligent model, outperforms its peers on most of the common evaluation benchmarks for AI systems, including undergraduate level expert knowledge (MMLU), graduate level expert reasoning (GPQA), basic mathematics (GSM8K), and more. It exhibits near-human levels of comprehension and fluency on complex tasks, leading the frontier of general intelligence.

All Claude 3 models show increased capabilities in analysis and forecasting, nuanced content creation, code generation, and conversing in non-English languages like Spanish, Japanese, and French.

Below is a comparison of the Claude 3 models to those of our peers on multiple benchmarks [1] of capability:

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	**86.8%** 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	**50.4%** 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	**84.9%** 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	**83.1** 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	**86.8%** 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	**96.4%** 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	**95.4%** 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Near-instant results

The Claude 3 models can power live customer chats, auto-completions, and data extraction tasks where responses must be immediate and in real-time.

Haiku is the fastest and most cost-effective model on the market for its intelligence category. It can read an information and data dense research paper on arXiv (~10k tokens) with charts and graphs in less than three seconds. Following launch, we expect to improve performance even further.

For the vast majority of workloads, Sonnet is 2x faster than Claude 2 and Claude 2.1 with higher levels of intelligence. It excels at tasks demanding rapid responses, like knowledge retrieval or sales automation. Opus delivers similar speeds to Claude 2 and 2.1, but with much higher levels of intelligence.

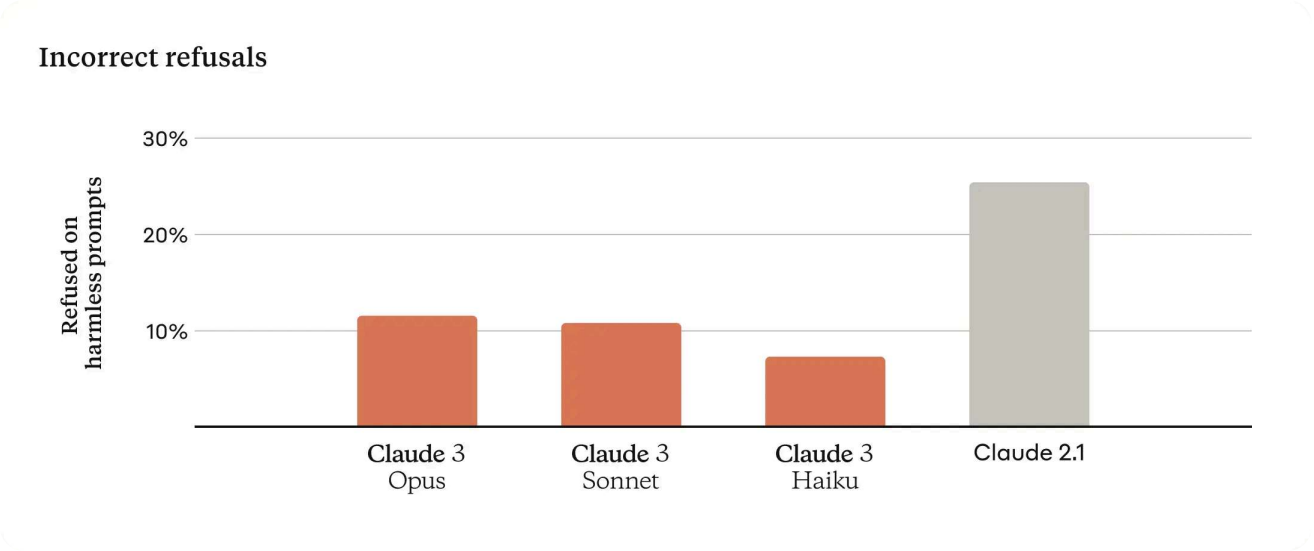
Strong vision capabilities

The Claude 3 models have sophisticated vision capabilities on par with other leading models. They can process a wide range of visual formats, including photos, charts, graphs and technical diagrams. We're particularly excited to provide this new modality to our enterprise customers, some of whom have up to 50% of their knowledge bases encoded in various formats such as PDFs, flowcharts, or presentation slides.

