ChatGPT

Anthropic's Claude 3 Sonnet and Opus: Next-Generation AI Models

Anthropic's **Claude 3** family (introduced March 2024) includes three large multimodal AI models – Haiku (smallest), **Sonnet** (mid-sized), and **Opus** (largest) [1]. Sonnet and Opus process both text and images, and excel at writing, reasoning, coding and analysis tasks. Opus is Anthropic's **most capable** model, achieving "near-human" comprehension and leading the frontier of general intelligence [2]. Sonnet is designed as a *balanced* model: roughly twice as fast as older Claude models with similarly strong intelligence [3], at a fraction of the cost [4]. In practice, Sonnet powers Claude's free tier, while Opus (the pro model) is used for the most demanding tasks [5] [6].

Both Sonnet and Opus build on the same underlying design principles. They are large **transformer-based** neural networks trained to predict the next word in text, and fine-tuned to follow instructions. They incorporate safety by design: trained with Anthropic's *Constitutional AI* approach (rules-based alignment) and reinforcement learning from human feedback to be helpful, honest, and harmless [7] [8]. Anthropic has also tuned these models to reduce unsafe outputs and bias. The Claude 3 models in general show **fewer inappropriate refusals** and give more nuanced answers than previous versions [9]. They also perform better on factual accuracy: Opus delivers about twice as many correct answers on hard open-ended questions as Claude 2.1, with far fewer mistakes [10].

Claude 3 Sonnet and Opus share many new capabilities. Both can process very **long contexts** – up to 200,000 tokens in a single prompt (nearly 150,000 words) [3] [11]. (Specialized customers can push this even further; Anthropic notes the models can accept over 1 million tokens if needed [12].) For example, Sonnet can read a 10,000-token research paper and charts in a few seconds [13]. They also "remember" details extremely well: in recall tests, Claude 3 Opus exceeded 99% accuracy on finding facts in long documents [14]. Both models support multilingual output (Spanish, Japanese, French, etc.) [15] and, importantly, **vision**. They can analyze images – photos, charts, graphs or diagrams – and answer questions about them [16] [17]. (In fact, a 2024 update *Sonnet 3.5* became the strongest vision model in the family, surpassing Opus on image benchmarks [18].)

Core Functionality. In everyday use, Claude 3 Sonnet and Opus act like advanced chat assistants. They excel at drafting and editing text (articles, emails, essays), generating creative content, summarizing or extracting information, and answering questions with reasoning. They are also strong at code: in tests Sonnet 3.5 solved ~64% of coding problems (versus 38% for Opus) [19], and Sonnet can write, edit, and execute code when given appropriate tools [19]. Both models support "tool use" (a form of function calling) so they can interact with external APIs or plugins for tasks like database queries or live data lookup [20]. In short, Sonnet and Opus can help automate complex workflows (e.g. data analysis, report generation, planning tasks) and serve as creative assistants or research collaborators.

Design Philosophy and Safety. Anthropic emphasizes **responsible AI** in Claude's design. The models are trained to follow a "constitution" of ethical guidelines rather than relying solely on human labels [8] [7]. For instance, the Claude 3 constitution (informed by human rights principles) explicitly teaches respect for

sensitive issues (e.g. disability rights) [21]. As a result, Claude 3 Sonnet and Opus are less likely to produce harmful or biased content, and in tests show reduced bias compared to prior models [22]. Dedicated safety teams continuously "red-teams" the models against misuse (misinformation, privacy breaches, dangerous instructions, etc.) [23]. Anthropic has publicly committed to U.S. government guidelines on AI safety: at release they classified Claude 3 at "Safety Level 2" (negligible catastrophic risk) [24]. In practical terms, Claude 3 will still refuse or safely manage disallowed queries, but it *knows the difference* between real harm and harmless topics, so it refuses far less often than older Claude models [25].

