

MOSES™ x Artificial Analysis Coding Agent Benchmarks

Operator-augmented Claude Code + Opus 4.7 vs. 13 published combinations · Field: artificialanalysis.ai/agents/coding-agents · 2026-05-14

MOSES™ leads all 5 measured economic categories — 7-day window (2026-05-08 → 2026-05-14)

<p>CACHE HIT RATE</p> <p>96.88%</p> <p>#1 — > field's 96.2%</p> <p>SRC: Token Dashboard 7d</p>	<p>OUTPUT : INPUT</p> <p>31.7x</p> <p>30d 42.5x · 90d 22.1x</p> <p>#1 — 83x field leader (0.38)</p> <p>SRC: 3.90M out + 123K in</p>	<p>TOKENS / TASK</p> <p>767K</p> <p>#1 — 3.6x more efficient</p> <p>SRC: 1.12B + 1,465 tasks</p>	<p>TIME / TASK</p> <p>1.84 min</p> <p>#1 — 3.2x faster</p> <p>SRC: ~45 hr = 1,465 tasks</p>	<p>COST / LOC</p> <p>\$0.0007</p> <p>plan · API: \$0.044 · ccusage: \$0.018</p> <p>#1 — < 1¢ per line</p> <p>SRC: \$23.33 = 35,242 LOC</p>
--	--	---	--	--

Testing methodology — not the same thing being measured: AA field: per-task isolated runs on SWE-Bench-Pro-Hard-AA, Terminal-Bench v2, SWE-Atlas-QnA. Each task = one bug/issue. **MOSES™:** sustained operator work on App Hub (real product) over 7-day measurement window (2026-05-08 → 2026-05-14), 21 sessions, 7,327 turns. App Hub = 5 build days within the window (5/10-5/14), 35,242 LOC shipped. Cost/task (\$0.017 sub - \$1.05 API) computed similarly but not on cards. \$/LOC: AA = cost_per_task + 20 LOC industry convention. MOSES = \$23.33/wk sub + 35,242 actual LOC.

1. Cache Hit Rate

Cache reuse % - higher better - MOSES sustained multi-project

Model	Cache Hit Rate
MOSES™ CC+Opus 4.7+op	96.9%
Cursor CLI Opus 4.7 (Medium)	96.2%
Claude Code Opus 4.7 (Medium)	96.2%
Claude Code Kimi K2.6	96.1%
Codex GPT-5.5 (Medium)	94.9%
Claude Code Sonnet 4.6 (Medium)	94.5%
Claude Code Opus 4.6 (Medium)	93.7%
Codex GPT-5.4 (Medium)	92.8%
Cursor CLI Composer 2	91.7%
Cursor CLI GPT-5.5 (Medium)	87.8%
Gemini CLI Gemini 3.1 Pro (High)	86.1%
Cursor CLI GPT-5.4 (Medium)	85.3%
Claude Code GLM-5.1	83.7%
Claude Code DeepSeek V4 Pro (High)	79.8%

SRC: cache_read 1.084B + (1.084B + 34.83M cache_create + 123K input) = 96.88%

2. Output : Fresh Input (log scale)

Output tokens per fresh-input token - higher = denser signal

Model	Output : Fresh Input
MOSES™ CC+Opus 4.7+op	31.7x
Cursor CLI Opus 4.7 (Medium)	0.38
Claude Code Kimi K2.6	0.25
Claude Code Sonnet 4.6 (Medium)	0.24
Claude Code Opus 4.7 (Medium)	0.24
Codex GPT-5.5 (Medium)	0.17
Cursor CLI Composer Opus 5.6 (Medium)	0.16
Claude Code Opus 5.6 (Medium)	0.15
Codex GPT-5.4 (Medium)	0.15
Gemini CLI Gemini 3.1 Pro (High)	0.14
Cursor CLI GPT-5.5 (Medium)	0.07
Cursor CLI GPT-5.4 (Medium)	0.06
Claude Code GLM-5.1	0.05
Claude Code DeepSeek V4 Pro (High)	0.04

SRC: 3,902,803 output + 123,246 fresh input = 31.7x · 30d 42.5x · 90d/all-time 22.1x

