

zapier/AutomationBench

PRIMARY

Publisher	GitHub
Published	2026
Retrieved	2026-10-01
Cite key	github2026zapieraut
URL	https://github.com/zapier/AutomationBench

Assessment. The benchmark maintainer's own public leaderboard. The decisive comparator: the best independently listed strict-metric score is 30.33%, and Argon is absent.

Public leaderboard over the 600-task public split. Metric is Pass Rate, 1.0 only if every assertion passes; models run at highest available reasoning effort. Claude Opus 4.8 30.33%; GPT-5.6 Sol 29.17%; GPT-5.6 Terra 25.83%; Claude Fable 5 25.83% (after the July 2026 stricter classifier); Claude Sonnet 5 24.00%; GLM 5.2 20.33%; Gemini 3.5 Flash 14.83%. Gemini 4 Argon does not appear. No explicit last-updated date in the README.