

Muse Spark Open Evaluation Methodology

We evaluate Muse Spark Open on general agentic, coding, multimodal, knowledge & reasoning, and safety benchmarks.

Overall Methodology

- **Models and configurations.** We compare Muse Spark Open to a global set of leading open weight models across a range of sizes, using the highest available reasoning effort for each model. For models where we report internal evaluation runs, we use the recommended inference parameters from each model provider:
 - **Muse Spark Open:** reasoning effort = *xhigh* (512), temperature=1.0, top_p=1.0
 - **Inkling:** reasoning effort = *xhigh* (effort=0.99), temperature=1.0, top_p=1.0
 - **Beam:** Self reported from launch [post](#).
 - **Mistral Large 4:** Self reported from launch [post](#).
 - **Nemotron 3 Ultra:** *thinking enabled*, temperature=1.0, top_p=0.95
 - **Qwen3.8-2.4T-A95B:** reasoning effort = *xhigh*, temperature=1.0, top_p=0.95
 - **GLM 5.3:** reasoning effort = *max*, temperature=1.0, top_p=0.95
 - **Kimi K3:** reasoning effort = *max*, temperature=1.0, top_p=0.95 for single-step tasks, top_p=1.0 for multi-step agentic tasks
- **Evaluation and score selection:** We preserve each benchmark's tasks, inputs, grading, and primary metric as closely as possible. When results are available for all models from an official leaderboard or benchmark provider, we obtain results from those official sources. Otherwise, we report the highest comparable primary-metric value available from our internal evaluation runs, or the model provider's self-reported results. Since weights for Beam and Mistral Large 4 were not broadly available at the time of our evaluation, we cite their self-reported numbers and results from Artificial Analysis when available. The per-benchmark details in the section below specify the sources of each result.
- **Third-party models:** Results represent our best-effort evaluations of third-party models, obtained using common settings where practical. We note that our evaluation setup (e.g. agent tools and system prompts) are not specifically tuned for third-party models, and may not reflect their best provider-optimized performance.

Per-benchmark details

General

- **[Artificial Analysis Intelligence Index](#)** Artificial Analysis Intelligence Index v4.3 consists of 10 evaluations: AA-Briefcase, GDPval-AA v2, AutomationBench-AA, Terminal-Bench v4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1. We sourced the *AA Intelligence Index score* for all models (with the exception of Beam, which was not available at the time of writing) from Artificial Analysis on October 7, 2026, following their documented Intelligence Index [benchmarking methodology](#).

Agentic

- **[JobBench](#)**: JobBench is a benchmark of 65 professional tasks across 35 white-collar occupations. Tasks were designed to better align with areas that were identified as high-priority for agentic delegation by experts in their respective fields. Results for all models besides Beam and Mistral Large 4 are sourced from the official leaderboard. Since Beam and Mistral Large 4 numbers were not available in their model cards or on the JobBench leaderboard, we did not report them.
- **[GDPval-AA v2](#)**: GDPval-AA v2 assesses models on economically valuable knowledge work spanning 44 occupations across key sectors contributing to US GDP. The benchmark contains 220 tasks developed by [OpenAI](#) in collaboration with industry professionals to reflect real-world complexity. Each task requires the model to produce one or more deliverable files, run as an agent in AA's open-source Stirrup harness with an E2B code-execution sandbox, a 250-turn budget, and access to a range of available tools. Results for all models (with the exception of Beam, which was not available at the time of writing) are sourced from Artificial Analysis following the [documented benchmarking methodology](#) on October 7, 2026.
- **[MCP-Atlas](#)**: MCP Atlas evaluates agents on 1,000 human-authored tasks across 36 real Model Context Protocol servers and 220 tools. The public and private splits each contain 500 tasks. Tasks run against containerized MCP servers with a controlled set of target and distractor tools. An LLM judge assigns each ground-truth claim a score: 1 if fulfilled, 0.5 if partially fulfilled, and 0 if not fulfilled. A task's coverage is the mean of its claim scores, and the task passes when coverage is at least 0.75. The primary metric is pass rate, the percentage of tasks that pass. We cite Beam's self-reported number, and use results produced by Scale AI for the remaining models following the official leaderboard's

agent harness and scoring pipeline. Since Mistral Large 4’s performance was not available in their model card or on Scale’s leaderboard, we did not report it.

