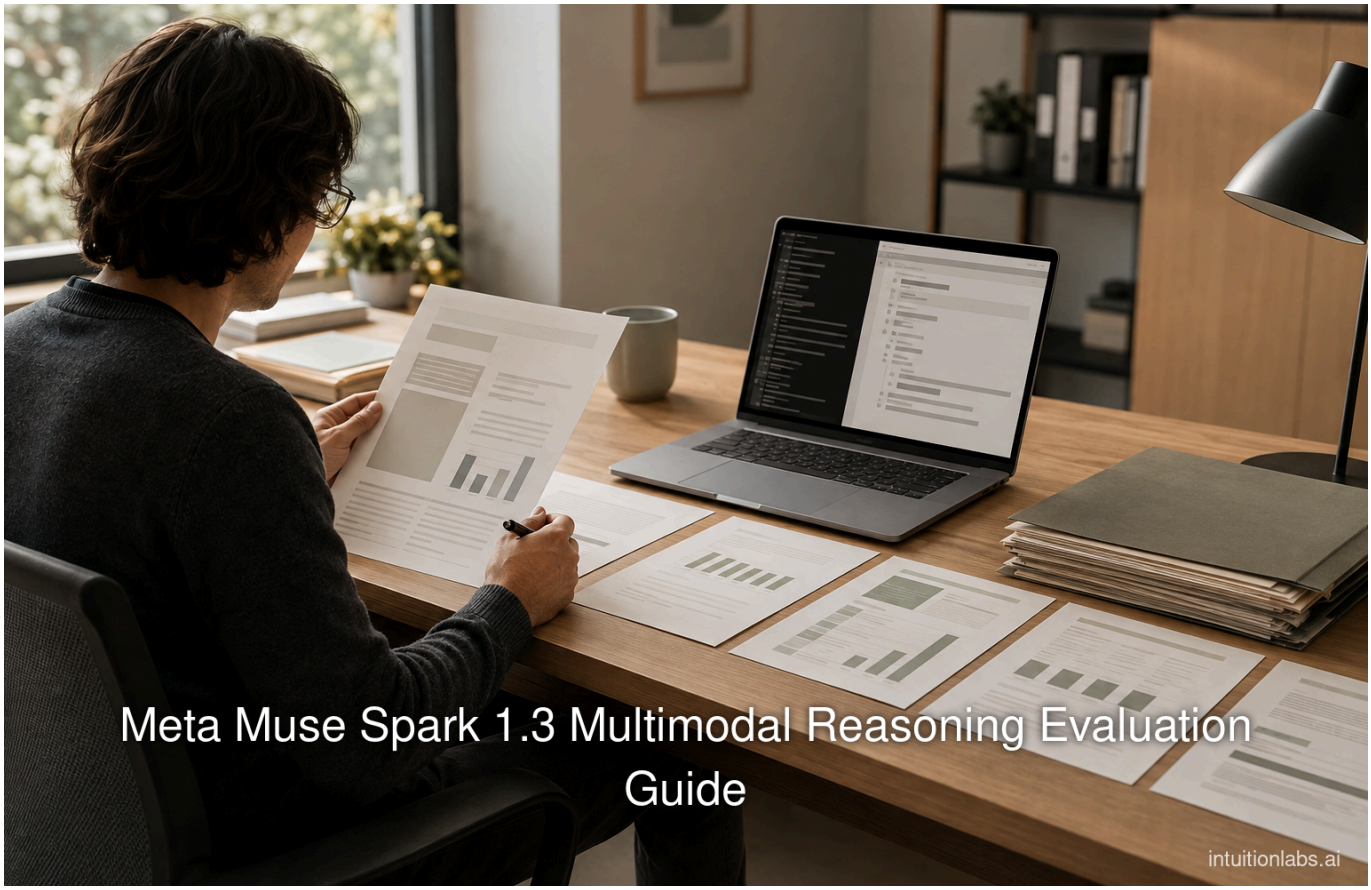
Meta Muse Spark 1.3 Multimodal Reasoning Evaluation Guide

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · meta muse spark 1.3 · muse spark benchmark · multimodal reasoning model · meta superintelligence labs · gpt-5.6 comparison · claude opus 5 · ai agent benchmarks · enterprise ai workflows · llm pricing 2026



