ChatGPT

Claude 4 (Opus 4 & Sonnet 4)

Anthropic’s **Claude 4** is a new generation of AI assistants, released in May 2025. It comprises two models – **Claude Opus 4** and **Claude Sonnet 4** – which set “new standards” for coding, reasoning, and AI agents [1]. Opus 4 is the larger, more powerful model (“the world’s best coding model” [2]), while Sonnet 4 is a highly capable, more efficient model (an upgrade to Sonnet 3.7) designed for widespread use and instruction-following [1] [3]. Both models accept text and image inputs [4] and can operate in two modes: a default quick-response mode and an *extended thinking* mode. In extended thinking mode, Claude generates explicit chain-of-thought reasoning (“internal monologue”) before replying, allowing it to tackle complex, multi-step problems. For example, Claude 4 can iterate through planning and tool use (like web search) many times, enabling it to solve tasks “thousands of times over a period of several hours” [5] [6].

Key capabilities of Claude 4 include: - **Hybrid reasoning:** Near-instant answers by default, with optional *“extended thinking”* for deeper reasoning [6] [7]. In extended mode the model produces step-by-step thought processes before giving a final answer [7].
- **State-of-the-art coding:** Claude Opus 4 leads on coding benchmarks (e.g. 72.5% on SWE-bench) and excels at understanding complex codebases [2]. Opus 4 “delivers sustained performance on long-running tasks…with the ability to work continuously for several hours” [2]. Sonnet 4 also achieves top coding scores (72.7% on SWE-bench) and outperforms Sonnet 3.7 in coding and reasoning [3].
- **Advanced problem-solving:** Both models handle complex reasoning tasks. They can maintain focus through thousands of reasoning steps [2]. Use cases include large-scale research, writing, and scientific discovery (Opus 4 “pushes boundaries” in research and writing, Sonnet 4 brings high-end reasoning to everyday tasks) [8].
- **Large context and memory:** Each Claude 4 model supports an enormous context window (up to **200,000 tokens**) [9], letting them process very long documents or histories. Opus 4 can also create *“memory files”* when given file access, storing notes or facts for later. For instance, while playing the game Pokémon Red, Opus 4 recorded a visual “navigation guide” note and used it to improve its play [10]. This ability to write and recall memory files gives Opus 4 far better long-term coherence and task awareness [10]. *(Figure shows an example memory note.)* - **Multimodal inputs:** Both models take text and images as input [4], enabling tasks like analyzing charts or describing photos.
- **Tool use:** In extended mode Claude 4 can use external tools via a plugin-like API. For example, it can perform web searches, execute code, access a text editor or file system, etc., while reasoning [11] [12]. This effectively turns Claude into an AI agent that can gather data and perform actions during problem solving.
- **Developer integrations:** Anthropic provides **Claude Code** – IDE plugins and SDKs – so Claude 4 can edit code inside VS Code or JetBrains and even run GitHub Actions. For example, new extensions show Claude’s proposed code edits inline in your editor [13].

```
<details>
<summary><strong>Practical use cases</strong></summary>

- **Coding agents:** Companies like Replit report that integrating Claude has
sharply improved their coding tools. Replit noted that Claude 4 Opus is “more
precise” and much better at making changes across multiple files than earlier
```

```
models [14].  GitHub plans to use Sonnet 4 in Copilot for its strong performance
on agentic coding tasks [15].
- **Long-running projects:** Opus 4 excels at tasks requiring many steps.  For
example, Rakuten ran Opus 4 for 7+ hours on an open-source code refactoring and
saw consistent success [16].
- **AI assistants:** Claude 4 can act as a personal assistant. Google Cloud
notes Opus 4 is suitable for “autonomous AI agents” and research, while Sonnet 4
(a “mid-size” model) is ideal for code review, bug fixes, research assistance,
and large content analysis [17] [18].

</details>
```

Design Philosophy

Anthropic designs Claude models around the principles of being **“helpful, honest, and harmless”** [19]. Both Opus 4 and Sonnet 4 were trained with extensive human feedback and advanced alignment methods. In particular, Anthropic uses **“Constitutional AI”** techniques (where a set of principles guides the model’s behavior, e.g. based on the UN Declaration of Human Rights [20]) alongside human review. The models were fine-tuned to exhibit desired traits and to follow rules and ethical guidelines [19]. In practice, this means Claude is encouraged to refuse inappropriate requests, be transparent about reasoning, and avoid manipulations.

