Meta: Will Muse Cause a Spark?

Muse Spark 1.3 puts Meta back on the frontier watchlist. It is not yet a revenue story.

CONSTRUCTIVE ON CAPABILITY. NEUTRAL ON MONETISATION.

Four months ago we put Meta's Llama-4 in the dead lane of our model map: slow and unreliable when problems got hard, with no reason to deploy it over cheaper or better alternatives. Muse Spark 1.3, released this month, changes that verdict. Across 31 rounds and 1650 graded answers on AMC 8 competition math, Muse scored 99.0% on the questions it answered, level with Claude Opus 5 and GPT-5.6 Terra Pro and inside a point of Gemini 3.8 Flash, and it was perfect on the two hardest tiers. It was the second-fastest model in the field and its metered bill, 0.27 cents a question, was a third of Opus 5's and a fifth of Terra Pro's. What it has not yet shown is consistency: its time per batch swung from 35 to 170 seconds on the hardest problems and one hard batch blew a three-minute deadline, which is exactly the trait that keeps a model out of the interactive products where the margin sits. We read Muse as proof that Meta's AI capex is producing a frontier-grade model and as cost-of-goods relief for Meta's own surfaces, not as a new revenue line. That is a start, and it is worth taking seriously.

1. **Meta now builds frontier-grade models.**
   In May, on the same problem pool, Llama-4 Maverick scored 84.2% overall and 65.5% on the hardest tier. Muse scores 99.0% and 100.0%. On the like-for-like view the entire field now sits within about three points and Muse is inside that band at every difficulty tier, with every hard and stretch question answered correctly. The "Meta cannot build at the frontier" pillar of the bear case does not survive this data.

2. **The gap to Anthropic is efficiency, not intelligence.**
   Think of a smart school math kid and an olympiad winner. Both get the answer. The olympiad winner has muscle memory: Opus 5 used 326 output tokens a question to Muse's 612, took 37 seconds a batch to Muse's 52, barely slowed as problems got harder and never ran past about a minute. Muse works everything from first principles, with 66% of its output spent thinking, and its time nearly triples from easy to stretch. Speed and consistency on hard work remain Anthropic's moat.

3. **This is a cost story for Meta, not a revenue story.**
   A model that is cheap, accurate and variable in latency is the profile you run inside your own free products at scale, not one you sell by the token to enterprises paying for predictability. Muse lowers what Meta AI, WhatsApp and Instagram cost to serve and removes Meta's dependence on rivals' models. It does not, on this evidence, earn Meta a seat in the API market. The next release, and whether the latency tail narrows, decides which of those two stories dominates.

KEY DATA

Models	6, via OpenRouter
Rounds / graded answers	31 / 1,650
Metered spend	$10.02
Muse accuracy, excl. failed	99.0% (#3 of 6)
Muse accuracy, overall	95.8% (#4)
Muse hard / stretch accuracy	100% / 100%
Muse speed, s per batch	52 (#2; Opus 37)
Muse batch range, stretch	35 to 170 s
Muse cost per question	0.27¢ (Opus 0.88¢, GPT 1.48¢)
Muse reasoning share	66% (Opus 39%)
Muse list price, in / out	$1.25 / $4.25 per M tokens

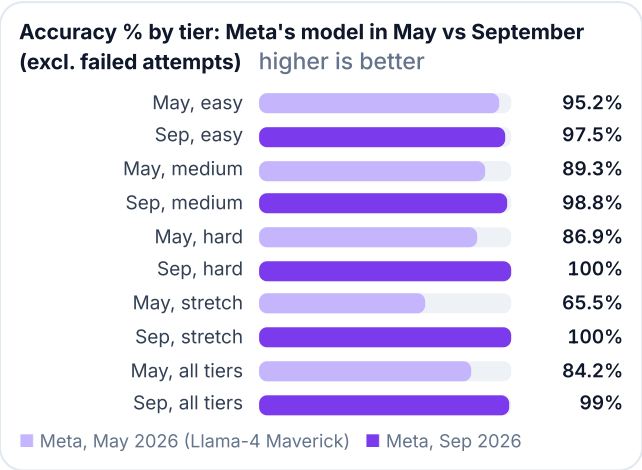
Muse Spark 1.3 (Meta), Claude Opus 5 (Anthropic), GPT-5.6 Terra Pro (OpenAI), Gemini 3.8 Flash (Google), GLM latest (Z.ai), Qwen3.8-27B (Alibaba). Provider-default settings. Costs as metered by OpenRouter. Two Opus batches lost to a credit lapse on our account are excluded.