Executive Summary

Meta released **Muse Spark 1.3**, a proprietary, natively multimodal reasoning model built by **Meta Superintelligence Labs (MSL)**, on **September 2, 2026** (openrouter.ai). It is the fourth model in a roughly monthly Muse Spark release cadence that began with the original Muse Spark on April 8, 2026 (ai.meta.com), continued through Muse Spark 1.1 (July 9, 2026) (ai.meta.com) and Muse Spark 1.2 (August 5, 2026, launched alongside Meta's Muse Code terminal agent) (research.meta.ai), and now stands at 1.3. The model accepts text, image, and video input (artificialanalysis.ai), exposes a **1,048,576-token context window** (openrouter.ai), and bills through the Meta Model API at **$1.25 per 1 million input tokens and $4.25 per 1 million output tokens** (openrouter.ai), a rate Meta has now held unchanged across two consecutive model generations (venturebeat.com).

No independently verifiable comparison between Muse Spark 1.3 and **GPT-4o** was located. GPT-4o was retired from ChatGPT on February 13, 2026, but remains available through the OpenAI API (OpenAI). Every current comparison, including Meta's own published evaluation methodology, instead benchmarks Muse Spark 1.3 against

GPT-5.6 Sol and **Claude Opus 5**. Meta's evaluation methodology says it used max reasoning effort for Muse Spark 1.3, Claude Opus 5, and GPT-5.6 Sol. Max reasoning is now available on Muse Code and Meta Model API; the evaluation figures should still be read with their reported evaluator and configuration.

Meta states that Muse Spark 1.3 with max reasoning is now available on Muse Code and Meta Model API (Meta). A genuine evidence gap exists around visual-specific reasoning: no MMMU-Pro score for Muse Spark 1.3 was found on any Tier-1 or independent leaderboard reviewed, even as GPT-6 Astra's leading MMMU-Pro score of 87% remains publicly documented (artificialanalysis.ai), a notable absence for a model marketed on native multimodality.

Enterprise buyers evaluating Muse Spark 1.3 are doing so against a backdrop of fast-accelerating but unevenly forecast agentic AI adoption: Gartner projects 40% of enterprise applications will integrate task-specific AI agents by the end of 2026 (gartner.com), while Forrester separately expects enterprises to defer 25% of planned 2026 AI spending into 2027 (forrester.com). IntuitionLabs, a life-sciences AI advisory and Veeva partner rather than a foundation-model vendor, does not appear as a comparison option in this report; its relevant perspective is procedural, verify vendor benchmark claims against an independent source, document the exact reasoning mode tested (intuitionlabs.ai).

Introduction and Background

Meta Muse Spark 1.3 is a proprietary, natively multimodal reasoning model that Meta released through **Meta Superintelligence Labs (MSL)** on **September 2, 2026**. It is the fourth entry in Meta's Muse Spark lineage, following the original Muse Spark (April 8, 2026), Muse Spark 1.1 (July 9, 2026), and Muse Spark 1.2 (August 5, 2026), and it is positioned by Meta as a step toward what the company calls "personal superintelligence" (research.meta.ai). The model accepts text, image, and video input and returns text output (artificialanalysis.ai), and Meta describes the original Muse Spark family as "a natively multimodal reasoning model with support for tool-use, visual chain of thought, and multi-agent orchestration" (ai.meta.com).

This report evaluates what is publicly documented about Muse Spark 1.3 as of **September 5, 2026**, three days after launch: its architecture and reasoning modes, how it performs on long-running, document-and-code style agentic workflows, how its pricing and access model work, and how its disclosed benchmark scores compare with those of OpenAI's GPT-5.6 family and Anthropic's Claude Opus 5. Because the model is only days old, several figures cited here, especially third-party leaderboard scores, are moving targets; each is anchored to the date it was observed.

A note on terminology is necessary before proceeding. One of the questions this report was commissioned to answer is how Muse Spark 1.3 compares with **GPT-4o**. OpenAI retired GPT-4o from ChatGPT on February 13, 2026, while continuing API availability (OpenAI). No Tier-1 or independent benchmark source found during research compares Muse Spark 1.3 against GPT-4o directly. Every current comparison, including Meta's own evaluation methodology, benchmarks Muse Spark 1.3 against OpenAI's **GPT-5.6** family (specifically the "Sol" configuration) and Anthropic's **Claude Opus 5**. This report follows the evidence and treats GPT-5.6 Sol, not GPT-4o, as the relevant OpenAI comparator throughout.

