Claude vs GPT-5 vs Gemini: 2026 Coding Benchmarks

Edition of November 13, 2026 — facts verified November 13, 2026. Current edition always at https://tygartmedia.com/claude-vs-gpt-vs-gemini-coding-benchmark/.

Direct Answer (November 13, 2026): Anthropic's current coding and reasoning lineup is **Fable 5.2** (`claude-fable-5-2`, $8/$40), **Opus 5.5** (`claude-opus-5-5`, $4/$20), **Sonnet 5.5** (`claude-sonnet-5-5`, $2/$10), and **Haiku 5.5** (`claude-haiku-5-5`, $0.10/$0.50 ≤100K) — all with 1M context and 128K max output. The primary-source leaderboard below preserves our verified June 13, 2026 cross-vendor benchmark audit (Fable 5, Opus 4.8, GPT-5.5, Gemini 3.1 Pro) without re-attributing historical scores to newer releases.

Last refreshed: November 13, 2026 · Historical benchmark scores verified June 13, 2026

Official references: Anthropic models overview · Claude API pricing

As of June 13, 2026, the four models most often compared for coding work are **Claude Fable 5** and **Claude Opus 4.8** from Anthropic, **GPT-5.5** from OpenAI, and **Gemini 3.1 Pro** from Google. This page is a leaderboard built on one rule: every score below is taken from a vendor's own page or the benchmark's official model card that we fetched on the verification date, or it is marked as not published. Several vendors publish their benchmark tables as images rather than machine-readable text; where we could not read an official figure directly, we list the metric as *not machine-verifiable* and link to the source document instead of estimating. The result is a smaller table than most roundups, but every number in it is one you can click through and check.

November 2026 lineup update: Since those June 13 scores were verified, Anthropic's lineup has moved to **Fable 5.2** ($8/$40), **Opus 5.5** ($4/$20), **Sonnet 5.5** ($2/$10), and **Haiku 5.5** ($0.10/$0.50 ≤100K; $0.50/$2.50 >100K), superseding Fable 5.1, Fable 5, Opus 4.8, and Haiku 4.5. The historical benchmark scores below remain labeled under their June 2026 models and have not been re-attributed.

Models and pricing (specs verified June 13, 2026)

Models compared — verified specs without sticky dollars.

These columns are confirmed from each vendor's official model documentation. Claude prices, context windows, and cutoffs come from Anthropic's models overview and the AWS Bedrock model card; GPT-5.5 from OpenAI's developer docs; Gemini 3.1 Pro from Google's DeepMind model card and the Gemini API pricing page.

Model	API ID	Input / Output (per Mtok)	Context	Max output	Knowledge cutoff
Claude Fable 5	`claude-fable-5`	$10 / $50	1M	128K	Not stated on overview*
Claude Opus 4.8	`claude-opus-4-8`	$5 / $25	1M	128K	Jan 2026
GPT-5.5	`gpt-5.5`	$5 / $30	1,050,000	128K	Dec 1, 2025
Gemini 3.1 Pro	`gemini-3.1-pro-preview`	$2 / $12 (â‰¤200K)**	1M	64K	Not stated on model card

Anthropic's models overview lists Fable 5's specs and price but does not publish a knowledge-cutoff date for it in the table we fetched. Gemini 3.1 Pro uses tiered pricing: $2 / $12 per Mtok for prompts up to 200K tokens, rising to $4 / $18 for prompts above 200K tokens (Google AI pricing page). GPT-5.5 pricing rises to 2x input / 1.5x output above 272K input tokens (OpenAI developer docs). Claude Opus 4.8 (legacy — still listed)* offers an optional fast mode at $10 / $50 per Mtok (Anthropic).

Coding benchmark scores (primary-source only, June 2026 models)

Coding benchmark scores — primary-source only.

Each cell is either a figure we read directly from a primary source on June 13, 2026, or marked “not machine-verifiable” with the source you should consult. A blank-equivalent entry never means zero â€” it means the official figure was not available in readable form during verification. Note the harness and version differences called out in the footnotes: they make cross-vendor cells not strictly comparable.

Benchmark	Claude Fable 5	Claude Opus 4.8	GPT-5.5	Gemini 3.1 Pro
SWE-bench Verified	Not machine-verifiable (see system card)	Not machine-verifiable (see system card)	Not published in retrievable primary source	80.6%
SWE-bench Pro (Public)	Not machine-verifiable (see system card)	Not machine-verifiable (see system card)	Not published in retrievable primary source	54.2%
Terminal-Bench	Not machine-verifiable (see system card)	Not machine-verifiable (see system card)	83.4% (v2.1, Codex CLI harness)â€	68.5% (v2.0, Terminus-2 harness)
LiveCodeBench Pro	Not published in retrievable primary source	Not published in retrievable primary source	Not published in retrievable primary source	2887 Elo

â€ GPT-5.5’s Terminal-Bench 2.1 figure of 83.4% is the score Anthropic attributes to GPT-5.5 “with the Codex CLI harness” in a footnote on its Claude Opus 4.8 announcement page. It is a competitor-reported comparison, not a number we read from OpenAI directly. Google reports Gemini 3.1 Pro on Terminal-Bench *2.0* under the Terminus-2 harness (68.5%); because the version and harness differ, the Gemini and GPT-5.5 Terminal-Bench cells are not directly comparable. Gemini’s SWE-bench Verified (80.6%), SWE-bench Pro Public (54.2%), and LiveCodeBench Pro (2887 Elo) are single-attempt figures from Google’s official Gemini 3.1 Pro model card.