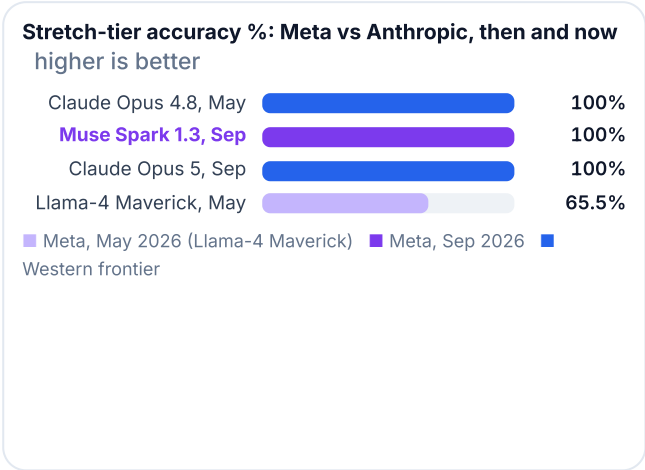
WHAT WE DID.

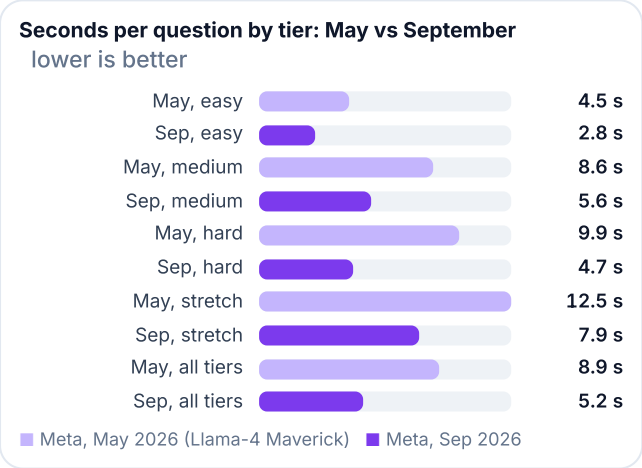
We ran six current models through our own benchmark: batches of ten random text-only AMC 8 problems from a single difficulty tier, sent to every model in one call at provider-default settings, with a 180-second deadline and only the final answer graded. 31 rounds, 1650 graded model-questions, $10.02 of metered API spend, built and run in one afternoon by one analyst with an AI coding agent. Full data, every model's working and the per-question results are at genaware.ai/bv/report3.

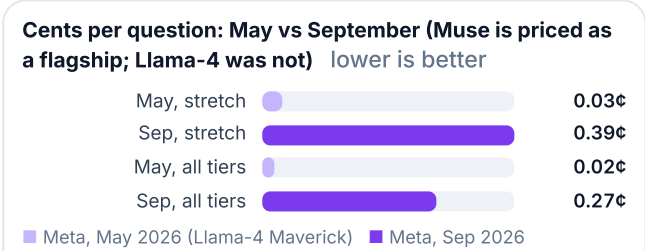
Four months on: Meta in May vs Meta in September

Same problem pool and tiers. May figures are Llama-4 Maverick from our 31 May 2026 study (12-question batches, 15-minute deadline), so accuracy is directly comparable and timing and cost are indicative. Light purple is May, dark purple is now. The hardest-tier jump, 65.5% to 100.0%, is the single most important number in this report; the cost chart is the honest counterweight: Meta traded a cheap, weak model for a frontier-priced, frontier-grade one.