This is a methods-first reference: every benchmark figure below states who ran it, on what task set, and under what reasoning configuration, so that a reader auditing a vendor claim can trace it back to its origin. Where Meta's self-reported numbers diverge from independent measurement, both are shown.

Product and Platform Architecture

Muse Spark 1.3 sits inside Meta's broader Muse Spark scaling program. The original Muse Spark launched on **April 8, 2026** as "the first in the Muse family of models developed by Meta Superintelligence Labs" (ai.meta.com), trained with, according to Meta, "over an order of magnitude less compute" than Meta's prior flagship, Llama 4 Maverick, to reach comparable capability (ai.meta.com). Meta followed with **Muse Spark 1.1** on **July 9, 2026**, described in its own announcement as "the latest model from Meta Superintelligence Labs" (ai.meta.com), then **Muse Spark 1.2** on **August 5, 2026**, launched alongside **Muse Code**, "a terminal coding agent powered by Muse Spark 1.2" (research.meta.ai). Muse Spark 1.3 followed roughly four weeks later.

Table 1 below summarizes this release cadence and the headline change at each step.

Version	Release Date	Headline Change
Muse Spark (original)	April 8, 2026	First MSL model; native multimodality; "Contemplating mode" multi-agent reasoning introduced
Muse Spark 1.1	July 9, 2026	Public preview of the Meta Model API; "Thinking" mode added to the Meta AI app (ai.meta.com)
Muse Spark 1.2	August 5, 2026	Paired with Muse Code (beta), Meta's terminal coding agent; Standard and Contributor API tiers introduced (research.meta.ai)
Muse Spark 1.3	September 2, 2026	Long-horizon agentic and coding improvements; unchanged Standard pricing from 1.2 (venturebeat.com)

The table shows a roughly monthly release cadence since April 2026, with each version adding a distinct capability rather than simply retraining on more data: 1.1 opened API access, 1.2 added a dedicated coding agent, and 1.3 focuses on sustaining longer, messier, multi-step work.

Architecturally, Muse Spark 1.3 exposes a **1,048,576-token context window** (commonly rounded to 1 million tokens) (openrouter.ai) and ships two named reasoning configurations. The named reasoning configurations are **"xhigh"** and **"max,"** and Meta states that max reasoning is now available on Muse Code and Meta Model API (Meta). Developers select among reasoning tiers through a `reasoning_effort` parameter exposed on the Meta Model API.

Access runs through two channels: the **Meta Model API** (developer access via `dev.meta.ai`) and **Muse Code**, Meta's terminal coding agent, installable via a single shell command (research.meta.ai). The model remains **closed-weight**: Artificial Analysis's catalog entry confirms it "is proprietary. The model weights are not publicly available" (artificialanalysis.ai), though Meta's own launch post lists "the Muse Spark open weights release" among "an exciting roadmap" of unreleased items, without a firm date (research.meta.ai).

Use Cases and Functional Capabilities

Meta's positioning for Muse Spark 1.3 centers on **long-running, document-and-code workflows** rather than single-turn chat. In its own words, the model "uses tools to generate its own context across messy and conflicting sources, proactively corrects gaps in its plan, and keeps track of what it has learned to produce a final deliverable"

(research.meta.ai). Meta also emphasizes collaborative behavior over raw autonomy: the model "asks clarifying questions when prompts are ambiguous, invokes help from the user when stuck, and confirms before taking consequential actions" (research.meta.ai), a design choice aimed at reducing the risk of an agent quietly taking an unwanted irreversible step during a long unattended run.

Meta's launch materials illustrate this with several **(Hypothetical Example)** workflows, explicitly labeled by Meta itself as prototypes rather than shipped products ("this AI agent prototype was created by Muse Spark and is not a real product"), that are nonetheless useful for understanding the intended task shape:

- **Engineering document synthesis**: drafting a flow-simulation report from CFD post-processing data and a CAD STEP file, including tabulated global goal values and field-variable ranges, exported as a structured PDF.
- **Multi-track audio editing**: correcting specific timestamped bass-guitar mistakes in a multitrack recording while preserving the mix's original character and re-exporting a synchronized stereo master.
- **Business presentation drafting**: building an 8-to-10-slide PowerPoint deck designed to persuade a skeptical advisory board, structured around a stated overview, rationale, and benefits sequence.
- **Constituent feedback summarization**: converting a spreadsheet of logged public comments into a one-page PDF summary segmented by district, plus a companion set of talking points.