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4V	Gemini 1.0 Ultra	Gemini 1.0 Pro
Math & reasoning *MMMU (val)*	59.4%	53.1%	50.2%	56.8%	59.4%	47.9%
Document visual Q&A *ANLS score, test*	89.3%	89.5%	88.8%	88.4%	90.9%	88.1%
Math *MathVista (testmini)*	50.5% CoT	47.9% CoT	46.4% CoT	49.9%	53.0%	45.2%
Science diagrams *AI2D, test*	88.1%	88.7%	86.7%	78.2%	79.5%	73.9%
Chart Q&A *Relaxed accuracy (test)*	80.8% 0-shot CoT	81.1% 0-shot CoT	81.7% 0-shot CoT	78.5% 4-shot CoT	80.8%	74.1%

Fewer refusals

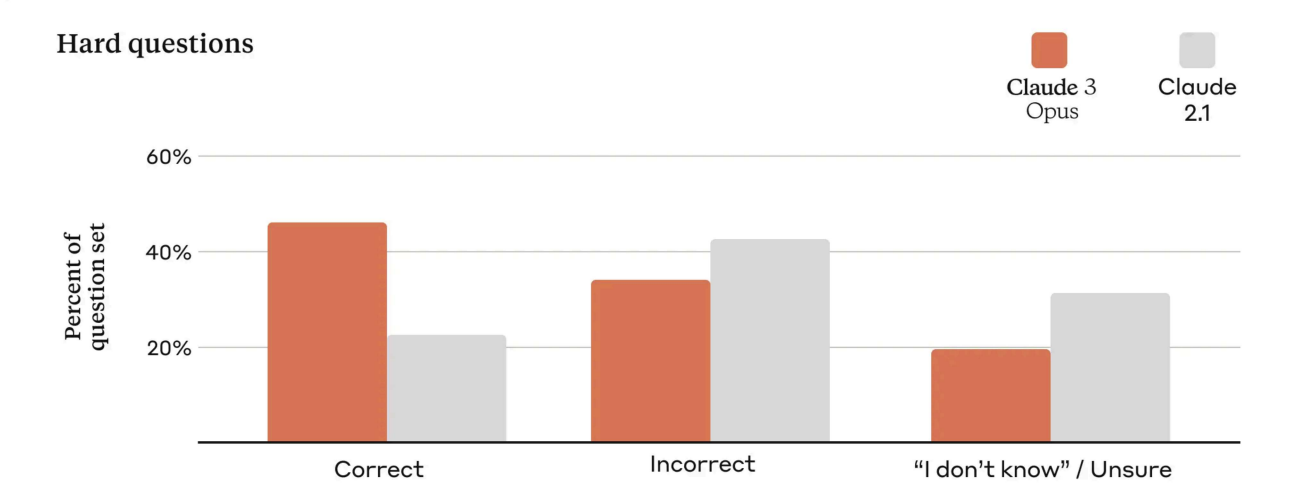
Previous Claude models often made unnecessary refusals that suggested a lack of contextual understanding. We've made meaningful progress in this area: Opus, Sonnet, and Haiku are significantly less likely to refuse to answer prompts that border on the system's guardrails than previous generations of models. As shown below, the Claude 3 models show a more nuanced understanding of requests, recognize real harm, and refuse to answer harmless prompts much less often.



Improved accuracy

Businesses of all sizes rely on our models to serve their customers, making it imperative for our model outputs to maintain high accuracy at scale. To assess this, we use a large set of complex, factual questions that target known weaknesses in current models. We categorize the responses into correct answers, incorrect answers (or hallucinations), and admissions of uncertainty, where the model says it doesn't know the answer instead of providing incorrect information. Compared to Claude 2.1, Opus demonstrates a twofold improvement in accuracy (or correct answers) on these challenging open-ended questions while also exhibiting reduced levels of incorrect answers.