Differences Between Sonnet and Opus. The main trade-off is **capability vs speed/cost**. Opus is bigger and more capable: it "sets a new standard" on reasoning, advanced math, and coding [2]. It can handle open-ended creative and analytical tasks with very high accuracy (e.g. strategic analysis of charts or scientific reasoning). However, Opus is roughly the same speed as Claude 2 and 2.1. It is also expensive: roughly \$15 per 1,000 input tokens and \$75 per 1,000 output tokens [6]. Sonnet, by contrast, is engineered for **scale**. It is about 2× faster than Claude 2/2.1 and nearly as intelligent, making it ideal for high-volume or real-time tasks [3]. Sonnet costs only \$3 per 1,000 input and \$15 per 1,000 output tokens [26] – one-fifth the input price of Opus – and is pitched for enterprise workloads where cost and speed matter [4]. Accordingly, Sonnet is used in high-throughput settings (like search & retrieval over knowledge bases, customer support automation, large-scale data parsing) [27], while Opus is reserved for specialized tasks needing maximal reasoning ability. In short, Sonnet "strikes the ideal balance between intelligence and speed," whereas Opus offers the very highest raw capability [4] [6].

Both models share a 200,000-token context window [28] [11], so neither loses context in typical use. (Opus can be configured to handle up to 1 million tokens for special customers [29].) They also share the same training data cut-off (mid-2023) and foundational architecture, so their knowledge base and underlying "world model" are comparable. In practice, non-specialist users will notice Opus giving slightly more insightful or creative answers on tough problems, at the cost of slower responses and higher usage fees. Sonnet, on the other hand, is nimbler and cheaper, often "good enough" or even better (as in vision tasks [18]) for most purposes.

Technical Appendix

Model Architecture

Claude 3 Sonnet and Opus are **autoregressive transformer** models, meaning they generate text one token at a time by attending to all prior tokens. Anthropic has not disclosed exact architecture details (layer count, embedding size, etc.), but they use state-of-the-art transformer designs. Both models were pre-trained using standard next-word (token) prediction on massive text corpora [30]. They likely have many billions of parameters (undisclosed), similar to or exceeding other top models. Some third-party estimates suggest Claude 3 Opus could be on the order of *10^12* parameters, while Sonnet might be smaller; however, Anthropic has not confirmed any parameter counts.

There is no public mention of radical architectural innovations (like mixture-of-experts, new activation functions, or sparse attention) in Claude 3's release. Available information suggests Claude 3 follows the familiar multi-layer self-attention + feedforward (MLP) block pattern. Internally, the network uses softmax attention and MLPs ("word prediction" training) along with positional encodings. It supports multilingual text and also takes **image inputs**: likely via an image encoder (similar to CLIP or other vision transformers) that maps images into tokens for the main transformer. In early experiments, Claude 3.7 Sonnet introduced

an *“extended thinking”* mode, allowing it to internally generate and display a chain-of-thought reasoning trace before answering [31]. Technically this suggests the model can pause and produce a reasoning sequence on demand, but in the base Sonnet and Opus this is implemented simply as additional decoding steps (a “hybrid reasoning model”) rather than a separate network.

Both models have very large context windows (200K tokens standard). This likely uses efficient transformer techniques (e.g. sliding-window attention or optimized hardware) to handle such long inputs. Anthropic’s documentation emphasizes tools like function-calling (“Tool Use”) and interactive coding (REPL) as features, implying the models can generate special JSON-formatted outputs or code that triggers external actions [20]. These integrations suggest the model tokenizes and produces special tokens for tool calls, though the exact mechanism is proprietary.

Training Data and Scale

Anthropic has revealed that Claude 3 was trained on a **mixture of data sources**. Specifically, the model card states: *“Claude 3 models are trained on a proprietary mix of publicly available information on the Internet as of August 2023, as well as non-public data from third parties, data from labeling services and paid contractors, and data we generate internally.”* [32]. In other words, training data includes a huge web crawl (text and code from websites), plus licensed or purchased datasets (e.g. proprietary books, academic content, news archives) and specialized data (like high-quality instruction pairs made by contractors). The data was carefully cleaned (filters, deduplication) [33]. Importantly, no user-submitted prompts or outputs were used in pre-training [34], only publicly-available and proprietary corpora.