- The models were trained on a **proprietary mix of data**: public web text (crawled up to March 2025) plus licensed and user-provided data [21]. Anthropic filtered and deduplicated data carefully. Training emphasizes broad knowledge plus high-quality, helpful output.
- Anthropic actively tests Claude for alignment. A *detailed assessment* in their system card shows **“no significant signs of systematic deception or hidden goals”** in Opus 4 [22]. In other words, Claude 4 showed no evidence of pursuing secret objectives or strategies during evaluation.
- Both models undergo rigorous evaluation against Anthropic’s Usage Policy. They implement guardrails to prevent misuse. For example, Anthropic notes that Claude 4 has **65% fewer instances of taking “shortcuts or loopholes”** on tricky tasks than Sonnet 3.7 did [23], reflecting efforts to reduce gaming or manipulation of tasks.

Safety and Robustness

Claude 4 incorporates multiple safety systems and has been heavily tested to reduce harmful or unintended outputs. Key safety features include:

- **AI Safety Levels:** Anthropic assigns *Opus 4* a high safety standard (AI Safety Level 3) and *Sonnet 4* Level 2 [24]. These levels reflect extensive pre-release testing. Anthropic explicitly states that Claude 4 went through “extensive testing and evaluation to minimize risk and maximize safety,” meeting stricter safety criteria than earlier models [25] [24].
- **Prompt-injection defenses:** The models incorporate specialized defenses against prompt-injection attacks. Anthropic reports using targeted reinforcement learning to recognize and avoid malicious prompts, plus detectors that can halt execution if an attack is detected [26]. In tests, these defenses improved Claude 4’s ability to resist code or instruction injection.

- **Alignment testing:** The system card documents many adversarial tests (e.g. malicious computer use, jailbreaking, bias) and reports that Claude 4 passes them at high rates. For example, aggregated malicious-use scores for Claude 4 models reach the mid-80% range (higher is safer), which is better than earlier versions. (Anthropic also continuously monitors for new issues in deployment.)
- **Reduced misbehavior:** In practice, Claude 4 rarely defies instructions or provides unsafe content. The extended thinking mode and training regimen encourage careful, rule-following behavior. Notably, Anthropic found *Opus 4* showed no strong sign of pursuing its own agenda [22], and when pushed to extreme scenarios it usually explained its actions rather than hide them.

Taken together, these safety measures mean Claude 4 is built to be a responsible assistant. The company highlights that both models now meet higher internal standards for safe deployment than earlier Claude versions [25] [24].

Summary of Claude 4's Innovations

- **Coding prowess:** Opus 4 is the most advanced model Anthropic has built for programming. It **outperforms all previous Claude models on coding benchmarks** (SWE-bench, Terminal-bench) [2]. Sonnet 4 also sets new records for its class [3].
- **Long-context workflows:** With a 200K token context window [9] and new memory features [10], Claude 4 can handle very large documents or multi-file codebases in one conversation.
- **Agentic reasoning:** The hybrid mode and tool use let Claude 4 act as an AI agent. It can iteratively plan, use tools (search, code execution, file access), and revise its plan. This lifts it far beyond a simple chat assistant toward an autonomous collaborator [5] [27].
- **Better alignment:** Claude 4 inherits Anthropic's latest safety work. Compared to Claude 3.7, it makes fewer "loophole" mistakes [23] and passed a battery of alignment evaluations. Opus 4 was released under even stricter safety criteria (ASL-3) than Sonnet 4 [24].
- **Efficiency vs. Power:** Sonnet 4 offers much of the power of Opus at lower cost. Anthropic describes Sonnet 4 as an *"optimal mix of capability and practicality"* [28]. It is even made available on the free tier of the Claude service. Opus 4 is reserved for paid plans.