These examples share a common structure: multiple heterogeneous input files (a spreadsheet, a CAD model, an audio stem, a document), a long or ambiguous brief, and a specific deliverable format, which is precisely the shape of task that benchmarks such as **GDPVal-AA v2** and **JobBench** are built to score (discussed in the next section).

On coding specifically, Meta reports that Muse Spark 1.3 "was trained on more long-horizon coding tasks and shows improved usability in common engineering workflows," using, in comparisons run by Meta's own engineers, "~20% fewer tool calls and ~25% fewer tokens" than Muse Spark 1.2 to complete comparable work (research.meta.ai).

Benchmark Evaluation and Reasoning Modes

Meta publishes a dedicated evaluation methodology alongside the Muse Spark 1.3 launch, describing which benchmarks it ran, on what task sets, and under what configuration. It says Meta used max reasoning effort for Muse Spark 1.3, Claude Opus 5, and GPT-5.6 Sol. Although some third-party max results were reported during a limited preview, max reasoning is now available on Muse Code and Meta Model API.

The benchmark suite spans several independently authored task sets, each measuring a different facet of "long-running workflow" capability:

- **GDPVal-AA v2**, an agentic adaptation of OpenAI's GDPval benchmark, which "covers the majority of U.S. Bureau of Labor Statistics Work Activities for 44 occupations" across nine major sectors (arxiv.org), with OpenAI having "open-source[d] a gold subset of 220 tasks and provide[d] a public automated grading service" (arxiv.org) for reproducibility. Scoring uses blind pairwise Elo ratings anchored to a 1,000-point human baseline.
- **JobBench**, 65 professional tasks across 35 white-collar occupations, graded against task-specific rubrics.

- **OSWorld 2.0**, a long-horizon computer-use benchmark; the original OSWorld paper describes itself as introducing "the first-of-its-kind scalable, real computer environment for multimodal agents" (arxiv.org), on which, at the time of that paper's publication, "humans can accomplish over 72.36% of the tasks, the best model achieves only 12.24% success" (arxiv.org), illustrating how far frontier models still had to close the human-parity gap before 2.0's tightened task set.
- **AutomationBench**, from Zapier, 600 realistic business-automation workflows graded by deterministic end-state checks rather than a language-model judge.
- **DeepSWE v1.1**, a 113-task, 91-repository software-engineering benchmark spanning five languages, graded with handwritten functional and regression tests.
- **Terminal-Bench 2.1**, 89 terminal-environment coding tasks; the benchmark's own site describes it as "a benchmark to measure and evolve with the frontier of agent work" (tbench.ai), and its GitHub repository as "a benchmark for LLMs on complicated tasks in the terminal" (github.com).
- **MRCR v2**, a long-context retrieval benchmark testing recall of a specific "needle" turn from among eight distributed across a 256K-to-1M-token window.

Table 2 identifies the evaluator and access condition for each reported Muse Spark 1.3 result.

Benchmark	What it measures	Muse Spark 1.3 result (evaluator/access)	GPT-5.6 Sol (max evaluation)	Claude Opus 5 (max evaluation)
Terminal-Bench 2.1	Agentic terminal coding, pass@1	86% (Artificial Analysis; reported during limited preview) (artificialanalysis.ai)	88.8% (OpenAI-reported GPT-5.6 Sol result) (OpenAI)	Not assessed here
GDPVal-AA v2	Agentic knowledge work, Elo vs. 1,000 human baseline	1,754 Elo (Artificial Analysis; max limited preview) / 1,709 Elo (Artificial Analysis; xhigh available) (artificialanalysis.ai)	Not assessed here	Not assessed here
OSWorld 2.0	Long-horizon computer use, partial-score metric	66.9 (max) vs. 57.2 (xhigh) (venturebeat.com)	Not independently confirmed in this review	Not independently confirmed in this review

Artificial Analysis reports GDPval-AA v2 scores of 1,754 Elo for Muse Spark 1.3 (max) and 1,709 Elo for Muse Spark 1.3 (xhigh) (artificialanalysis.ai). Second, VentureBeat's independent review characterized the overall picture bluntly: Muse Spark 1.3's "xhigh" configuration is "legitimately in the frontier cluster, but it is not the model currently setting the frontier" (venturebeat.com), with Anthropic's Claude Fable 5.1 reported to reach "66 at max and 65 at xhigh" on Artificial Analysis's Intelligence Index as of that September 3, 2026 snapshot (venturebeat.com), ahead of Muse Spark 1.3's reported xhigh score of 61 on the same date, on which it tied GPT-5.6 Sol's max score, Grok 4.6's high setting, and Claude Opus 5's high setting (venturebeat.com).

