

Source dossier

Nobody else ran it — Gemini 4 Argon claims first place on a benchmark with two metrics, and the duty that would ask you to supervise is fourteen months out

Edition 2026-10-01 · narrative: A vendor-reported benchmark score is a claim about a metric and about who ran it; the EU duty in force today makes you declare an agent, not supervise it. · compiled 2026-10-01

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-01-01	Google announced Gemini 4 Argon on 30 September 2026, rolling out first to a set of trusted cyber defenders through its Fairwind Program, with broader access to follow for paid API customers and Google AI Ultra subscribers.	verified	bloggoogle2026gemini, axios2026googleb
CLAIM-2026-10-01-02	In its own announcement Google reports Gemini 4 Argon at 77.9% on DeepSWE v1.1, ranked #1 on AutomationBench with a score of 51.3%, tied for first on CWE-bench v1 at 68%, and 91.7% on LVBench; introductory API pricing is \$2 per million input tokens and \$10 per million output tokens, rising to \$4 and \$20 after the introductory period.	verified	bloggoogle2026gemini
CLAIM-2026-10-01-03	AutomationBench is a Zapier benchmark (Daniel Shepard and Robin Salimans, both Zapier, April 2026) evaluating AI agents on cross-application workflow orchestration via REST APIs across Sales, Marketing, Operations, Support, Finance and HR; it comprises 600 public tasks plus 600+ private, roughly 500 endpoints across 47 apps, and grading is programmatic and end-state only with binary pass/fail and no partial credit.	verified	arxiv2026automation, github2026zapieraut
CLAIM-2026-10-01-04	The AutomationBench paper states that even the best frontier models currently score below 10%, reporting Opus 4.7 at 9.9%, Gemini 3.1 Pro at 9.6% and GPT 5.4 at 7.6%.	verified	arxiv2026automation
CLAIM-2026-10-01-05	Zapier's public AutomationBench leaderboard over the 600-task public split, using a strict pass rate scored 1.0 only if every assertion passes, is topped by Claude Opus 4.8 at 30.33%, followed by GPT-5.6 Sol at 29.17%, GPT-5.6 Terra and Claude Fable 5 at 25.83%, Claude Sonnet 5 at 24.00%, GLM 5.2 at 20.33% and Gemini 3.5 Flash at 14.83%. Gemini 4 Argon does not appear on it.	verified	github2026zapieraut

Claim	Statement	Status	Sources
CLAIM-2026-10-01-06	AutomationBench-AA, Artificial Analysis's variant, covers 657 tasks across the same six domains using simulated Gmail, Google Sheets, Slack, Salesforce, Zendesk, Jira and HubSpot, and scores the average share of each task's objectives a model completes without triggering guardrail violations rather than full-task completion. Its top entries are Claude Sonnet 5.5 at 71.3%, Claude Opus 5.5 at 69.5% and DeepSeek V4.1 Flash at 68.9%. Gemini 4 Argon does not appear on it.	verified	artificiala2026automatio
CLAIM-2026-10-01-07	Google's 51.3% AutomationBench result and its #1 ranking are self-reported and were not independently reproduced on either public leaderboard at the time of writing; which of the two scoring metrics the figure was computed under is not stated in Google's announcement, though a score of 51.3% is consistent with first place only on Zapier's strict pass-rate metric, since AutomationBench-AA's leading entries exceed it.	unverified	bloggoogle2026gemini, github2026zapiernaut, artificiala2026automatio
CLAIM-2026-10-01-08	AI Act Article 52(1) requires a provider to notify the Commission without delay and in any event within two weeks after a general-purpose AI model meets the systemic-risk classification requirement; Article 52(2) and (3) allow the provider to contest the classification and the Commission to reject insufficiently substantiated arguments; Article 52(6) requires the Commission to publish and keep updated a list of general-purpose AI models with systemic risk.	verified	artificiali2026article
CLAIM-2026-10-01-09	AI Act Article 50 transparency obligations have applied since 2 August 2026 and the machine-readable marking obligation for generated content applies from 2 December 2026. The Digital Omnibus, Regulation (EU) 2026/1744, in force 27 July 2026, moved Annex III high-risk obligations to 2 December 2027 and Annex I to 2 August 2028, so Article 14 human oversight for Annex III systems including employment and workers management is not binding until 2 December 2027.	verified	artificiali2026article
CLAIM-2026-10-01-10	GDPR Article 5(1)(d) requires personal data to be accurate and, where necessary, kept up to date, and that every reasonable step be taken to ensure that inaccurate personal data are erased or rectified without delay. It is subject to no deferral under the Digital Omnibus.	verified	gdprinfoeu2026art

Only claims marked **verified** may carry a post. **single-source** may appear as hedged context; **disputed** must not be published at all.

