SOFTWARE ENGINEERING / FRONTIER MODELS

Choose the model for the shape of the work.

Claude Opus 5, Claude Fable 5, and every GPT-5.6 tier and effort level - compared with a strict boundary between measured evidence and editorial guidance.

Practical verdict

Sol max for the hardest terminal work. Fable for ambitious autonomous projects. Terra for the default queue.

Use Luna for speed. Opus 5 is a premium daily driver, but public evidence does not yet support a clean numeric rank against the others.

Read this correctly. Published numbers compare selected high-end configurations. No source publishes a complete model-by-effort matrix. The effort curves in this report are operating guidance, not invented scores.

The public coding-agent snapshot

GPT-5.6 Sol	80.0	Terminal-Bench 2.1: 88.8% max / 91.9% ultra; DeepSWE: 72.7%
GPT-5.6 Terra	77.4	Terminal-Bench 2.1: 87.4%; DeepSWE: 69.6%
Claude Fable 5	77.2	SWE-Bench Pro: 80.0%; Terminal-Bench 2.1: 83.1%
GPT-5.5	76.4	Terminal-Bench 2.1: 85.6%; DeepSWE: 67.0%
GPT-5.6 Luna	74.6	Terminal-Bench 2.1: 84.7%; DeepSWE: 67.2%
Claude Opus 5	Not reported	Anthropic internal SWE-bench subset: 91.7%, inside noise of Fable 91.3%; not comparable to public leaderboard

Interpretation. Sol max leads the shared Coding Agent Index by 2.8 points over Fable. Terra essentially ties Fable (77.4 vs 77.2) at much lower list price. Luna remains above Opus 4.8 in the same OpenAI table. Opus 5 requires repository-specific evaluation because its available SWE-bench subset is explicitly not comparable to the public leaderboard.

Vendor-run and vendor-selected evaluations are informative, not neutral. Different harnesses, tools and test subsets can reverse rankings.

All supported reasoning levels

Level	Coding use	Principal trade-off
low	Small fixes, search, formatting, fast subagent leaves	Speed and cost; weakest exploration
medium	Routine features, tests, ordinary refactors	Balanced daily engineering
high	Ambiguous bugs and multi-file implementation	Strong default; more latency/tool use
xhigh	Large migrations and repeated verification	Long horizon; diminishing returns on easy work
max	Frontier correctness, security, architecture	Deepest single agent; highest cost/latency
ultra	Parallel research, implementation and independent audit	Coordination overhead; only Sol and Terra in current Codex lineup

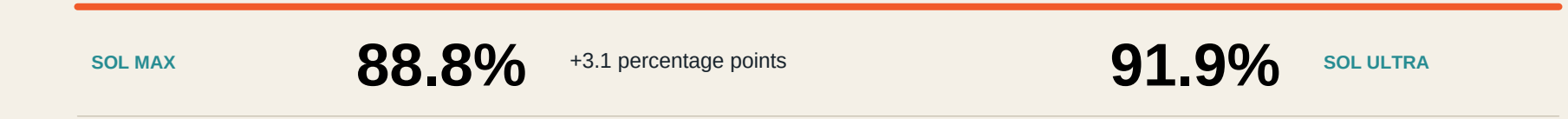
Model	low	medium	high	xhigh	max	ultra	Default
GPT-5.6 Sol	yes	yes	yes	yes	yes	yes	low
GPT-5.6 Terra	yes	yes	yes	yes	yes	yes	medium
GPT-5.6 Luna	yes	yes	yes	yes	yes	-	medium
Claude Opus 5	yes	yes	yes	yes	yes	-	high
Claude Fable 5	yes	yes	yes	yes	yes	-	high

Anthropic and OpenAI use similar effort names, but the names do not guarantee equal compute, latency or intelligence. Sweep effort separately for each model and workload.

03 / ULTRA

Parallelism helps only when the work splits cleanly

OpenAI says Ultra coordinates four agents in parallel by default. On Terminal-Bench 2.1, the published Sol result improves from 88.8% at max to 91.9% at ultra. The same page reports SEC-Bench Pro rising from 71.2% to 74.3%.



Research	Implement	Verify	Alternatives
Map modules, constraints and prior art.	Build independent low-collision leaves.	Run tests, inspect output and attack assumptions.	Explore competing fixes or architectures.

Use Ultra: broad audits, multi-component implementations, parallel research, independent verification, or work with clear file ownership.
Stay on max: tightly coupled debugging, a single architectural hotspot, sequential reproduction, or tasks where merge and synthesis overhead dominate.

A practical default stack

Work shape	Start here	Escalate when
Fast, scoped loop	Luna medium	Tests fail ambiguously or repository context expands
Everyday engineering	Terra high	The task crosses subsystems or needs deep verification
Hard terminal / correctness	Sol max	Independent branches can work in parallel -> Sol ultra
Large autonomous project	Fable high or xhigh	Maximum intelligence is worth its premium -> Fable max
Premium general daily driver	Opus 5 high	Demanding long-horizon coding -> xhigh or max

Price snapshot (USD per 1M input / output tokens)

Sol $5 / $30 | Terra $2.50 / $15 | Luna $1 / $6 | Opus 5 $5 / $25 | Fable 5 $10 / $50

Before caching, batch, fast-mode or regional modifiers. Total task cost depends on reasoning tokens, tool loops, retries, context reuse and orchestration overhead - not just list price.

Evaluation recipe

Take 20-50 representative repository tasks. Hold the harness, tools, permissions and success checks constant. Test at least medium/high/max in separate sessions; add Ultra only to decomposable tasks. Score functional correctness, regression rate, human review time, elapsed time and total cost. Route by observed task shape, then retry failures one level up.

Primary references

1. OpenAI - GPT-5.6 launch, benchmarks, pricing and Ultra
https://openai.com/index/gpt-5-6/

2. Anthropic - Claude Fable 5 overview and pricing
https://www.anthropic.com/claude/fable

3. Anthropic - What is new in Claude Opus 5
https://platform.claude.com/docs/en/models/opus-5/whats-new-opus-5

4. Anthropic - Effort parameter and model guidance
https://platform.claude.com/docs/en/build-with-claude/effort

5. Anthropic - Optimizing for cost and intelligence
https://platform.claude.com/docs/en/about-claude/models/optimizing-for-cost-and-intelligence

Method note. This report gives precedence to official product documentation and primary benchmark reporting. Where a model is missing from a shared table, it says so. Descriptions such as "daily driver" and the routing recommendations are reasoned editorial judgments, not vendor benchmark facts.