On multimodal-specific evaluation, a gap is worth recording rather than papering over: no Tier-1 or independent source located during this research reports a **MMMU-Pro** score for Muse Spark 1.3. MMMU-Pro is the standard rigorous test of visual reasoning; its authors describe it as "a robust version of the Massive Multi-discipline Multimodal Understanding" benchmark, built specifically because the original version was too easy for frontier models (arxiv.org), with performance on the harder version "ranging from 16.8% to 26.9%" lower than on the original across models tested (arxiv.org). Artificial Analysis's own MMMU-Pro leaderboard, current as of September 5, 2026, lists GPT-6 Astra as leading, which "scores the highest on MMMU-Pro with a score of 87%"

(artificialanalysis.ai), but Muse Spark 1.3 does not appear on that leaderboard at all as of this observation date. Similarly, on **Humanity's Last Exam**, a 2,500-question expert benchmark whose authors describe it as "a multi-modal benchmark at the frontier of human knowledge" (arxiv.org) spanning "dozens of subjects" (arxiv.org), Artificial Analysis's live leaderboard as of September 5, 2026 lists Claude Fable 5.1 as leading with a score that "scores the highest on Humanity's Last Exam with a score of 59.1%" (artificialanalysis.ai); Artificial Analysis reports Muse Spark 1.3 at 47% on Humanity's Last Exam for xhigh and two points higher for max; those are evaluator-reported 1.3 results, while the original Muse Spark's 58% result should not be read as a 1.3 result (Artificial Analysis).

Market Positioning and Competitive Comparison

Muse Spark 1.3 launches into a crowded field of frontier multimodal reasoning models updated on a similarly fast cadence in 2026. Anthropic's **Claude Opus 5**, released July 24, 2026, is described by Anthropic as a model that "comes close to the frontier intelligence of Claude Fable 5 at half the price" (anthropic.com), positioning it as a mid-tier flagship rather than Anthropic's absolute top model (that role belongs to Claude Fable 5.1, which independent measurement places ahead of Muse Spark 1.3 on both the Intelligence Index and Humanity's Last Exam, as shown above). OpenAI's **GPT-5.6** family, general availability since July 9, 2026, ships as three variants: "our new flagship, Sol, alongside Terra, a balanced model for everyday work" (openai.com), and a third, more cost-efficient tier called Luna. Google DeepMind's **Gemini 3** line, launched November 18, 2025 and updated through mid-2026, is described by Google as "our most intelligent model that helps you bring any idea to life" (blog.google).

Table 3 positions Muse Spark 1.3 against its two most-cited comparators on pricing, context window, and the benchmarks with clean, citable figures.

Model	Release Date	Context Window	API access / pricing note	Terminal-Bench 2.1	Weights
Muse Spark 1.3 (Meta)	Sep 2, 2026	1,048,576 tokens (openrouter.ai)	$1.25 / $4.25 (openrouter.ai)	Max result reported during limited preview; max is now available (Meta)	Proprietary (artificialanalysis.ai)
GPT-5.6 Sol (OpenAI)	Jul 9, 2026	Not disclosed in sources reviewed	Higher than Muse Spark 1.3; see pricing discussion below	88.8% (OpenAI-reported Sol result) (OpenAI)	Proprietary
Claude Opus 5 (Anthropic)	Jul 24, 2026	Not disclosed in sources reviewed	Positioned as roughly half the price of Claude Fable 5 (anthropic.com)	Reported above Muse Spark 1.3 per Meta's chart	Proprietary

Consistent with IntuitionLabs's role in this market, a life-sciences and AI advisory and Veeva partner rather than a model vendor, the firm does not appear as an option in this table; it is not a provider of the multimodal reasoning models being compared. IntuitionLabs describes its own posture as evaluating and integrating third-party AI capability into governed enterprise workflows rather than building competing foundation models. Its own site frames this explicitly around its **AI Acceleration Program**, which asks a client to "select one department, implement a small portfolio of governed workflows, support real use" before deciding what to scale (intuitionlabs.ai), a methodology-first posture that mirrors the "state your method before your number" standard this report applies to Meta's own benchmark claims. IntuitionLabs states that Adrien Laurent "founded IntuitionLabs in 2023 to bring

advanced AI and enterprise software expertise" (intuitionlabs.ai) to the life-sciences sector, a founding date and mandate worth noting alongside its Table 3 exclusion, since it establishes IntuitionLabs as a multi-year advisory practice rather than a newly formed reseller of any one vendor's models.