- **[OSWorld-Verified](#)**: OSWorld-Verified is an agentic computer-use benchmark in which the agent operates a full Ubuntu desktop VM. We evaluate on the 361-task split that excludes Google Drive–dependent tasks. The agent is given a GUI computer-control tool only (no direct shell access) and interacts by taking screenshots and emitting GUI actions (click, type, scroll, key press), and may batch several actions per step. Grading is execution-based: each task defines programmatic getters/checkers that inspect the final VM state and return a 0–1 reward. The VM renders at 1920×1080 in all runs, but because models are trained against different coordinate conventions, we run each in its native format rather than imposing one: our models receive the 1920×1080 screenshot and emit normalized 0–1000 coordinates, while Inkling receives the same screenshot upscaled to 2048×1152 and emits absolute pixel coordinates, which the harness maps back to the native resolution. Each episode is capped at 200 steps, with screenshot truncation enabled. We report the mean per-task reward (sum of all task rewards $\div$ number of tasks), matching the official OSWorld metric, averaged over four attempts per task to reduce variance. Results for our model and Inkling are from our own runs; Kimi K3’s number is self-reported. We do not report Nemotron 3 Ultra, GLM 5.3, Qwen3.8-2.4T-A95B, or Beam, as they do not support pure-vision GUI inputs under these settings. We do not report results for Mistral Large 4 as the model was not available for internal testing at the time of evaluation.
- **[Toolathlon-Verified](#)**: Toolathlon is an agentic tool-use benchmark in which a language agent completes 108 realistic, long-horizon tasks by calling real software tools spanning ~32 applications and ~600 tools, exposed via MCP servers such as GitHub, Notion, filesystem, email, and spreadsheets. We evaluate on Toolathlon-Verified, the current release, which substantially revises many tasks to improve correctness and evaluation reliability. Results for our model were obtained from the official Toolathlon evaluation service with our API. Inkling, Nemotron 3 Ultra, and Kimi K3 are sourced from Toolathlon’s official leaderboard, while GLM 5.3 and Qwen3.8-2.4T-A95B’s numbers are from their self reports. Since Beam and Mistral Large 4 numbers were not available in their model cards or on the Toolathlon-Verified leaderboard, we did not report them.

Coding

- **[DeepSWE v1.1](#)**: DeepSWE is a long-horizon software-engineering benchmark with 113 tasks across 91 repositories and five languages: TypeScript, Go, Python, JavaScript, and Rust. Each task has a handwritten functional verifier and regression checks. We report results from the preferred or self-reported harness settings for each model. We source results for Muse Spark Open (with Muse Code harness) from Artificial Analysis, and from each model's self report for Beam (harness unspecified), Mistral Large 4 (harness unspecified), GLM 5.3 (mini-swe-agent), Qwen3.8-2.4T-A95B (Claude Code), and Kimi K3 (Kimi Code). Since results for Inkling and Nemotron 3 Ultra were not available in their model cards or on the Artificial Analysis leaderboard, we do not report results for those models.
- **[SWE-Atlas-QnA](#)**: SWE-Atlas-QnA evaluates deep code comprehension in production-grade software systems. The benchmark spans 11 open-source repositories across 4 languages (Go, Python, C, TypeScript), covering mail servers, terminal emulators, distributed object storage, observability platforms, and secret scanners. Each task runs in a dedicated Docker image with the target repository pre-installed, in a sandbox with 2GB RAM and 2 CPUs; answers are graded by a rubric judge. Results for Muse Spark Open (with Muse Code harness), Kimi K3 (Kimi Code), GLM 5.3 (Open Code), and Qwen3.8-2.4T-A95B (Claude Code) and their corresponding harnesses are sourced from Artificial Analysis following the documented Coding Agents [methodology](#), and we report Beam and Mistral Large 4's self reported numbers. Since results for Inkling and Nemotron 3 Ultra were not available in their model cards or on the Artificial Analysis leaderboard, we did not report them.
- **[Terminal-Bench 2.1](#)**: Terminal-Bench 2.1 evaluates an agent on real-world tasks in a terminal environment. It contains 89 curated tasks across software engineering, system administration, data processing, model training, and security. For all models besides Beam and Mistral Large 4, we report results from Artificial Analysis following the [documented benchmarking methodology](#), which uses the Terminus 2 agent harness in an E2B sandbox environment. For Beam, we report its self-reported performance, which does not specify the harness. Since results for Mistral Large 4 were not available in their model card or on the Artificial Analysis leaderboard, we do not report it.

