AI Agent Harness Benchmark, November 2026

Public SWE-bench and Terminal-Bench scores for agent harnesses and coding agents, each with the date on the official source. Read November 5, 2026.

Last updated November 5, 2026

This page copies public scores for coding agents and the harnesses around them. A harness is the loop, tools, and limits around a model. The tables use the official SWE-bench leaderboard and the official Terminal-Bench 4.0 leaderboard. Each row keeps the date on that source. Nothing here is estimated, and a score from another website is not filled in when the official board does not list it.

SWE-bench Verified is 500 real GitHub issues. The score is the percent resolved: the share of issues whose tests passed. The bash-only rows all use mini-SWE-agent, so the harness stays the same while the model changes. The SWE-bench site says release 2.x calls tools, while release 1.x reads the action out of the model's text, and that those two generations are not necessarily comparable. The version column is the release named on the row.

Terminal-Bench 4.0 is the current board on tbench.ai. The agent column changes by row, so those scores include Claude Code, Codex, Grok Build, mini-SWE-agent, or Muse Code, not one shared harness. Accuracy and the 95% half-width are the fields on the official leaderboard. The cost cell is copied as printed, including a partial-trial note where the board includes one.

The newest result date on the SWE-bench Verified file, as read on November 5, 2026, is 2026-02-26. The Terminal-Bench 4.0 board record says it was updated 2026-10-23. This November report is a reading of those boards. It is not a claim that either project published a new run in November.

SWE-bench Verified, bash only

47 rows from the Verified list where the harness is mini-SWE-agent. "Team checked" means the SWE-bench page marks the run as performed or checked by the SWE-bench team. A "No" means the row is on the official board without that mark. Source: swebench.com, read November 5, 2026.

SWE-bench Verified bash-only scores

Run	Harness	Model	% resolved	Result date	mini-SWE-agent	Effort	Run cost	Team checked	Open weights
Claude 4.5 Opus (high)	mini-SWE-agent	Claude 4.5 Opus	76.8%	2026-02-17	2.0.0	high	$376.95	No	No
Gemini 3 Flash (high)	mini-SWE-agent	Gemini 3 Flash	75.8%	2026-02-17	2.0.0	high	$177.98	No	No
MiniMax M2.5 (high)	mini-SWE-agent	MiniMax M2.5	75.8%	2026-02-17	2.0.0	high	$36.64	No	Yes
Claude 4.6 Opus	mini-SWE-agent	Claude 4.6 Opus	75.6%	2026-02-17	2.0.0	Not stated	$275.76	No	No
Claude 4.5 Opus (20251101) (medium)	mini-SWE-agent	Claude 4.5 Opus	74.4%	2025-11-24	1.16.0	medium	$360.62	Yes	No
Gemini 3 Pro Preview (2025-11-18)	mini-SWE-agent	Gemini 3 Pro Preview	74.2%	2025-11-18	1.15.0	Not stated	$229.98	Yes	No
GLM 5 (high)	mini-SWE-agent	GLM 5	72.8%	2026-02-17	2.0.0	high	$267.19	No	Yes
GPT 5.2 (high)	mini-SWE-agent	GPT 5.2	72.8%	2026-02-17	2.0.0	high	$236.78	No	No
GPT 5.2 Codex	mini-SWE-agent	GPT 5.2 Codex	72.8%	2026-02-19	2.0.0	Not stated	$224.71	No	No
GPT 5.2 (2025-12-11) (high)	mini-SWE-agent	GPT 5.2	71.8%	2025-12-11	1.17.2	high	$260.13	No	No
Claude 4.5 Sonnet (high)	mini-SWE-agent	Claude 4.5 Sonnet	71.4%	2026-02-17	2.0.0	high	$328.95	No	No
Kimi K2.5 (high)	mini-SWE-agent	Kimi K2.5	70.8%	2026-02-17	2.0.0	high	$73.28	No	Yes
Claude 4.5 Sonnet (20250929)	mini-SWE-agent	Claude 4.5 Sonnet	70.6%	2025-09-29	1.13.3	Not stated	$279.17	Yes	No
DeepSeek V3.2 (high)	mini-SWE-agent	DeepSeek V3.2	70%	2026-02-17	2.0.0	high	$223.92	No	Yes
Gemini 3 Pro	mini-SWE-agent	Gemini 3 Pro	69.6%	2026-02-26	2.0.0	high	$480.01	No	No
GPT 5.2 (2025-12-11)	mini-SWE-agent	GPT 5.2	69%	2025-12-11	1.17.2	Not stated	$134.83	No	No
Claude 4 Opus (20250514)	mini-SWE-agent	Claude 4 Opus	67.6%	2025-08-02	1.0.0	Not stated	$565.64	Yes	No
Claude 4.5 Haiku (high)	mini-SWE-agent	Claude 4.5 Haiku	66.6%	2026-02-17	2.0.0	high	$165.46	No	No
GPT 5.1 (2025-11-13) (medium)	mini-SWE-agent	GPT 5.1	66%	2025-11-20	1.15.0	medium	$153.08	No	No
GPT 5.1 Codex (medium)	mini-SWE-agent	GPT 5.1 Codex	66%	2025-11-24	1.16.0	medium	$294.44	No	No
GPT 5 (2025-08-07) (medium)	mini-SWE-agent	GPT 5	65%	2025-08-07	1.7.0	medium	$140.19	Yes	No
Claude 4 Sonnet (20250514)	mini-SWE-agent	Claude 4 Sonnet	64.93%	2025-07-26	1.0.0	Not stated	$185.73	Yes	No
Kimi K2 Thinking	mini-SWE-agent	Kimi K2 Thinking	63.4%	2025-12-10	1.17.2	Not stated	$219.15	Yes	Yes
MiniMax M2	mini-SWE-agent	MiniMax M2	61%	2025-11-24	1.17.0	Not stated	$214.13	Yes	Yes
DeepSeek V3.2 Reasoner	mini-SWE-agent	DeepSeek V3.2 Reasoner	60%	2025-12-01	1.17.1	Not stated	$14.04	Yes	Yes
GPT 5 mini (2025-08-07) (medium)	mini-SWE-agent	GPT 5 mini	59.8%	2025-08-07	1.7.0	medium	$17.74	Yes	No

