Claude Sonnet 3.7 vs Claude Sonnet 4

Compare Anthropic's Claude Sonnet 3.7 and Claude Sonnet 4, explore key upgrades in reasoning, contextual memory, and user experience to find the best AI model for your needs.

Anthropic's Claude Sonnet 3.7 and the newer Claude Sonnet 4 both aim to deliver strong performance across reasoning, creativity, and task completion, but with key upgrades in the latest release.

Claude Sonnet 3.7 set a solid foundation with reliable language understanding and fast responses. Claude Sonnet 4 builds on that with improved contextual memory, deeper reasoning, and more natural conversation flow, making it better suited for complex tasks and longer interactions.

In this article, we compare Claude Sonnet 3.7 and Sonnet 4 across several dimensions, speed, accuracy, coding ability, and overall user experience, to help you decide which model best fits your needs.

Specifications and Technical Details

Model	**Claude Sonnet 3.7**	**Claude Sonnet 4**
Alias	claude-3-7-sonnet-20250219	claude-sonnet-4-20250514
Description (provider)	Our most intelligent model to date and the first hybrid reasoning model on the market.	Our high-performance model with exceptional reasoning and efficiency
Release date	February 2025	22 May 2025
Developer	Anthropic	Anthropic
Primary use cases	RAG, search & retrieval, code generation, content creation	code generation, advanced AI chatbots, knowledge and Q&A
Context window	200k tokens	200k tokens

Model	Claude Sonnet 3.7	Claude Sonnet 4
Max output tokens	8192 tokens	64k
Knowledge cutoff	November 2024	March 2025
Multimodal	Accepted input: text, image	Accepted input: text, image
Fine tuning	No	No

Sources:

- Anthropic news release: Claude Sonnet 4 \ Anthropic
- Anthropic Documentation: Intro to Claude - Anthropic

Performance Benchmarks

Claude Sonnet 4 significantly outperforms Sonnet 3.7 in software engineering, scoring **72.7%** (or **80.2%** with parallel compute) on SWE-bench versus **62.3%** (70.3% with compute) for 3.7. It also dominates in high school-level math (AIME), achieving **70.5% / 85.0%** compared to Sonnet 3.7's **54.8%**.

In contrast, Sonnet 3.7 slightly edges out Sonnet 4 in graduate-level reasoning (**78.2%** vs **75.4%**) and visual reasoning (**75.0%** vs **74.4%**). Both perform similarly in multilingual Q&A (**~86%**) and agentic tool use (Retail: **~81%**, Airline: **~59–60%**).

	Claude Opus 4	Claude Sonnet 4	Claude Sonnet 3.7	OpenAI o3	OpenAI GPT-4.1	Gemini 2.5 Pro *Preview (05-06)*
Agentic coding *SWE-bench Verified*[1,5]	72.5% / 79.4%	72.7% / 80.2%	62.3% / 70.3%	69.1%	54.6%	63.2%
Agentic terminal coding *Terminal-bench*[2,5]	43.2% / 50.0%	35.5% / 41.3%	35.2%	30.2%	30.3%	25.3%
Graduate-level reasoning *GPQA Diamond*[5]	79.6% / 83.3%	75.4% / 83.8%	78.2%	83.3%	66.3%	83.0%
Agentic tool use *TAU-bench*	Retail 81.4% Airline 59.6%	Retail 80.5% Airline 60.0%	Retail 81.2% Airline 58.4%	Retail 70.4% Airline 52.0%	Retail 68.0% Airline 49.4%	— —
Multilingual Q&A *MMMLU*[3]	88.8%	86.5%	85.9%	88.8%	83.7%	—
Visual reasoning *MMMU (validation)*	76.5%	74.4%	75.0%	82.9%	74.8%	79.6%
High school math competition *AIME 2025*[4,5]	75.5% / 90.0%	70.5% / 85.0%	54.8%	88.9%	—	83.0%

Methodology

1. Opus 4 and Sonnet 4 achieve 72.5% and 72.7% pass@1 with bash/editor tools (averaged over 10 trials, single-attempt patches, no test-time compute, using nucleus sampling with a top_p of 0.95).
2. Opus 4 and Sonnet 4 score 39.2% and 33.5% pass@1 with the same agent as non-Claude models, the above reported 43.2% and 35.5% with Claude Code as agent framework.
3. Claude scores on MMMLU are the average over 14 non-English languages.
4. Opus 4 and Sonnet 4 were run on AIME using nucleus sampling with a top_p of 0.95.
5. On SWE-Bench, Terminal-Bench, GPQA and AIME, we additionally report results that benefit from parallel test-time compute by sampling multiple sequences and selecting the single best via an internal scoring model.