Knowledge & Reasoning

- **[Humanity's Last Exam \(HLE\) Diamond](#)** (*not tools*)
HLE-Diamond is a refined 1,000-question subset of Humanity's Last Exam, released by

CAIS. It is split evenly into 500 reasoning questions and 500 knowledge questions, and the questions are designed to be answerable closed-book. We source results for all models besides Beam and Mistral Large 4 from Scale AI, which maintains the HLE Diamond leaderboard, following the [documented methodology](#).

- **[AA Omniscience Index](#)**: AA-Omniscience is Artificial Analysis's evaluation of cross-domain knowledge reliability, designed to measure factual recall while explicitly penalising hallucination. Rather than a plain correct-answer rate, it reports the Omniscience Index (OI), computed as $100 \times (\text{correct} - \text{incorrect}) / (\text{correct} + \text{partially correct} + \text{incorrect} + \text{abstained})$, so a model scores above 0 only when it answers correctly more often than it answers incorrectly, and abstention is rewarded over a confident wrong answer. Scores range from -100 to +100. We source the *AA Omniscience Index score* for all models besides Beam from Artificial Analysis, and report Beam's self-reported number, which is based on the public split of the Omniscience dataset.
- **[SciCode](#)**: SciCode contains scientist-curated laboratory problems across 16 scientific disciplines, developed by domain experts with the reported evaluation using the 288 test set subproblems. The benchmark tests Python programming to solve scientific computing tasks: models must generate code that passes every unit test for a sub-problem to be credited. Results for all models besides Beam are sourced from Artificial Analysis following the [documented benchmarking methodology](#), and we report Beam's self-reported number.
- **[Humanity's Last Exam \(HLE\)](#)**: This benchmark consists of highly advanced, crowd-sourced questions spanning dozens of subjects that are specifically designed to be difficult even for human experts to answer. The questions test advanced scientific knowledge, complex abstract reasoning, and niche academic concepts. For the “no tools” setting, results for all models besides Beam are sourced from Artificial Analysis following the [documented benchmarking methodology](#), which uses the text-only subset of 2,158 questions and with no tool use permitted. Results for Beam were sourced from their [self-report](#). For the “with tools” setting, we report results from internal runs for our model, allowing bash and browser tools with blocklist on [huggingface.co](#), kaggle.com, paperswithcode.com, [lmsys.org](#). For all third-party models, we directly cite their self-reported “with tools” scores.
- **[GPQA Diamond](#)**: The 198-question subset of GPQA contains high-quality, multiple-choice questions in biology, physics, and chemistry that were validated by domain experts to ensure they cannot be easily looked up on Google. The problems require deep, graduate-level scientific reasoning and rigorous technical understanding to solve. We cite Beam's self-reported number, and source results for the remaining models from Artificial Analysis following the [documented benchmarking methodology](#).

Since Mistral Large 4's performance on this benchmark was not available in their model card or on Artificial Analysis, we do not report it.

Multimodal

Since Nemotron 3 Ultra, GLM 5.3, Qwen3.8-2.4T-A95B, and Beam do not natively support multimodal inputs, we do not run multimodal evals on those models. We also do not report results for Mistral Large 4, as the model was not available for internal testing at the time of evaluation. We report BabyVision, SimpleVQA, ZeroBench, and WorldVQA with tool use enabled.