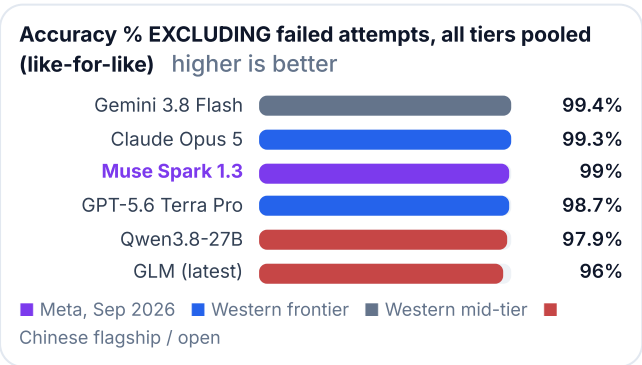


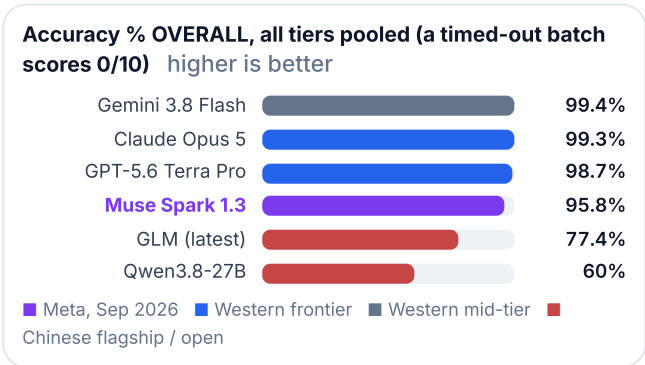




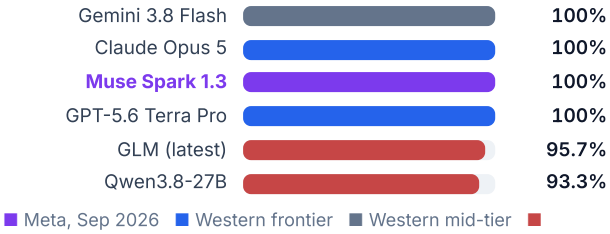
The evidence

Muse is shown in purple throughout. The left-hand accuracy chart drops failed attempts (the like-for-like view); the right-hand one scores a timed-out batch as zero (what a buyer under a time budget experiences).



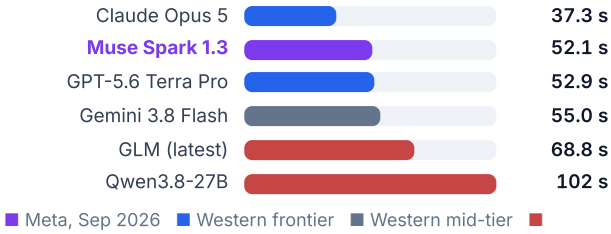


STRETCH tier - accuracy % excluding failed attempts
higher is better



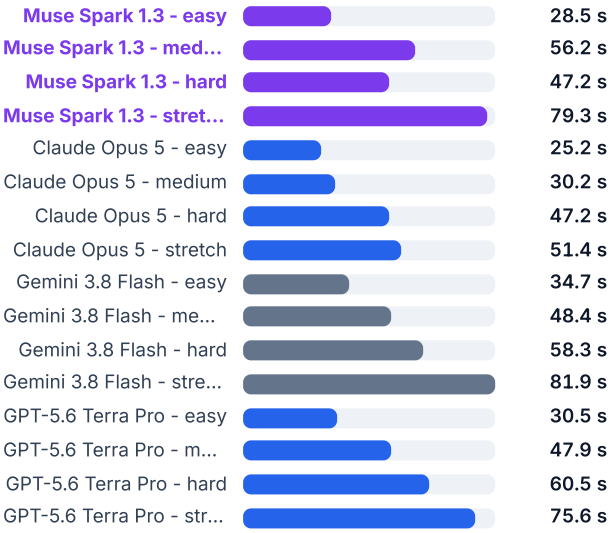
Meta, Sep 2026 · Western frontier · Western mid-tier · Chinese flagship / open

Seconds per 10-question batch, all tiers pooled
lower is better



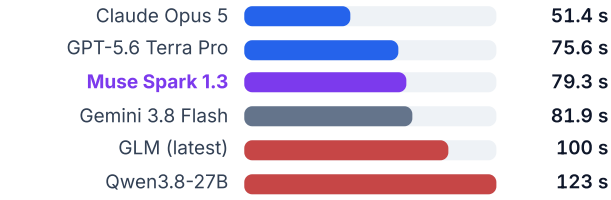
Meta, Sep 2026 · Western frontier · Western mid-tier · Chinese flagship / open

Seconds per batch by tier - Muse vs the fast models (how steeply time climbs with difficulty) lower is better



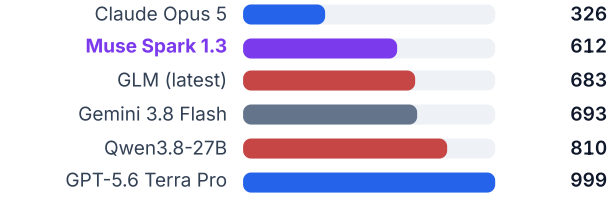
Meta, Sep 2026 · Western frontier · Western mid-tier

STRETCH tier - seconds per 10-question batch
lower is better



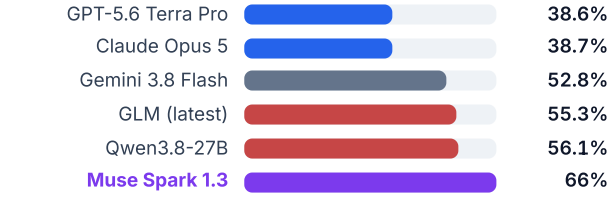
Meta, Sep 2026 · Western frontier · Western mid-tier · Chinese flagship / open

Output tokens per question, all tiers pooled (incl. hidden reasoning) lower is better