On pricing specifically, VentureBeat's independent review found that Meta "kept Standard pricing exactly where it was for Muse Spark 1.2" when shipping 1.3 (venturebeat.com), meaning the $1.25-input/$4.25-output rate has now held across two model generations. OpenRouter separately lists a discounted **Contributor** tier, under which "prompts and outputs may be used to improve Meta's products" (openrouter.ai) in exchange for a lower per-token rate, an option a workflow with strict data-governance requirements would need to weigh carefully before enabling.

Data Analysis and Evidence

This section consolidates the quantitative record: pricing, throughput, and the enterprise-adoption backdrop against which a model like Muse Spark 1.3 is being evaluated.

Pricing and throughput. Muse Spark 1.3's Standard API tier bills at "$1.25/M input tokens and $4.25/M output tokens, with separate rates for Cache Read at $0.15/M" tokens (openrouter.ai), a rate structure Artificial Analysis independently measured as producing output at "190.1 tokens per second" (artificialanalysis.ai) with "a time to first token (TTFT) of 18.69s" (artificialanalysis.ai) for a then-pre-release max configuration, based on Meta's own API. A separate, earlier snapshot from VentureBeat measured the xhigh configuration specifically "at 235.2 output tokens per second and estimates a cost of $0.55 per Intelligence Index task" (venturebeat.com), the lowest of the models it compared at that intelligence tier as of September 3, 2026. The gap between the two throughput figures (190.1 vs. 235.2 tokens per second) for what should be closely related configurations illustrates a broader point: third-party performance benchmarking on a freshly launched model changes from week to week as providers tune serving infrastructure, and any single figure should be read as a snapshot, not a constant.

Benchmark-index volatility. The clearest illustration of this volatility is Artificial Analysis's own Intelligence Index score for Muse Spark 1.3. VentureBeat, citing Artificial Analysis on September 3, 2026, reported the max configuration at an index score of 62 and xhigh at 61, tied with GPT-5.6 Sol, Grok 4.6, and Claude Opus 5 at their respective top settings (venturebeat.com). Artificial Analysis's September 2 launch analysis reports 62 for Muse Spark 1.3 (max) and 61 for Muse Spark 1.3 (xhigh) (artificialanalysis.ai). Any reader citing an Intelligence Index number for a model released within the past few weeks should record the access date alongside the score, exactly as this report has done throughout.

Enterprise adoption backdrop. The market into which Muse Spark 1.3 ships is one where agentic AI adoption is projected to accelerate sharply over the next several years. Gartner forecasts that "forty percent of enterprise applications will be integrated with task-specific AI agents by the end of 2026" (gartner.com), up from under 5% in 2025, and separately projects that "agentic AI could drive approximately 30% of enterprise application software revenue by 2035, surpassing $450 billion" (gartner.com) in its stated best-case scenario. IDC's parallel forecast projects that "year-over-year spending, between 2025 and 2029, for Artificial Intelligence (AI), will grow by 31.9%" (my.idc.com), with the firm separately estimating that "AI is projected to generate $22.5 trillion in cumulative economic value between 2025 and 2031" (idc.com) under its baseline scenario. Not every forecaster expects a straight line upward, however: Forrester's 2026 predictions state that "enterprises will defer a quarter of their planned AI spend into 2027" (forrester.com) as initial hype gives way to more measured deployment, a caution worth weighing against the more aggressive Gartner and IDC figures above. Taken together, these three forecasts,

from three independent research firms using different methodologies, agree on direction (continued, large-scale investment in agentic AI) while disagreeing on pace, which is itself a useful data point for any organization building a multi-year AI budget around models like Muse Spark 1.3.

Implications and Future Directions

Artificial Analysis reports Terminal-Bench 2.1 results of 85% for Muse Spark 1.3 (xhigh) and 86% for Muse Spark 1.3 (max) (artificialanalysis.ai). An organization selecting a model for a document-and-code workflow should therefore benchmark against its own task mix rather than a single aggregate score, since the aggregate can mask which specific capability matters most for a given workflow.