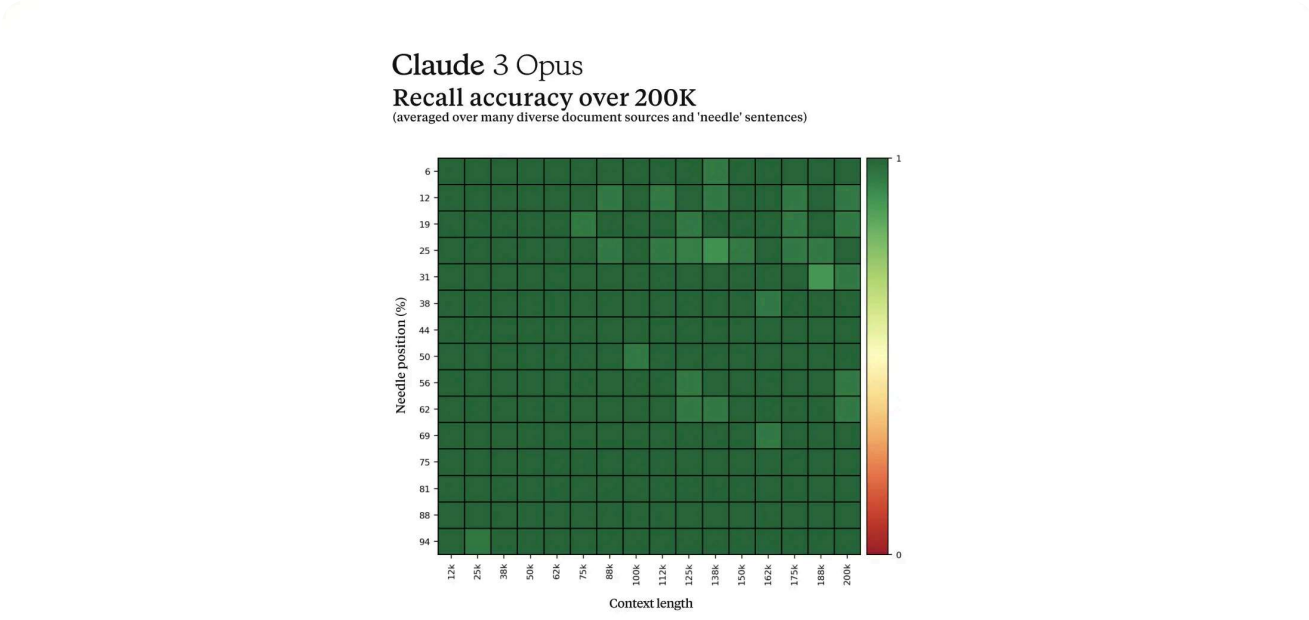
In addition to producing more trustworthy responses, we will soon enable citations in our Claude 3 models so they can point to precise sentences in reference material to verify their answers.



Long context and near-perfect recall

The Claude 3 family of models will initially offer a 200K context window upon launch. However, all three models are capable of accepting inputs exceeding 1 million tokens and we may make this available to select customers who need enhanced processing power.

To process long context prompts effectively, models require robust recall capabilities. The 'Needle In A Haystack' (NIAH) evaluation measures a model's ability to accurately recall information from a vast corpus of data. We enhanced the robustness of this benchmark by using one of 30 random needle/question pairs per prompt and testing on a diverse crowdsourced corpus of documents. Claude 3 Opus not only achieved near-perfect recall, surpassing 99% accuracy, but in some cases, it even identified the limitations of the evaluation itself by recognizing that the "needle" sentence appeared to be artificially inserted into the original text by a human.



Responsible design

We've developed the Claude 3 family of models to be as trustworthy as they are capable. We have several dedicated teams that track and mitigate a broad spectrum of risks, ranging from misinformation and CSAM to biological misuse, election interference, and autonomous replication skills. We continue to develop methods such as Constitutional AI that improve the safety and

transparency of our models, and have tuned our models to mitigate against privacy issues that could be raised by new modalities.

Addressing biases in increasingly sophisticated models is an ongoing effort and we've made strides with this new release. As shown in the model card, Claude 3 shows less biases than our previous models according to the Bias Benchmark for Question Answering (BBQ). We remain committed to advancing techniques that reduce biases and promote greater neutrality in our models, ensuring they are not skewed towards any particular partisan stance.

While the Claude 3 model family has advanced on key measures of biological knowledge, cyber-related knowledge, and autonomy compared to previous models, it remains at AI Safety Level 2 (ASL-2) per our Responsible Scaling Policy. Our red teaming evaluations (performed in line with our White House commitments and the 2023 US Executive Order) have concluded that the models present negligible potential for catastrophic risk at this time. We will continue to carefully monitor future models to assess their proximity to the ASL-3 threshold. Further safety details are available in the Claude 3 model card.

Easier to use

The Claude 3 models are better at following complex, multi-step instructions. They are particularly adept at adhering to brand voice and response guidelines, and developing customer-facing experiences our users can trust. In addition, the Claude 3 models are better at producing popular structured output in formats like JSON—making it simpler to instruct Claude for use cases like natural language classification and sentiment analysis.