In short, Claude 4 represents a major step up for Anthropic's AI, especially in coding and agentic applications, while continuing the company's emphasis on reliability and safety. It is designed to serve both as a powerful co-pilot for developers and as a versatile reasoning assistant for a wide range of tasks.

Technical Appendix

Model Architecture and Size

Anthropic has not publicly released the exact architecture or parameter count of Claude Opus 4 or Sonnet 4. They are presumed to be very large transformer-based models, following the lineage of Claude 3.x (which were decoder-only transformers). Industry observers estimate they likely contain on the order of hundreds of billions of parameters, but no official figures are available. (By comparison, OpenAI's GPT-4 is speculated to be many hundreds of billions or even over a trillion parameters.) The models support both text and image inputs [4], indicating a multi-modal training.

Training Data and Pretraining

Claude 4 was pretrained on a vast dataset combining public and private data. According to Anthropic, training data included:

- **Public Internet data (Web crawls)**: A general-purpose web crawler collected public web pages up to *March 2025* [21]. Anthropic respects robots.txt and avoids private or gated content, then applies deduplication and filtering.
- **Licensed and private datasets**: Proprietary third-party data (e.g. books, specialized corpora) and contributions from paid annotators or partners were included [21].
- **User data (opted-in)**: Data from Claude users who consented to contribute their chat transcripts to training.
- **Synthetic and augmented data**: Anthropic also generated or expanded data using its own models during training.

Overall, the data is "large" and "diverse" to build general language understanding [19]. The training cutoff is early 2025 (March), so the models have current knowledge up to that date [29].

Tokenization and Context Window

Claude 4 uses a token-based approach. Both Opus 4 and Sonnet 4 have an extremely large context window of **200,000 tokens** [9], far exceeding previous models. This allows feeding documents of roughly 100,000+ words. Output length is also flexible: Opus 4 can generate up to ~32,000 tokens (about 20,000 words) in one response, while Sonnet 4 supports up to ~64,000 output tokens by default (according to the API docs) [9]. (API options allow extending these limits further if needed.)

Fine-Tuning and Alignment

After pretraining, Anthropic applied a sophisticated fine-tuning and alignment pipeline. As in prior Claude versions, they used a combination of techniques:

- **Human Feedback / Reinforcement Learning**: Human reviewers rated outputs and instructed the model, likely using techniques similar to RLHF (Reinforcement Learning from Human Feedback), to encourage helpful, safe answers. The system card refers to "human feedback" explicitly [20].
- **Constitutional AI**: They applied a "constitutional" fine-tuning phase, where Claude is guided by a written constitution of rules (e.g. "Be helpful" / "Avoid harassment") rather than direct human labels for every example [20]. Anthropic cites research on Constitutional AI (Bai et al. 2022) in the system card [20].
- **Trait training**: The card mentions training "selected character traits" to shape model behavior (e.g. being more cautious or cooperative) [20].
- **Adversarial and safety training**: They exposed Claude 4 to adversarial scenarios (e.g. malicious prompts) during training to help it learn to refuse harmful requests. The system card notes adding "specialized reinforcement learning" to improve robustness against attacks [26].

Throughout fine-tuning, Anthropic paid special attention to the model's **alignment**. The system card reports extensive testing for misalignment risks (hidden goals, sabotage, bias, jailbreaking, etc.) and updates the model to mitigate discovered issues. For example, no strong reward-hacking or deceptive behaviors were found in final training snapshots [22].

