Meta

Muse Ember

Evaluation Methodology

We evaluate Muse Ember on benchmarks spanning general intelligence, agents, coding, and multimodal understanding.

Overall Methodology

- **Models and configurations.** Muse Ember results are generated with max reasoning effort through the Meta Model API. We compare with Claude Haiku 5.5 and GPT-6 Luna at max reasoning effort. For controlled internal comparisons, every model uses the same benchmark-specific harness. External coding results retain the harness or agent product used by the reporting vendor or model provider, and that source is identified below.
- **Evaluation and score selection.** We preserve each benchmark's tasks, inputs, grading, and primary metric as closely as possible. Benchmark names link to official leaderboards where available. Truncated responses are graded as-is and model refusals receive zero credit and remain in the denominator rather than being dropped; samples lost to infrastructure or grading errors are not scored. Material departures from the official setup, including any benchmark that scores these cases differently, are noted below.
- **Competitor-model results.** Internally generated competitor runs use common settings where practical but remain best-effort: prompts, tools, and runtime may not be optimally tuned for proprietary models.

Per-benchmark details

General

- **Artificial Analysis Index.** The Artificial Analysis Intelligence Index v5.0 combines AA-Briefcase v1.1, AutomationBench, GDP.pdf, FrontierCode 2.0, Terminal-Bench

4.0, Terminal-Bench-Science 0.1, CritPt v2, AA-Omniscience v1.1 and HLE-Diamond into a single score. The result source is Artificial Analysis.

Agents

- **AutomationBench-AA.** AutomationBench-AA is Artificial Analysis's run of Zapier's private set of 657 business-workflow tasks across simulated SaaS applications such as Gmail, Salesforce and Slack. Agents complete each workflow through the applications' APIs, and deterministic checks inspect the final state of those systems. The result source is Artificial Analysis, and the primary metric is pass@1 (Artificial Analysis's Tasks Completed): the share of tasks completed in full, with every objective met and no guardrail violated.
- **Agents' Last Exam.** Agents' Last Exam evaluates demanding end-to-end scientific and professional tasks performed on Linux virtual machines. We use ALE v1.0 and its Linux task split, excluding five of its 105 tasks for every model: those whose input data exceeds the evaluation environment's storage limit or whose evaluation scripts did not return a score. Each task is graded by its native evaluation script, which awards partial credit between 0 and 1; the few LLM-judged checks use GPT-5.5. We run every model in a common internal agent loop; the official leaderboard instead compares agent products such as Codex and Claude Code. The primary metric is the mean task score.
- **Personal Superintelligence Index (PSI).** PSI is an internal composite of eight personal and knowledge-work evaluations: Agents' Last Exam, AutomationBench-AA, JobBench, OfficeQA Pro V2, GDP.pdf, GDP.xlsx, VBT and FocusBench.
- **JobBench.** JobBench contains 65 professional tasks across 35 white-collar occupations. Each task asks the agent to produce a practical work product and is scored against task-specific rubrics by a Grok 4.3 judge, in the OpenCode harness used on the official leaderboard. The primary metric is mean reward across tasks.
- **OfficeQA Pro V2.** OfficeQA Pro V2, from Databricks, tests grounded reasoning over a second, harder corpus: 90 questions on the U.S. Federal Accounts of Receipts and Expenditures (1,435 documents, 1793–2024), each answer graded deterministically via the official scorer with a 0% tolerance. We run it in a common internal framework: the agent works in a bash sandbox with the documents each

question cites (a median of six) staged as parsed text and a browser tool with benchmark sites blocked. The primary metric is pass@1.

- **GDP.xlsx.** GDP.xlsx, from Surge, evaluates professional reasoning inside real-world spreadsheets and CSV files across 70 economically valuable tasks in 13 domains, graded against expert rubrics. We run the 70 tasks in a common internal framework; Gemini 3.8 Flash grades each task's rubric on the final response, as in the official harness. Our agent loop and tools replace the official OpenHands loop.
- **Verifiable Browsing Task (VBT).** VBT is an internal agentic web-browsing evaluation in which models answer list-style information-seeking questions, written as realistic user requests, using browser search, open, and find tools.
- **FocusBench.** FocusBench is an internal evaluation of personal agentic tasks across a variety of domains like housing, travel, IT, and finance. It evaluates how well agents can persistently work through real-world blockers across multiple interactions and adapt to dynamically changing environments.

Coding

- **Terminal-Bench 4.0.** Terminal-Bench 4.0 evaluates 66 harder tasks completed in isolated terminal environments, with executable verifiers grading the final container state. Muse Ember's result is Artificial Analysis's run in the mini-swe-agent harness, and the primary metric is accuracy. For Claude Haiku 5.5 we use Anthropic's self-reported Claude Code result from the Claude Haiku 5.5 system card, and for GPT-6 Luna the Codex result on the Terminal-Bench leaderboard.
- **Coding Agent Index-AA.** The Artificial Analysis Coding Agent Index v2.0 combines Terminal-Bench 4.0, SWE-Atlas-QnA and FrontierCode v1.1, each run inside a coding agent: Muse Code for Muse Ember, Claude Code for Claude Haiku 5.5 and Codex for GPT-6 Luna. The result source is Artificial Analysis.
- **FrontierCode 2.0-AA.** FrontierCode 2.0, from Cognition, evaluates real engineering tasks in real repositories, each graded on whether a maintainer would merge the change. Artificial Analysis runs its private 132-task set. The result source is Artificial Analysis, and the primary metric is accuracy.

Multimodal

- **GDP.pdf.** GDP.pdf, from Surge, evaluates professional work grounded in supplied PDF source material. Artificial Analysis runs its own version: each model receives the PDF's extracted text plus page images, answers without tools, and a judge checks every rubric criterion separately. The result source is Artificial Analysis, and the primary metric is all-pass: the share of answers that meet every criterion.
- **OSWorld 2.1.** OSWorld 2.1 is the latest release of OSWorld 2.0, from XLANG Lab: 108 long-horizon computer-use workflows in which the agent operates a live Ubuntu desktop through screenshots and mouse and keyboard actions. Each task's evaluator awards partial credit on the final state of the machine, and model-based checks are judged by Claude Sonnet 4.6. We run all 108 tasks in a common internal framework rather than the official runner, with each model in its provider's computer-use configuration. The primary metric is the partial score, the mean credit per task; the strict pass rate, the share of tasks with full credit, is reported second.
- **PerceptionBench.** PerceptionBench, from Moonshot AI, tests atomic visual perception: 3,000 verified questions with short, uniquely determined answers across ten capabilities, including counting, localization, depth, OCR and perception-related hallucination. We run all 3,000 questions in a common internal framework where the model can inspect the image in a sandbox with a bash terminal and a file viewer and editor. The official judge, gpt-oss-120b, checks each answer against the reference with the official judge prompt. The primary metric is pass@1.