o3 (2025-04-16)	mini-SWE-agent	o3	58.4%	2025-07-26	1.0.0	Not stated	$166.83	Yes	No
Devstral Small (2512)	mini-SWE-agent	Devstral Small (2512)	56.4%	2025-12-09	1.17.2	Not stated	Not listed	Yes	Yes
GPT 5 mini	mini-SWE-agent	GPT 5 mini	56.2%	2026-02-17	2.0.0	Not stated	$23.60	No	No
Qwen3-Coder 480B/A35B Instruct	mini-SWE-agent	Qwen3-Coder 480B/A35B Instruct	55.4%	2025-08-02	1.0.0	Not stated	$123.96	Yes	Yes
GLM 4.6 (T=1)	mini-SWE-agent	GLM 4.6	55.4%	2025-12-01	1.17.1	Not stated	$48.31	Yes	Yes
GLM 4.5 (2025-08-22)	mini-SWE-agent	GLM 4.5	54.2%	2025-08-22	1.9.1	Not stated	$148.60	Yes	Yes
Devstral (2512)	mini-SWE-agent	Devstral (2512)	53.8%	2025-12-09	1.17.2	Not stated	Not listed	Yes	Yes
Gemini 2.5 Pro (2025-05-06)	mini-SWE-agent	Gemini 2.5 Pro	53.6%	2025-07-26	1.0.0	Not stated	$144.19	Yes	No
Claude 3.7 Sonnet (20250219)	mini-SWE-agent	Claude 3.7 Sonnet	52.8%	2025-07-20	0.0.0	Not stated	$177.31	Yes	No
o4-mini (2025-04-16)	mini-SWE-agent	o4-mini	45%	2025-07-26	1.0.0	Not stated	$104.99	Yes	No
Kimi K2 Instruct	mini-SWE-agent	Kimi K2 Instruct	43.8%	2025-08-07	1.7.0	Not stated	$265.80	Yes	Yes
GPT 4.1 (2025-04-14)	mini-SWE-agent	GPT 4.1	39.58%	2025-07-26	1.0.0	Not stated	$73.16	Yes	No
GPT 5 nano (2025-08-07) (medium)	mini-SWE-agent	GPT 5 nano	34.8%	2025-08-07	1.7.0	medium	$19.04	Yes	No
Gemini 2.5 Flash (2025-04-17)	mini-SWE-agent	Gemini 2.5 Flash	28.73%	2025-07-26	1.0.0	Not stated	$66.94	Yes	No
gpt-oss-120b	mini-SWE-agent	gpt-oss-120b	26%	2025-08-07	1.7.0	Not stated	$28.56	Yes	Yes
GPT 4.1 mini (2025-04-14)	mini-SWE-agent	GPT 4.1 mini	23.94%	2025-07-20	0.0.0	Not stated	$217.58	Yes	No
GPT 4o (2024-11-20)	mini-SWE-agent	GPT 4o	21.62%	2025-07-20	0.0.0	Not stated	$765.57	Yes	No
Llama 4 Maverick Instruct	mini-SWE-agent	Llama 4 Maverick Instruct	21.04%	2025-07-20	0.0.0	Not stated	$156.85	Yes	No
Gemini 2.0 Flash	mini-SWE-agent	Gemini 2.0 Flash	13.52%	2025-07-26	0.0.0	Not stated	$0.00	Yes	No
Llama 4 Scout Instruct	mini-SWE-agent	Llama 4 Scout Instruct	9.06%	2025-07-20	0.0.0	Not stated	$57.82	Yes	No
Qwen2.5-Coder 32B Instruct	mini-SWE-agent	Qwen2.5-Coder 32B Instruct	9%	2025-08-03	1.0.0	Not stated	$34.06	Yes	Yes