Gemini 4 Argon: our next era of frontier intelligence

PRIMARY

Publisher blog.google

Published 2026-09-30

Retrieved 2026-10-01

Cite key bloggogoogle2026gemini

URL <https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-argon/>

Assessment. The vendor's own announcement and the only source for the benchmark figures quoted. Primary for what Google claims; not independent evidence that the claims hold.

Argon is rolling out to a set of trusted cyber defenders through the Fairwind Program, with paid API customers and Google AI Ultra subscribers next. Reported results: DeepSWE v1.1 77.9%; AutomationBench ranked #1 with a score of 51.3%; CWE-bench v1 tied for first with a top score of 68%; LVBench 91.7%; leading on Gray Swan's indirect prompt injection benchmark, and on the Vals Index, Vals Finance Agent v2 and Harvey's Legal Agent Benchmark. Introductory pricing \$2 per million input tokens and \$10 per million output tokens, cached input at 95% off, rising to \$4 and \$20 after the introductory period. References the Frontier Safety Framework and robustness testing by internal and external red teams. No statement about EU availability.

Google unveils Gemini 4, long-awaited answer to OpenAI and Anthropic

SECONDARY

Publisher	Axios
Published	2026-09-30
Retrieved	2026-10-01
Cite key	axios2026google
URL	https://www.axios.com/2026/09/30/google-gemini-4

Assessment. Independent same-day coverage. Used for the staged-release framing and the disputed Bloomberg report; no benchmark numbers of its own.

Reports the initial rollout to a small group of cybersecurity partners and Google's participation in the US government's voluntary pre-release process, targeting complex long-running tasks in software engineering, finance, legal work and cybersecurity. Notes Bloomberg's report that some employees found performance lacking in internal testing, which Google disputes. Contains no verbatim benchmark figures.

AutomationBench

PRIMARY

Publisher	arXiv
Published	2026-04
Retrieved	2026-10-01
Cite key	arxiv2026automatio
URL	https://arxiv.org/abs/2604.18934

Assessment. The benchmark's own paper. Affiliations taken from the work's front matter: both authors at Zapier. Establishes the construction, the grading rule and the April baseline.

Abstract: 'Existing AI benchmarks for software automation rarely combine cross-application coordination, autonomous API discovery, and policy adherence. Real business workflows demand all three: a single task may span a CRM, inbox, calendar, and messaging platform - requiring the agent to find the right endpoints, follow a policy document, and write correct data to each system. To address this gap, we introduce AutomationBench, a benchmark for evaluating AI agents on cross-application workflow orchestration via REST APIs. Drawing on real workflow patterns from Zapier's platform, tasks span Sales, Marketing, Operations, Support, Finance, and HR domains. Agents must discover relevant endpoints themselves, follow layered business rules, and navigate environments with irrelevant and sometimes misleading records. Grading is programmatic and end-state only: whether the correct data ended up in the right systems. Even the best frontier models currently score below 10%. AutomationBench provides a challenging, realistic measure of where current models stand relative to the agentic capabilities businesses actually need.' 600 public tasks plus 600+ private; roughly 500 endpoints across 47 apps; binary pass/fail, no partial credit. Reported scores: Opus 4.7 9.9%, Gemini 3.1 Pro 9.6%, GPT 5.4 7.6%.

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.

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.

Article 52: Procedure

PRIMARY

Publisher	artificialintelligenceact.eu
Published	2026
Retrieved	2026-10-01
Cite key	artificiali2026article
URL	https://artificialintelligenceact.eu/article/52/

Assessment. Consolidated text of the Article. Used for the notification clock and the Commission's publication duty.

52(1): providers must notify the Commission 'without delay and in any event within two weeks after that requirement is met'. 52(2): the provider may submit arguments that the model does not present systemic risks due to its specific characteristics. 52(3): where arguments are insufficiently substantiated the Commission rejects them and the model is classified as presenting systemic risk. 52(6): the Commission must publish and keep updated a list of general-purpose AI models with systemic risk, while protecting intellectual property and confidential information.

Art. 5 GDPR — Principles relating to processing of personal data

PRIMARY

Publisher gdpr-info.eu

Published 2026

Retrieved 2026-10-01

Cite key gdprinfoeu2026art

URL <https://gdpr-info.eu/art-5-gdpr/>

Assessment. Used for the accuracy principle, the limb that reaches a wrong record written by an agent and that carries no deferral.

Article 5(1)(d): personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay.