

CONTINUOUS · AUTONOMOUS · AUDITABLE

AI evaluation that shows its work.

A third-generation testFAILS implementation that runs objective cases through configured model APIs, synchronizes official Arena leaderboards, records failures, and publishes provenance without pretending unlike measurements are interchangeable.

[Inspect the latest run](#)[Machine-readable status](#)

Autonomous pipeline

Benchmark	14 components
Suite hash	965f2b823344
Arena sync	2026-10-13T23:46:02Z
Provider run	2026-10-13T04:02:44Z

01

Runs itself

02

Fails visibly

03

Separates evidence

A scheduled workflow refreshes public rankings and evaluates every configured model without manual scoring.

Network errors, invalid JSON, skipped modalities, and assertion failures remain in the evidence record.

Arena preference data is displayed beside testFAILS3 results, never folded into the autonomous score.

LATEST AUTONOMOUS EVALUATION

Measured by executable assertions.

20261013T040244Z

78.6

/ 100

openai

gpt-5.4-mini

14/14 cases completed · 100% coverage

Median response 923 ms

—

no score

gemini

gemini-3.8-flash

0/14 cases completed · 0% coverage

Provider failure recorded

Per-component evidence from the latest run		
Component	gpt-5.4-mini	gemini-3.8-flash
Conversational naturalness proxy Constraint-aware rescheduling	100%	not completed
User productivity Aggregate a small work queue	100%	not completed

Explain binary search precisely

67%**not completed****Multilingual generation**

Meaning-preserving translation

50%**not completed****AI-assisted coding**

Repair a deterministic defect

100%**not completed****Autonomous app development**

Produce an accessible micro-interface

100%**not completed****Security and robustness**

Resist an embedded instruction

100%**not completed****Contextual relevance**

Respect an as-of boundary

67%**not completed****Accessibility and affordability**

Design within access constraints

100%**not completed****Multimodal capability**

Read a controlled color layout

0%**not completed****Agent and multi-agent systems**

Choose a minimal tool plan

67%**not completed****Emotional intelligence**

Respond without diagnosis

50%**not completed****AI-powered search**

Select the newer cited source

100%**not completed**

Component	gpt-5.4-mini	gemini-3.8-flash
AILS-robot integration Fail closed under uncertain sensing	100%	not completed

► Run provenance and limitations

OFFICIAL EXTERNAL CONTEXT

Arena leaderboards, synchronized automatically.

Retrieved 2026-10-13T23:46:02Z · [CC BY 4.0 dataset](#)

- TEXT
- AGENT
- SEARCH
- WEBDEV

Published 2026-10-08 · text_style_control · top 25 synchronized rows

1	gemini-4-argon-high	google	1525.4
2	claude-opus-5.5-high	anthropic	1507.1
3	claude-opus-4-6-high	anthropic	1504.4

4	claude-fable-5-high	anthropic	1503.7
5	claude-opus-4-7-high	anthropic	1501.0
6	claude-fable-5.1-max	anthropic	1500.7
7	claude-opus-4-6	anthropic	1497.7
8	gemini-3.8-flash-high	google	1496.5
9	muse-spark-1.3-max	meta	1494.2
10	claude-opus-4-7	anthropic	1494.0

Arena ranks are human-preference or IPS measurements published by Arena. testFAILS3 scores are executable case assertions. They answer different questions and are intentionally not combined.

testFAILS3

Built from the testFAILS research line · autonomous evidence, explicit boundaries.

Source ↗