Other harnesses the SWE-bench team checked

30 Verified rows that use a harness other than mini-SWE-agent and that the board marks as team-checked.

Team-checked SWE-bench Verified rows outside the bash-only harness

Run	Harness	Model	% resolved	Result date	mini-SWE-agent	Effort	Run cost	Team checked	Open weights
OpenHands + GPT-5	OpenHands	GPT-5	71.8%	2025-08-07	-	Not stated	Not listed	Yes	No
Moatless Tools + Claude 4 Sonnet	Moatless Tools	Claude 4 Sonnet	70.8%	2025-06-11	-	Not stated	Not listed	Yes	No
OpenHands + Claude 4 Sonnet	OpenHands	Claude 4 Sonnet	70.4%	2025-05-24	-	Not stated	Not listed	Yes	No
SWE-agent + Claude 4 Sonnet	SWE-agent	Claude 4 Sonnet	66.6%	2025-05-22	-	Not stated	Not listed	Yes	No
OpenHands	OpenHands	Undisclosed	65.8%	2025-04-15	-	Not stated	Not listed	Yes	No
OpenHands + Kimi K2	OpenHands	Kimi K2	65.4%	2025-07-16	-	Not stated	Not listed	Yes	Yes
SWE-agent + Claude 3.7 Sonnet w/ Review Heavy	SWE-agent	Claude 3.7 Sonnet w/ Review Heavy	62.4%	2025-02-25	-	Not stated	Not listed	Yes	No
OpenHands + 4x Scaled (2024-02-03)	OpenHands	4x Scaled	60.8%	2025-02-03	-	Not stated	Not listed	Yes	No
Nemotron-CORTEXA	Nemotron-CORTEXA	Multiple	58.2%	2025-04-10	-	Not stated	Not listed	Yes	No
OpenHands + CodeAct v2.1 (claude-3-5-sonnet-20241022)	OpenHands	CodeAct v2.1 (claude-3-5-sonnet-20241022)	53%	2024-10-29	-	Not stated	Not listed	Yes	No
Composio SWE-Kit (2024-10-25)	Composio SWE-Kit	Multiple	48.6%	2024-10-25	-	Not stated	Not listed	Yes	No
AppMap Navie v2	AppMap Navie v2	Multiple	47.2%	2024-11-06	-	Not stated	Not listed	Yes	No
Skywork-SWE-32B + TTS(Bo8)	Skywork-SWE-32B	TTS(Bo8)	47%	2025-06-16	-	Not stated	Not listed	Yes	Yes
OpenHands + DevStral Small 2505	OpenHands	DevStral Small 2505	46.8%	2025-05-20	-	Not stated	Not listed	Yes	Yes
Agentless Lite + O3 Mini (20250214)	Agentless Lite	O3 Mini	42.4%	2025-02-14	-	Not stated	Not listed	Yes	No
SWE-agent + SWE-agent-LM-32B	SWE-agent	SWE-agent-LM-32B	40.2%	2025-05-11	-	Not stated	Not listed	Yes	Yes
Skywork-SWE-32B	Skywork-SWE-32B	Qwen2.5 Coder 32B Instruct	38%	2025-06-16	-	Not stated	Not listed	Yes	Yes

SWE-agent + DevStral Small 2507	SWE-agent	DevStral Small 2507	38%	2025-07-25	-	Not stated	Not listed	Yes	Yes
SWE-agent + Claude 3.5 Sonnet	SWE-agent	Claude 3.5 Sonnet	33.6%	2024-06-20	-	Not stated	Not listed	Yes	No
SWE-Fixer (Qwen2.5-7b retriever + Qwen2.5-72b editor)	SWE-Fixer	Qwen2.5 (7B + 72B)	32.8%	2025-03-06	-	Not stated	Not listed	Yes	Yes
AppMap Navie + GPT 4o (2024-05-13)	AppMap Navie	GPT-4o	26.2%	2024-06-15	-	Not stated	Not listed	Yes	No
SWE-agent + GPT 4o (2024-05-13)	SWE-agent	GPT-4o	23.2%	2024-07-28	-	Not stated	Not listed	Yes	No
SWE-agent + GPT 4 (1106)	SWE-agent	GPT-4 (1106)	22.4%	2024-04-02	-	Not stated	Not listed	Yes	No
SWE-agent + Claude 3 Opus	SWE-agent	Claude 3 Opus	15.8%	2024-04-02	-	Not stated	Not listed	Yes	No
RAG + Claude 3 Opus	RAG baseline	Claude 3 Opus	7%	2024-04-02	-	Not stated	Not listed	Yes	No
RAG + Claude 2	RAG baseline	Claude 2	4.4%	2023-10-10	-	Not stated	Not listed	Yes	No
RAG + GPT 4 (1106)	RAG baseline	GPT-4 (1106)	2.8%	2024-04-02	-	Not stated	Not listed	Yes	No
RAG + SWE-Llama 7B	RAG baseline	SWE-Llama 7B	1.4%	2023-10-10	-	Not stated	Not listed	Yes	No
RAG + SWE-Llama 13B	RAG baseline	SWE-Llama 13B	1.2%	2023-10-10	-	Not stated	Not listed	Yes	No
RAG + ChatGPT 3.5	RAG baseline	GPT-3.5	0.4%	2023-10-10	-	Not stated	Not listed	Yes	No

SWE-bench Multilingual, same harness

300 instances, 13 mini-SWE-agent rows on the same official file. This is a separate task set from Verified, so the percents are not a second reading of the 500 Verified issues.

SWE-bench Multilingual mini-SWE-agent scores

Run	Harness	Model	% resolved	Result date	mini-SWE-agent	Effort	Run cost	Team checked	Open weights
Gemini 3 Flash	mini-SWE-agent	Gemini 3 Flash	72.7%	2026-02-13	2.0.0a0	Not stated	$105.55	Yes	No
Claude 4.6 Opus	mini-SWE-agent	Claude 4.6 Opus	72%	2026-02-13	2.0.0a0	Not stated	$198.89	Yes	No
Claude 4.5 Opus	mini-SWE-agent	Claude 4.5 Opus	70.7%	2026-02-13	2.0.0a0	Not stated	$250.07	Yes	No
GLM 5	mini-SWE-agent	GLM 5	69.7%	2026-02-13	2.0.0a0	Not stated	$192.74	Yes	Yes
Gemini 3 Pro	mini-SWE-agent	Gemini 3 Pro	68.7%	2026-02-13	2.0.0a0	Not stated	$306.33	Yes	No
MiniMax M2.5	mini-SWE-agent	MiniMax M2.5	68.3%	2026-02-16	2.0.0a0	Not stated	$30.06	Yes	Yes
Kimi K2.5	mini-SWE-agent	Kimi K2.5	67.3%	2026-02-13	2.0.0a0	Not stated	$207.85	Yes	Yes
Claude 4.5 Sonnet	mini-SWE-agent	Claude 4.5 Sonnet	67%	2026-02-13	2.0.0a0	Not stated	$200.86	Yes	No
GPT 5.2 (high)	mini-SWE-agent	GPT 5.2	66.7%	2026-02-13	2.0.0a0	high	$160.92	Yes	No
GPT 5.2 Codex	mini-SWE-agent	GPT 5.2 Codex	66.3%	2026-02-20	2.0.0	Not stated	$198.65	No	No
Claude 4.5 Haiku	mini-SWE-agent	Claude 4.5 Haiku	64.7%	2026-02-13	2.0.0a0	Not stated	$113.85	Yes	No
DeepSeek V3.2	mini-SWE-agent	DeepSeek V3.2	59%	2026-02-13	2.0.0a0	Not stated	$115.15	Yes	Yes
GPT 5 mini	mini-SWE-agent	GPT 5 mini	39.7%	2026-02-13	2.0.0a0	Not stated	$15.49	Yes	No

Terminal-Bench 4.0

35 rows. The official leaderboard for Terminal-Bench 4.0. The board name on Harbor Hub is 4-0-0, and the record was updated 2026-10-23. Read November 5, 2026.