Model details

Claude 3 Opus is our most intelligent model, with best-in-market performance on highly complex tasks. It can navigate open-ended prompts and sight-unseen scenarios with remarkable fluency and human-like understanding. Opus shows us the outer limits of what's possible with generative AI.

Cost

[Input $/million tokens | Output $/million tokens]
$15 | $75

Context window
200K*

Potential uses

- Task automation: plan and execute complex actions across APIs and databases, interactive coding
- R&D: research review, brainstorming and hypothesis generation, drug discovery
- Strategy: advanced analysis of charts & graphs, financials and market trends, forecasting

Differentiator
Higher intelligence than any other model available.

data

*1M tokens available for specific use cases, please inquire.

Claude 3 Sonnet strikes the ideal balance between intelligence and speed—particularly for enterprise workloads. It delivers strong performance at a lower cost compared to its peers, and is engineered for high endurance in large-scale AI deployments.

Cost

[Input $/million tokens | Output $/million tokens]
$3 | $15

Context window

200K

Potential uses

- Data processing: RAG or search & retrieval over vast amounts of knowledge
- Sales: product recommendations, forecasting, targeted marketing
- Time-saving tasks: code generation, quality control, parse text from images

Differentiator

More affordable than other models with similar intelligence; better for scale.

data

Claude 3 Haiku is our fastest, most compact model for near-instant responsiveness. It answers simple queries and requests with unmatched speed. Users will be able to build seamless AI experiences that mimic human interactions.

Cost

[Input $/million tokens | Output $/million tokens]
$0.25 | $1.25

Context window

200K

Potential uses

- Customer interactions: quick and accurate support in live interactions, translations
- Content moderation: catch risky behavior or customer requests
- Cost-saving tasks: optimized logistics, inventory management, extract knowledge from unstructured data

Differentiator

Smarter, faster, and more affordable than other models in its intelligence category.

data

Model availability

Opus and Sonnet are available to use today in our API, which is now generally available, enabling developers to sign up and start using these models immediately. Haiku will be available soon. Sonnet is powering the free experience on claude.ai, with Opus available for Claude Pro subscribers.

Sonnet is also available today through Amazon Bedrock and in private preview on Google Cloud's Vertex AI Model Garden—with Opus and Haiku coming soon to both.

Smarter, faster, safer

We do not believe that model intelligence is anywhere near its limits, and we plan to release frequent updates to the Claude 3 model family over the next few months. We're also excited to release a series of features to enhance our models' capabilities, particularly for enterprise use cases and large-scale deployments. These new features will include Tool Use (aka function calling), interactive coding (aka REPL), and more advanced agentic capabilities.

As we push the boundaries of AI capabilities, we're equally committed to ensuring that our safety guardrails keep apace with these leaps in performance. Our hypothesis is that being at the frontier of AI development is the most effective way to steer its trajectory towards positive societal outcomes.

We're excited to see what you create with Claude 3 and hope you will give us feedback to make Claude an even more useful assistant and creative companion. To start building with Claude, visit anthropic.com/claude.

Footnotes

1. This table shows comparisons to models currently available commercially that have released evals. Our model card shows comparisons to models that have been announced but not yet released, such as Gemini 1.5 Pro. In addition, we'd like to note that engineers have worked to optimize prompts and few-shot samples for evaluations and reported higher scores for a newer GPT-4T model. Source.

News

How Anthropic teams use Claude Code

Jul 25, 2025

News

Thoughts on America's AI Action Plan

Jul 24, 2025

News

Anthropic partners with the University of Chicago's Becker Friedman Institute on AI economic research

Jul 23, 2025

Claude.ai login

Research

Research overview

Economic Index

Claude models

Claude Opus 4

Claude Sonnet 4

Claude Haiku 3.5

Commitments

Transparency

Responsible scaling policy

Security and compliance

Solutions

AI agents

Coding

Customer support

Education

Financial services

Learn

Anthropic Academy

Customer stories

Engineering at Anthropic

MCP Integrations

Partner Directory

Explore

About us

Become a partner

Careers

Events

News

Startups program

Help and security

Status

Availability

Support center

Terms and policies

Privacy choices

Privacy policy

Responsible disclosure policy

Terms of service - consumer

Terms of service - commercial

Usage policy