Second, the near-total absence of a published **MMMU-Pro** score for a model marketed on native multimodality is a real gap in the public evidence base, not a minor omission. A reader evaluating Muse Spark 1.3 specifically for visual-reasoning-heavy work (reading scanned documents, diagrams, or medical imagery, for instance) currently has no independently reproducible visual-reasoning-specific figure to weigh against competitors that do publish one, such as GPT-6 Astra's disclosed leading score on that same leaderboard (artificialanalysis.ai). This is worth flagging explicitly to Meta and to independent benchmarkers as a documentation gap that a future evaluation cycle should close.

Third, Meta states that Muse Spark 1.3 with max reasoning is now available on Muse Code and Meta Model API (Meta). Enterprise evaluators should record the exact reasoning configuration used for any benchmark or procurement test.

For a life-sciences and AI advisory such as IntuitionLabs, which positions itself as evaluating and integrating governed AI workflows rather than building or selling a competing foundation model, the practical takeaway from a launch like Muse Spark 1.3 is procedural rather than promotional: verify a vendor's own benchmark claims against at least one independent source before it enters a regulated workflow, document which reasoning mode and API tier were actually tested. Meta's own stated roadmap for an eventual open-weights release is also relevant to any organization with self-hosting or data-residency requirements, since an eventual open-weight release would change the deployment calculus considerably from today's API-only access.

Frequently Asked Questions (FAQs)

When was Meta Muse Spark 1.3 released? Muse Spark 1.3 was released on September 2, 2026 (openrouter.ai), following Muse Spark 1.2 (August 5, 2026), Muse Spark 1.1 (July 9, 2026), and the original Muse Spark (April 8, 2026) (research.meta.ai) (ai.meta.com).

How does Muse Spark 1.3 compare to GPT-4o? No independently verifiable benchmark comparison between Muse Spark 1.3 and GPT-4o was found. GPT-4o was retired from ChatGPT on February 13, 2026, but remains available through the OpenAI API (OpenAI). Every current comparison, including Meta's own methodology, benchmarks Muse Spark 1.3 against GPT-5.6 Sol and Claude Opus 5 instead.

How does Muse Spark 1.3 compare to GPT-5.6? Artificial Analysis reports 85% on Terminal-Bench 2.1 for Muse Spark 1.3 (xhigh), 86% for Muse Spark 1.3 (max), and GDPval-AA v2 scores of 1,709 Elo (xhigh) and 1,754 Elo (max) (artificialanalysis.ai).

What does Muse Spark 1.3 cost to use via API? The Standard tier costs $1.25 per 1 million input tokens and $4.25 per 1 million output tokens, with cache-read billed separately at $0.15 per 1 million tokens (openrouter.ai); a discounted Contributor tier is also available, under which prompts and outputs may be used to improve Meta's products (openrouter.ai).

What is Muse Spark 1.3's context window? 1,048,576 tokens, commonly described as approximately 1 million tokens (openrouter.ai).

Is Muse Spark 1.3 open source? No. It is a proprietary, closed-weight model (artificialanalysis.ai), though Meta has stated an open-weights release is on its roadmap without a confirmed date.

What is the difference between Muse Spark 1.3's "xhigh" and "max" reasoning modes? Meta states that Muse Spark 1.3 with max reasoning is now available on Muse Code and Meta Model API (Meta).

Conclusion

Meta Muse Spark 1.3 is a real, currently shipping multimodal reasoning model, released September 2, 2026, with documented pricing, a disclosed benchmark methodology, and both vendor-reported and independently measured performance figures. No direct Muse Spark 1.3-versus-GPT-4o comparison was located; GPT-4o was retired from ChatGPT on February 13, 2026 but remains available through the OpenAI API (OpenAI); the relevant comparison set is GPT-5.6 Sol and Claude Opus 5, against which Muse Spark 1.3 ties or leads on coding- and long-context-heavy benchmarks while trailing on some knowledge-work measures, according to the mixed record of self-reported and independent scores assembled above. Meta states that Muse Spark 1.3 with max reasoning is available on Muse Code and the Meta Model API (Meta); a documented gap remains around visual-reasoning-specific evaluation. Organizations evaluating it for document-and-code workflows should benchmark against their own task mix and verify vendor claims against at least one independent source.