Terminal-Bench 4.0 official leaderboard

Rank	Model	Agent	Effort	Accuracy	95% half-width	Successes / trials	Cost	Date
1	Opus 5.5	Claude Code	max	64.85%	3.11	214 / 330	$4,718.65	2026-09-22
2	Sonnet 5.5	Claude Code	max	61.82%	2.94	204 / 330	$7,333.64	2026-09-28
3	GPT-6 Astra	Codex	max	58.18%	2.79	192 / 330	$3,267.18	2026-09-03
3	GPT-6.1 Sol	Codex	max	58.18%	3.11	192 / 330	$634.24	2026-09-29
5	Fable 5.1	Claude Code	max	57.88%	3.76	191 / 330	$6,243.50	2026-09-01
5	GPT-6 Astra	Codex	xhigh	57.88%	2.72	191 / 330	$2,350.51	2026-09-03
5	GPT-6 Astra	Codex	high	57.88%	2.97	191 / 330	$2,269.42	2026-09-03
5	Fable 5.1	Claude Code	xhigh	57.88%	3.36	191 / 330	$4,872.04	2026-09-01
9	Fable 5.1	Claude Code	high	54.55%	3.44	180 / 330	$3,985.38	2026-09-01
10	GPT-6 Astra	Codex	medium	54.24%	2.66	179 / 330	$1,914.80	2026-09-03
11	Fable 5.1	Claude Code	medium	53.94%	3.39	178 / 330	$2,832.90	2026-09-01
11	Opus 5	Claude Code	xhigh	53.94%	3.17	178 / 330	$6,086.22	2026-07-24
13	Opus 5	Claude Code	max	51.82%	3.39	171 / 330	$5,969.11	2026-07-24
14	GPT-6 Astra	Codex	low	50.61%	2.75	167 / 330	$1,557.30	2026-09-03
15	Opus 5	Claude Code	high	50.3%	3.73	166 / 330	$4,662.27	2026-07-24
16	GPT-6 Sol	Codex	max	49.39%	3.23	163 / 330	$1,451.32	2026-09-22
17	Opus 5	Claude Code	medium	44.85%	3.83	148 / 330	$3,191.59	2026-07-24
18	Fable 5	Claude Code	max	44.55%	3.85	147 / 330	$7,265.01	2026-06-09
19	Fable 5.1	Claude Code	low	43.33%	3.61	143 / 330	$2,358.72	2026-09-01
20	GLM-5.3	Claude Code	max	41.82%	3.23	138 / 330	$2,727.63	2026-08-14
21	Grok 4.7	Grok Build	xhigh	37.58%	3.54	124 / 330	$3,683.29 (partial: 324/330 trials)	2026-09-21
22	GPT-5.6 Sol	Codex	max	37.27%	3.78	123 / 330	$2,541.70	2026-06-26
23	GLM-5.3-Flash	Claude Code	none	35.76%	3.54	118 / 330	$118.74	2026-08-26
24	Opus 5	Claude Code	low	34.85%	3.94	115 / 330	$2,393.88	2026-07-24
25	Qwen3.8-Max-0902	Claude Code	max	26.97%	3.78	89 / 330	$4,635.67	2026-09-02

26	Opus 4.8	Claude Code	max	23.64%	3.56	78 / 330	$6,481.26	2026-05-28
27	GPT-5.6 Terra	Codex	max	21.52%	3.25	71 / 330	$1,733.52	2026-06-26
28	Grok 4.6	Grok Build	high	20.3%	3.09	67 / 330	$3,591.58	2026-08-12
29	Gemini 3.8 Flash	mini-SWE-agent	high	19.09%	3.36	63 / 330	$1,828.77	2026-09-02
30	GPT-5.6 Luna	Codex	max	17.27%	2.85	57 / 330	$346.67	2026-06-26
31	GPT-6 Luna	Codex	max	16.36%	2.72	54 / 330	$72.14	2026-09-22
32	Muse Spark 1.3	Muse Code	xhigh	14.55%	2.94	48 / 330	$0.00	2026-09-02
33	Grok 4.5	Grok Build	high	12.42%	2.62	41 / 330	$2,094.11	2026-07-16
33	Sonnet 5	Claude Code	max	12.42%	3.06	41 / 330	$9,603.86	2026-06-30
35	Gemini 3.7 Flash	mini-SWE-agent	high	11.21%	2.45	37 / 330	$1,261.87	2026-08-13

Left out

- Third-party score roundups, including pages that list SWE-bench results above the official board, are not used. If the official file does not contain the row, it is not on this page.
- SWE-bench Full, Lite, and Multimodal are on swebench.com and were not copied into this month's tables.
- OSWorld, WebArena, GAIA, and SWE-bench Pro were not read from an official board for this edition, so they are not given scores here.
- The November 3, 2026 paper "You don't need long context" reports its own SWE-bench Verified and Terminal-Bench 2.1 runs. Those are the authors' runs, not rows on the public boards, and they stay on the research note.