What we could not verify from a primary source

Anthropic publishes its coding comparison tables for Claude Opus 4.8 and Claude Fable 5 as images inside its announcement pages, and the full Claude Opus 4.8 System Card PDF exceeded our fetch size limit, so we could not machine-read those percentages on the verification date. OpenAI’s GPT-5.5 announcement page returned an access error to our fetcher, and its developer-docs model page lists specs and pricing but no benchmark scores. We have therefore left Claude’s and GPT-5.5’s SWE-bench figures out of the table rather than reproduce numbers we could not confirm at the source. For those figures, consult the primary documents linked in our source list: the Claude Opus 4.8 System Card, the Claude Fable 5 and Mythos 5 announcement, and OpenAI’s GPT-5.5 page. If you are choosing a model today, the verified spec table above (price, context, output, cutoff) is the part you can rely on without caveat.

How to read a coding leaderboard

How to read a coding leaderboard.

Three cautions apply to any 2026 coding comparison. First, harness matters: the same model scores differently on Terminal-Bench depending on whether it runs under Terminus-2, a Codex CLI scaffold, or a vendor’s internal agent, which is why we annotate every Terminal-Bench cell. Second, version matters: “Terminal-Bench 2.0” and “Terminal-Bench 2.1” are different test sets, and “SWE-bench Pro” public and full splits differ â€” a single percentage with no version is close to meaningless. Third, a headline score is one slice of behavior; long-horizon agentic coding, tool-call reliability, and context handling over a long session often decide real-world usefulness more than a single pass rate. Treat the verified cells here as a starting point, then test the shortlist on your own repository.

Which model has the highest published coding benchmark score in June 2026?

We cannot crown a single winner from primary sources alone, because Anthropic and OpenAI publish their coding scores in formats we could not machine-verify on June 13, 2026. From figures we could read directly, Google’s Gemini 3.1 Pro model card reports 80.6% on SWE-bench Verified and 54.2% on SWE-bench Pro (Public). Anthropic’s and OpenAI’s comparable figures are in their system cards and announcement pages, which we link in the sources; we did not reproduce them here because they were not readable at the source during verification.

What does Claude Fable 5 cost, and how is it different from Opus 4.8 *(legacy — still listed)*?

Claude Fable 5 (`claude-fable-5`) is priced at $10 per million input tokens and $50 per million output tokens, with a 1M-token context window and up to 128K output tokens (Anthropic models overview). Claude Opus 4.8 *(legacy — still listed)* (`claude-opus-4-8`) is the Opus-tier flagship at $5 / $25 per Mtok, also 1M context and 128K output, with a January 2026 knowledge cutoff. Fable 5 is Anthropic's most capable widely released model; Opus 4.8 *(legacy — still listed)* is the lower-priced model most teams will use for everyday agentic coding.

Lineup currency (November 13, 2026): Current API list: Fable 5.2 $8/$40, Opus 5.5 $4/$20, Sonnet 5.5 $2/$10, Haiku 5.5 $0.10/$0.50 (≤100K). Legacy (still listed on Anthropic's card): Fable 5.1 $10/$50, Haiku 4.5 $1/$5, Opus 4.8 $5/$25, Sonnet 4.6 $3/$15.

Why are some benchmark cells marked "not machine-verifiable" instead of showing a number?

Because this page only prints scores we could confirm from a primary source on the verification date. Several vendors render their benchmark tables as images, and one large system-card PDF exceeded our fetch limit, so the underlying percentages were not readable to us. Rather than copy figures from third-party trackers, we mark the cell and point you to the official document. It keeps the leaderboard honest at the cost of being shorter.

How do the context windows compare?

Claude Fable 5, Claude Opus 4.8, and Gemini 3.1 Pro each offer a 1M-token context window; GPT-5.5 offers 1,050,000 tokens. Maximum output is 128K tokens for Claude Fable 5, Claude Opus 4.8, and GPT-5.5, and 64K tokens for Gemini 3.1 Pro. Note that Claude Opus 4.8's context window is 200K on Microsoft Foundry specifically, per Anthropic's documentation.

Is Terminal-Bench comparable across these models?

Not cell-for-cell. Google reports Gemini 3.1 Pro on Terminal-Bench 2.0 under the Terminus-2 harness (68.5%), while the GPT-5.5 figure we show (83.4%) is Terminal-Bench 2.1 under a Codex CLI harness, as attributed by Anthropic. Different versions and different harnesses mean the two numbers should not be read as a head-to-head result.

Related on Tygart Media: Claude Code vs Cursor · Claude Code vs Codex.

■ Deploying Claude or AI Infrastructure in Your Business?

At Tygart Media, we engineer custom Model Context Protocol (MCP) servers, multi-model content pipelines, and AI operational systems. Explore our Claude AI Team Implementation Services or check out our complete Restoration Operations & AI Kit.

Primary & Reference Links Cited on This Page

- Anthropic models overview — `https://platform.claude.com/docs/en/about-claude/models/overview`
- Claude API pricing — `https://platform.claude.com/docs/en/about-claude/pricing`
- Tygart Media: Claude Code vs Cursor
- Tygart Media: Claude Code vs Codex
- Tygart Media: Claude AI Team Implementation Services
- Tygart Media: Restoration Operations & AI Kit