

2022-01
(GPT-3.5)

2022-11
(GPT-4)

2023-09
(gpt-4-1106)

2024-02
(Claude 3.5 Sonnet)

2024-07
(o1-preview)

2024-12
(Claude 3.7 Sonnet)

2025-05
(Claude 4 Sonnet)

2025-10
(Claude 4.5 Opus)

METR

Frontier: 1.4 hr
Avg a: 3.92, Avg b: -1.44

SWE-bench

Frontier: 2.0 hr
Avg a: 1.80, Avg b: 0.14

MLE-bench

Frontier: 2.5 hr
Avg a: 3.15, Avg b: 1.30

GDPval

Frontier: 1.8 hr
Avg a: 1.10, Avg b: -0.50

All Tasks

Frontier: 2.0 hr
Avg a: 2.17, Avg b: -0.13

Success Probability

15 sec 1 min 4 min 15 min 1 hr 4 hrs 16 hrs

15 sec 1 min 4 min 15 min 1 hr 4 hrs 16 hrs

15 sec 1 min 4 min 15 min 1 hr 4 hrs 16 hrs

15 sec 1 min 4 min 15 min 1 hr 4 hrs 16 hrs

15 sec 1 min 4 min 15 min 1 hr 4 hrs 16 hrs

Task Length