Safety Evaluations

Anthropic's **System Card** (123 pages) documents many internal safety evaluations. Key points include:

- **AI Safety Levels (ASL):** Based on these tests, Claude Opus 4 met *ASL-3* criteria and Sonnet 4 met *ASL-2*, where higher ASL requires stricter controls [24] .
- **Misbehavior reduction:** Compared to Claude 3.7, the new models are **65% less likely** to exploit "shortcuts or loopholes" on difficult tasks [23] .
- **Refusal rates:** Both models show higher refusal rates on disallowed queries (e.g. violent or sexual) than older models. (Exact numbers are in the card tables.)
- **Bias and fairness:** The card includes bias evaluations and reports no severe issues. Specialized classifiers and training were used to reduce biased associations.
- **Malicious use:** Claude 4 was tested on "agentic" use cases like malware coding. The system card's Table 3.3A shows Opus 4 and Sonnet 4 achieving very high safety scores (~88–90%) on malicious coding tests, outperforming Sonnet 3.7 [30] .

Context and Memory Handling

Claude 4's large context lets it refer back to lengthy inputs. Additionally, Opus 4's **memory files** feature extends context across interactions: when an application grants file access, Opus 4 can write and read files to store facts. Anthropic demonstrated Opus 4 creating a "Navigation Guide" note while playing Pokémon, showing it can recall details (e.g. town locations) days or weeks later [10] . The models also summarize long reasoning chains: if an internal chain-of-thought becomes too long, a smaller model condenses it into a summary [31] . This "thinking summary" occurs in roughly 5% of cases, so most chains are shown in full for transparency [31] .

Tools and API Features

Claude 4 supports a variety of tools via Anthropic's **Tools API** (beta). Notable tools include: - **Web search:** Claude can issue search queries and read results during extended reasoning. - **Code execution:** A sandboxed Python executor lets Claude run code, analyze data, and produce charts [12] . - **Text editor / file tools:** Claude can open and edit files (e.g. code files) via an API. - **Shell/Computer use:** Command-line and other simulated computer tools let Claude interact with virtual environments. - **Files API:** Developers can upload files for Claude to use or store as memories.

When these tools are enabled, Claude alternates between thinking and using tools to gather information or run tasks. (The system card reported improved performance in "agentic" benchmarks when tool use was allowed.) Using these tools effectively makes Claude 4 a powerful agent: for instance, it could research on the web, load and modify a code repository, and iterate on solutions without human intervention.

Model Usage and Pricing

Claude Opus 4 is available on paid plans; Sonnet 4 is also offered free to all users [6] . The Anthropic API pricing is roughly \$15 per million tokens (input+output) for Opus 4 and \$3 per million for Sonnet 4 (base rate) [32] . (Additional reduced rates apply for cached prompts and background tasks as documented.)

Benchmarks

Anthropic reports that Claude 4 achieves state-of-art results on benchmarks such as **SWE-bench Verified** (software engineering tasks). For example, Opus 4 scored 72.5% on SWE-bench and 43.2% on Terminal-bench [2], both improvements over prior models. Sonnet 4 scored 72.7% on SWE-bench [3]. These are among the highest published scores on coding benchmarks as of mid-2025. The models also excel at reasoning benchmarks, but those numbers are not publicly detailed.

References: Anthropic's official announcement and documentation [1] [2] [23] [19], system card and API docs [9] [12] [26] [22], and reporting from trusted sources [7] [33] were used to compile this report.

---

[1] [2] [3] [6] [8] [10] [11] [13] [15] [16] [23] [25] [27] [28] [31] [32] Introducing Claude 4 \ Anthropic
https://www.anthropic.com/news/claude-4

[4] [9] [29] Models overview - Anthropic
https://docs.anthropic.com/en/docs/about-claude/models/overview

[5] [7] [14] Anthropic Releases Claude 4, 'the World's Best Coding Model'
https://www.inc.com/ben-sherry/anthropic-releases-claude-4-the-worlds-best-coding-model/91192856

[12] Code execution tool - Anthropic
https://docs.anthropic.com/en/docs/agents-and-tools/tool-use/code-execution-tool

[17] [18] [33] Anthropic's Claude Opus 4 and Claude Sonnet 4 on Vertex AI | Google Cloud Blog
https://cloud.google.com/blog/products/ai-machine-learning/anthropics-claude-opus-4-and-claude-sonnet-4-on-vertex-ai

[19] [20] [21] [22] [24] [26] [30] Claude 4 System Card
https://anthropic.com/model-card