

AutomationBench-AA: Agentic SaaS Workflow Benchmark

PRIMARY

Publisher	Artificial Analysis
Published	2026
Retrieved	2026-10-01
Cite key	artificiala2026automatio
URL	https://artificialanalysis.ai/evaluations/automationbench-aa

Assessment. The second benchmark sharing the name. Establishes that 'AutomationBench' is ambiguous and that the two metrics are not comparable. No explicit update date published, so the leaderboard snapshot is as retrieved.

657 tasks across Finance, HR, Marketing, Operations, Sales and Support, using simulated Gmail, Google Sheets, Slack, Salesforce, Zendesk, Jira and HubSpot. Scoring differs from Zapier's: where Zapier reports the percentage of tasks completed fully, AutomationBench-AA uses the average share of each task's objectives a model completes without triggering guardrail violations. Top entries: Claude Sonnet 5.5 (Adaptive Reasoning, Max Effort) 71.3%; Claude Opus 5.5 69.5%; DeepSeek V4.1 Flash 68.9%. Gemini 4 Argon does not appear on the visible leaderboard.