3. Tokens per Task

Total tokens - lower better

Model	Tokens per Task
MOSES™ CC+Opus 4.7+op	767K
Cursor CLI GPT-5.5 (Medium)	2.74M
Cursor CLI Opus 4.7 (Medium)	2.93M
Gemini CLI Gemini 3.1 Pro (High)	3.24M
Cursor CLI Composer 2	3.33M
Claude Code Opus 4.7 (Medium)	3.33M
Cursor CLI GPT-5.4 (Medium)	3.75M
Claude Code Opus 4.6 (Medium)	4.27M
Claude Code Sonnet 4.6 (Medium)	4.41M
Codex GPT-5.4 (Medium)	4.92M
Codex GPT-5.5 (Medium)	5.42M
Claude Code DeepSeek V4 Pro (High)	6.20M
Claude Code Kimi K2.6	7.28M
Claude Code GLM-5.1	8.88M

SRC: 1.123B 7d total + 1,465 task-equivs (7,327 turns + 5) = 767K

4. Time per Task

Wall time - lower better

Model	Time per Task
MOSES™ CC+Opus 4.7+op	1.8m
Claude Code Opus 4.7 (Medium)	5.8m
Cursor CLI GPT-5.5 (Medium)	6.2m
Codex GPT-5.4 (Medium)	6.9m
Claude Code Opus 4.6 (Medium)	7.0m
Codex GPT-5.5 (Medium)	7.1m
Cursor CLI GPT-5.4 (Medium)	7.6m
Gemini CLI Gemini 3.1 Pro (High)	7.6m
Cursor CLI Opus 4.7 (Medium)	7.8m
Cursor CLI Composer 2	8.7m
Claude Code Sonnet 4.6 (Medium)	9.2m
Claude Code DeepSeek V4 Pro (High)	18.0m
Claude Code GLM-5.1	21.6m
Claude Code Kimi K2.6	41.5m

SRC: ~45 hr active + 1,465 = 1.84 min

5. Cost per LOC — all field models + MOSES™ (log scale)

USD per line shipped - lower better - MOSES plan basis \$0.0007 leads field

Model	Cost per LOC
MOSES™ Plan basis	\$0.0007
Cursor CLI Composer 2	\$0.0035
Claude Code DeepSeek V4 Pro (High)	\$0.018
Claude Code Kimi K2.6	\$0.038
MOSES™ API equiv basis	\$0.044
Cursor Industry low est.	\$0.050
Claude Code Sonnet 4.6 (Medium)	\$0.051
Claude Code Opus 4.7 (Medium)	\$0.062
Claude Code Opus 4.6 (Medium)	\$0.063
Cursor CLI Opus 4.7 (Medium)	\$0.073
Cursor CLI GPT-5.4 (Medium)	\$0.076
Gemini CLI Gemini 3.1 Pro (High)	\$0.080
Cursor CLI GPT-5.5 (Medium)	\$0.080
Codex GPT-5.4 (Medium)	\$0.10
Codex GPT-5.5 (Medium)	\$0.11
Claude Code GLM-5.1	\$0.11
Cursor Industry high est.	\$0.20
Devin Industry low est.	\$0.26
Devin Industry high est.	\$3.30

AA models: cost_per_task + 20 LOC · Cursor/Devin: industry estimates - MOSES plan: \$23.33/wk + 35,242 = \$0.000662 · MOSES API: \$1,564.47 + 35,242 = \$0.0444

RAW DATA — 7-DAY MEASUREMENT WINDOW (2026-05-08 → 2026-05-14)

<p>INPUT</p> <p>123,246</p> <p>FRESH TOKENS IN</p> <p>drives cache hit · out:in</p>	<p>OUTPUT</p> <p>3.90 M</p> <p>3,902,803</p> <p>drives out:in</p>	<p>CACHE CREATE</p> <p>34.83 M</p> <p>34,826,779</p> <p>drives cache hit</p>	<p>CACHE READ</p> <p>1.084 B</p> <p>1,084,399,183</p> <p>drives cache hit</p>	<p>TOTAL TOKENS</p> <p>1.123 B</p> <p>1,123,252,011</p> <p>drives tokens/task · time/task</p>
--	--	---	--	--

ancillary inputs > 21 sessions · 7,327 turns · 1,465 tasks · ~45 hr active · 35,242 LOC · \$23.33/wk plan \$1,564.47 API equiv

MOSES™ = operator-augmented Claude Code, concurrent sessions, \$100/mo Max plan · 7d window (2026-05-08 → 2026-05-14): 21 sessions · 7,327 turns · 1.12B tokens · \$1,564.47 API-equivalent · App Hub: 35,242 LOC in 5 days · Sources: Token Dashboard API, ccusage report, App Hub wc -l count, artificialanalysis.ai/agents/coding-agents