Overall, Sonnet 4 is clearly stronger for technical and coding-intensive tasks, while Sonnet 3.7 remains competitive in reasoning and general-purpose use.

Sources:

- Anthropic news release: https://www.anthropic.com/claude/sonnet

Practical Applications and Use Cases

Claude 3.7 Sonnet ：

- **Developer Support:** Enhances engineering workflows with contextual code generation, smart debugging assistance, and natural language explanations of complex codebases.

- **Business Operations:** Streamlines data handling by summarizing documents, extracting key insights from emails, and efficiently organizing feedback or survey results.
- **Customer Engagement:** Serves as a responsive virtual assistant, resolving service queries promptly and professionally, contributing to a smoother and more satisfying user experien

Claude Sonnet 4

- **Code Generation:** Powers end-to-end software development with high performance in planning, writing, debugging, and refactoring code—ideal for complex, agentic programming workflows.
- **Customer-Facing AI:** Delivers advanced reasoning, precise tool use, and strong instruction following—perfect for intelligent, reliable customer service agents and AI-driven support systems.
- **Knowledge Q&A:** Handles large documents and codebases with ease, offering accurate, low-hallucination responses thanks to its large context window and refined comprehension.

Using the Models with APIs

For developers interested in building custom AI solutions with Claude Sonnet 3.7 and 4, they are available on the Anthropic API, Amazon Bedrock, and Google Cloud's Vertex AI.

Accessing APIs Directly

Claude 3.7 request example

Python request example for chat with Anthropic API:

```
import anthropic


client = anthropic.Anthropic(
    # defaults to os.environ.get("ANTHROPIC_API_KEY")
    api_key="my_api_key",
)
message = client.messages.create(
    model="claude-3-7-sonnet-20250219",
```

```
    max_tokens=1024,
    messages=[
        {"role": "user", "content": "Hello, Claude"}
    ]
)
print(message.content)
```

Claude 4 request example

Python request example for chat with Anthropic API:

```
import anthropic

client = anthropic.Anthropic(
    # defaults to os.environ.get("ANTHROPIC_API_KEY")
    api_key="my_api_key",
)
message = client.messages.create(
    model="claude-sonnet-4-20250514",
    max_tokens=1024,
    messages=[
        {"role": "user", "content": "Hello, Claude"}
    ]
)
print(message.content)
```

Streamlined Access to Claude Models with Eden AI

Eden AI offers a unified platform to easily access Claude Sonnet 3.7 and Claude Sonnet 4 through a single, simplified API, eliminating the need to manage multiple keys or complex integrations. This makes it easier than ever to deploy powerful Claude models in your applications, with built-in support for custom data sources via an intuitive UI and Python SDK.

Designed with developers in mind, Eden AI provides transparent, usage-based pricing with no hidden fees or subscription costs. You pay only for what you use, with unlimited API call volume and clear supplier-side margins.

Whether you're building with Claude Sonnet 3.7's balanced reasoning or Claude Sonnet 4's high-performance capabilities, Eden AI helps teams scale AI-powered solutions efficiently with robust monitoring tools to track performance and ensure reliability at every step.

Eden AI Example Workflow

Python request (multimodal chat) example for chat with Eden AI API:

```
import requests


url = "https://api.edenai.run/v2/multimodal/chat"


payload = {
    "fallback_providers": ["DeepSeek-R1"],
    "response_as_dict": True,
    "attributes_as_list": False,
    "show_base_64": True,
    "show_original_response": False,
    "temperature": 0,
    "max_tokens": 16384,
    "providers": ["claude-3-7-sonnet-20250219"]
}
```

```
headers = {
   "accept": "application/json",
   "content-type": "application/json"
}

response = requests.post(url, json=payload, headers=headers)

print(response.text)
```

Cost Analysis

Cost (per 1M tokens)	**Claude Sonnet 3.7**	**Claude Sonnet 4**
Input	$3	$3
Output	$15	$15

Claude Sonnet 3.7 and Claude Sonnet 4 have the same pricing: **$3 per 1M input tokens** and **$15 per 1M output tokens**. This means there's no cost difference between the two models.

Users can upgrade to Claude Sonnet 4 for improved performance without paying extra, making it a straightforward, value-focused choice.

Conclusion

Claude Sonnet 4 represents a significant upgrade over Claude Sonnet 3.7, with enhanced reasoning, a vastly expanded output token limit, and improved performance in coding and complex task handling.

Both models share the same pricing and support multimodal inputs, making Claude Sonnet 4 a clear value proposition for users seeking more advanced capabilities without additional cost.

Whether for developers, businesses, or customer service applications, Claude Sonnet 4's improved context handling and natural interaction make it the better choice for demanding AI workflows and longer conversations, while Claude 3.7 remains a reliable option for solid general performance.