The training set was enormous. Anthropic’s model card implies *“trillions”* of tokens. Industry analyses estimate that models like Claude 3 used on the order of 10^12–10^13 tokens of text [35]. For example, one review notes “state-of-the-art models (GPT-4, Claude 3, etc.) each exceed 2 T tokens after aggressive filtering” [35]. Assuming Claude 3’s training pipeline was similar, it likely saw on the order of several trillion words (with more weight given to higher-quality content). Roughly half the data would be web text, along with code (e.g. GitHub), books, articles, and some mixed-language text. Images were included via paired text-image datasets (so the model learns to caption or analyze images). In short, Sonnet and Opus were trained on a vast, diverse corpus spanning web text, code, books, and image-text pairs up to mid-2023 [32] [35].

Model Size and Parameters

Anthropic has **not published** the exact number of parameters in Sonnet or Opus. However, the huge context window and high performance imply they are very large. By analogy, GPT-4 (a competitor) is rumored to be in the hundreds of billions of parameters, and some commentators have speculated that Claude 3 Opus may reach into the *trillion-parameter* range. For context, a separate estimate from Alan Cheung (the “Life Architect” newsletter) speculated Opus might be ~2 trillion parameters with Sonnet ~70 billion, but these are unconfirmed rumors. In any case, both models likely have multiple Transformer layers and large embedding sizes typical of top-tier LLMs. The models use subword tokenization and an internal vocabulary similar in size to other general LMs (tens of thousands of tokens).

Since Sonnet and Opus belong to the same family, Sonnet is probably smaller (fewer layers or narrower) and Opus larger. Some clues: Sonnet 3.5 is described as a “mid-tier” model [36], whereas Opus is Anthropic’s flagship. Regardless of size, both models fit on modern TPU/GPU clusters (they were trained on AWS/GCP

supercomputers [37]) and can run inference for users. Performance tuning (like faster attention or layer optimizations) was likely applied to make Sonnet faster for enterprise use, but specifics are not disclosed.

Alignment and Safety Techniques

Anthropic has invested heavily in aligning Claude 3. Both Sonnet and Opus were fine-tuned after pre-training using **human feedback and Constitutional AI** [38]. The process was: (1) supervised fine-tuning on expert-written examples, (2) applying a "constitution" of ethical rules (drawn from human rights declarations) to self-critique outputs, and (3) reinforcement learning from human feedback (RLHF). In practice, this means the models learned to refuse or safely answer harmful queries and to be helpful on allowed tasks [38] [23]. Claude's constitution now includes principles for respecting disability rights, among others [21].

After training, extensive **evaluation and red-teaming** was performed. Anthropic ran benchmarks and adversarial tests (including in line with U.S. government AI safety commitments) to check for issues like bias, privacy leakage, misinformation, or dangerous capabilities [23] [24]. The Claude 3 model card reports significant reductions in bias (e.g. on the BBQ bias benchmark) compared to prior Claude versions [22]. The team also continues to monitor outputs during deployment, using automated classifiers and human oversight to catch problems in real time.

In summary, alignment for Claude 3 involved a combination of **training-time safeguards** (Constitutional AI rules in RLHF) and **post-training scrutiny** (red teaming and continuous monitoring) [8] [38]. The result is that Sonnet and Opus aim to be helpful and grounded, while avoiding disallowed content. When they do refuse a request, it is usually for a legitimate reason (e.g. hateful or illegal content) rather than "hallucinating" ignorance.

Other Internal Mechanisms

- **Chain-of-thought reasoning:** While Claude 3 itself does not always show intermediate reasoning by default, Anthropic's 3.7 Sonnet introduced an *"extended thinking"* mode where the model internally generates a step-by-step reasoning chain before answering [31]. This suggests the model has an inherent ability to produce and use multi-step reasoning, although in Sonnet/Opus this was controlled via modes in the interface. In general, the model's outputs often reflect hidden "chains of thought" in its activations (see recent research on Claude's neuron interpretability).