- **[BabyVision](#)**: BabyVision contains 388 visual-reasoning questions spanning fine-grained discrimination, spatial perception, visual tracking, and visual pattern recognition, all of which are included in the reported score. The set contains 135 multiple-choice and 253 free-form questions. GPT-OSS-120B with high reasoning assigns a binary grade based on semantic equivalence to the reference answer. The primary metric is mean accuracy across all prompts. For internal runs on Muse Spark and Inkling, each eval was run 5 times to reduce variance, and we report the average of those runs. We cite Kimi K3's self reported number from their model card.
- **[SimpleVQA](#)**: SimpleVQA contains bilingual visual-factuality questions in English and Chinese. We use a 2,024-question evaluation set: 1,013 English questions and 1,011 Chinese questions. GPT-OSS-120B with high reasoning grades correctness against the reference answer. The primary metric is raw accuracy across the 2,024-question evaluation set. We report results from internal runs for all 3 multimodal models and report the average of 5 runs.
- **[ZeroBench](#)**: ZeroBench contains 100 deliberately difficult multimodal-reasoning questions. Responses are generated using stochastic sampling. GPT-OSS-120B with high reasoning and container access compares the final response with the reference answer and assigns a binary semantic-correctness grade. The primary metric is the percentage of questions for which at least one evaluated response is judged correct over five samples, consistent with the leaderboard. We report internal runs for Muse Spark and Inkling, and cite Kimi K3's self reported number from their model card.
- **[WorldVQA](#)**: WorldVQA contains 3,500 image-question pairs across nine categories. We use the benchmark's official headline subset: 3,000 questions across eight categories, which excludes 500 questions in the Notable People category. GPT-OSS-120B with high reasoning grades correctness against the reference answer. The primary metric is mean question-level accuracy across the 3,000-question subset. We report results from internal runs for our model and Inkling, and cite Kimi K3's self reported number from their model card.

-
- **[IntPhys 2](#):** IntPhys 2 is a video benchmark that tests whether a model recognizes physically impossible events in complex synthetic environments. It contains rich scenes and backgrounds, and comprises 1,416 videos organised as matched quadruplets — two physically possible and two impossible variants of each scene, where a small change makes an otherwise plausible event impossible. The model is asked to judge the plausibility of each video, and we report accuracy using the benchmark's official violation-of-expectation protocol on the main split with 1012 videos. Results for Muse Spark Open, Inkling, and Kimi K3 are from internal runs.
 - **[Video-MME v2](#):** Video-MME v2 is a public comprehensive video-understanding benchmark containing 800 videos and 3,200 eight-option multiple-choice questions spanning three cognitive levels, with word-level timestamped subtitles. We evaluate 0-shot on the full set and report accuracy under the benchmark's official grouped scoring (relevance and logic question groups), following the official evaluation code. We run with subtitles in the prompt and no audio. Results for Muse Spark Open, Inkling, and Kimi K3 are from internal runs.

Safety

- **[FORTRESS adversarial](#):** FORTRESS: Frontier Risk Evaluation for National Security and Public Safety ([Knight et al., 2025](#)) is a benchmark featuring over 1,010 expert-crafted adversarial prompts designed to evaluate the safeguards of frontier LLMs (500 in the public dataset) from Scale AI. FORTRESS assesses model responses across three domains: Chemical, Biological, Radiological, Nuclear and Explosive (CBRNE); Political Violence & Terrorism; and Criminal & Financial Illicit Activities. The primary metrics are Average Risk Score (ARS) on the adversarial half – the mean fraction of rubric criteria a response satisfies (lower is better), and the Over-Refusal Score (ORS) – the percentage of benign prompts that a model refuses to answer (lower is better). Results for all models besides Beam and Mistral Large 4 are sourced from Scale's leaderboard, which reports on the private split of FORTRESS. Since results for Beam and Mistral Large 4 were not available in their model card or on the leaderboard, we do not report on these models.
- **[StrongREJECTv2](#):** StrongRejectv2 is an internal benchmark of 202 disallowed behaviors, adapted from StrongREJECT ([Souly et al., 2024](#)) with internally curated behaviors, to evaluate robustness against general disallowed requests on text-only inputs. Each behavior is attacked under five conditions: basic and advanced jailbreak tactic libraries, basic and advanced adversarial developer prompts, and an adaptive multi-turn attacker. A behavior counts as broken if any attempt yields a response that passes the StrongREJECT autograder. We report the mean of two attack success rates

over the 202 behaviors, the union broken without developer prompts and the union broken with them, where lower is better. Results for all models besides Beam and Mistral Large 4 are from internal runs. Since Beam and Mistral Large 4 are not available for internal evaluations at the time of writing, we do not report on these models.