- **Tool/Function Calling:** As noted above, Claude 3 supports function calling ("Tool Use"), meaning it can output structured data that triggers external actions [20]. Under the hood, the model generates JSON or code-like tokens which are parsed by the Claude API. For example, the model can output a JSON payload to query a database or call a coding REPL. This enables advanced applications like Claude Code (a coding assistant CLI) [39].

- **Memory and Context:** Claude 3 does not have a persistent "memory store" apart from the current context window. It treats all conversation history and documents in the prompt as context for next-token prediction [40]. Therefore, any long-term memory must be managed externally (e.g. by feeding relevant notes into the prompt). The very large context size (200K) does serve as a short-term "working memory," allowing the model to reference much more information at once.

- **Performance Optimization:** Sonnet is specifically optimized for speed. According to Anthropic, Sonnet's inference latency is roughly half that of Opus (2× faster than Claude 2) for comparable prompts [3]. They likely achieved this through architecture tweaks and engineering (e.g. pipeline parallelism, optimized kernels). Opus is slower per token but focuses on maximum accuracy. Both models can stream output to the user, and Claude's API recommends streaming messages for very long outputs to avoid timeouts [41].

- **Vision Processing:** For images, Claude 3 Sonnet/Opus use a vision component. While details are not public, the model card implies image input is processed similarly to other multimodal models: the image is converted into tokens (via convolutional or patch embeddings) and fed into the transformer. The model has been tested on charts, graphs, photos, and figures [16]. At least from user-facing descriptions, the vision capability appears native (no external vision API called).

In conclusion, Claude 3 Sonnet and Opus are highly advanced AI assistants built on state-of-the-art LLM technology. They differ mainly in scale and performance (Opus being the more powerful, Sonnet being more efficient), but share the same careful safety-focused design. Together, they push the frontier of what AI can do in understanding and generating language and even visual content, while aiming to remain aligned with human intentions [2] [23].

Sources: Authoritative details about Claude 3 Sonnet and Opus come from Anthropic's official announcements and model cards [1] [3] [14] [6] [39] [18] [32] [7], as well as technology analyses and reporting [35] [20]. Citations are included above for verification.

---

[1] [2] [3] [4] [5] [6] [9] [10] [12] [13] [14] [15] [16] [20] [22] [23] [24] [25] [26] [27] [28] [29] Introducing the next generation of Claude \ Anthropic
https://www.anthropic.com/news/claude-3-family

[7] [17] [21] [30] [32] [33] [34] [37] [38] [40] 〖英文原版〗Claude3技术报告-Claude3模型系列作品、十四行诗、俳句,行业研报,研究报告,行业报告,行业分析,行业研究报告
https://www.qingbaoyuan.vip/doc/42360/

[8] Constitutional AI: Harmlessness from AI Feedback \ Anthropic
https://www.anthropic.com/news/constitutional-ai-harmlessness-from-ai-feedback

[11] [41] Models overview - Anthropic
https://docs.anthropic.com/en/docs/about-claude/models/overview

[18] [19] [36] Introducing Claude 3.5 Sonnet \ Anthropic
https://www.anthropic.com/news/claude-3-5-sonnet

[31] [39] Claude 3.7 Sonnet and Claude Code \ Anthropic
https://www.anthropic.com/news/claude-3-7-sonnet

[35] Inside the Great AI Data Grab — Comprehensive Analysis of Public and Proprietary Corpora Utilizsed in Training Contemporary Foundation Language Models | by Adnan Masood, PhD. | May, 2025 | Medium
https://medium.com/@adnanmasood/inside-the-great-ai-data-grab-comprehensive-analysis-of-public-and-proprietary-corpora-utilised-49b4770abc47