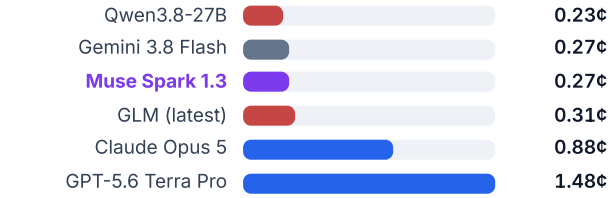
Meta, Sep 2026 · Western frontier · Western mid-tier · Chinese flagship / open

Reasoning share of output %, all tiers pooled
lower is better



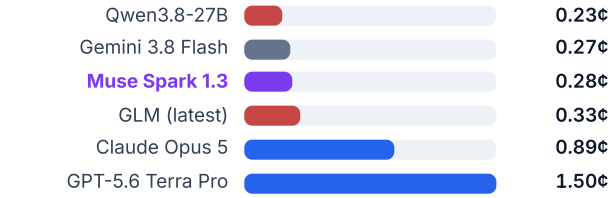
Meta, Sep 2026 · Western frontier · Western mid-tier · Chinese flagship / open

Cents per question, all tiers pooled (metered)
lower is better



Meta, Sep 2026 · Western frontier · Western mid-tier · Chinese flagship / open

Cents per CORRECT answer, all tiers pooled
lower is better



Meta, Sep 2026 · Western frontier · Western mid-tier · Chinese flagship / open

Scorecard

	MUSE SPARK 1.3	VS. FIELD	READ
Accuracy, excl. failed	**99.0%**	#3 of 6; leader 99.4%	Parity. Perfect on hard and stretch tiers.

	MUSE SPARK 1.3	VS. FIELD	READ
Accuracy, overall (timeouts = 0)	**95.8%**	#4; Opus 99.3%, GPT 98.7%	One blown deadline in 31 batches costs it 3 points.
Speed, s per 10-question batch	**52**	#2; Opus 37, GPT 53, Gemini 55	Fast on average…
Speed range on hardest tier	**35 to 170 s**	Opus never exceeded about a minute	…but wildly variable. This is the weakness.
Cost, cents per question	**0.27¢**	#3; Opus 0.88¢ (3.3x), GPT 1.48¢ (5.5x)	Budget-tier bill at flagship list price.
Reasoning share of output	**66%**	Highest in field; Opus 39%	Buys accuracy with compute. Works; caps price cuts.

Attempt accounting

MODEL	BATCHES ATTEMPTED	ANSWERED	TIMED OUT (180 S)	CREDIT/POLICY ERRORS	ANSWERED % OF VALID ATTEMPTS	ACC % EXCL. FAILED	ACC % OVERALL
Gemini 3.8 Flash	31	31	0	0	100	99.4	99.4
Claude Opus 5	31	29	0	2	100	99.3	99.3
Muse Spark 1.3	31	30	1	0	97	99.0	95.8
GPT-5.6 Terra Pro	31	31	0	0	100	98.7	98.7
Qwen3.8-27B	31	19	12	0	61	97.9	60.0
GLM (latest)	31	25	6	0	81	96.0	77.4

Timeouts are recorded when a single 10-question call exceeds 180 s. They may reflect provider/OpenRouter routing conditions rather than the model, so the excluding-failed view is the like-for-like comparison; the overall view is the buyer's experience under a time budget.

Views, risks and method

What would change our view

- *Upgrade to "threat to the frontier":* next release halves the hard-tier latency spread and holds accuracy with a lower reasoning share.
- *Downgrade to "ignore again":* the contributor tier (10x cheaper, data-for-price) is where the volume goes and it underperforms, or Muse fails on coding/agentic tests where this study is silent.

Risks to the view

One domain (competition math, text only), a 99% ceiling that cannot separate the leaders, small per-tier samples (one wrong answer moves a tier 10 points), provider routing inside every timing. This is a first look, not a coverage initiation.

How the study was run

Who	One person directing an AI coding agent (Claude Code). No engineering team, no data-science team, no vendor access.
Time	Start to published report: about two hours. The model runs themselves took 78 minutes.

Cost	$10.02 of metered OpenRouter spend across all six models, plus a $20 top-up mid-run when the account hit zero.
Questions	Our own curated store of about 700 text-only AMC 8 / AJHSME problems with verified answer keys, built earlier for a kids' math site.
Infrastructure	A $20-a-month Vultr box we already had. The GenAware Model Lab that ran our May and June papers was switched back on, its lab app patched (metered cost, reasoning tokens, a 3-minute deadline, provider defaults), and pointed at the new model list.
Runs	31 rounds, 1650 graded model-questions, every model seeing the identical ten questions per round, all metrics captured live and published to this page while the run was still going.
Reproducibility	Every batch, every model's full working, and every metered cost sits in the raw CSV and the live session pages linked at the top. Change one line to test a different model.

GenAware Model Lab, 03 Sep 2026. Independent benchmark; no vendor involvement or funding. Not investment advice. Full data edition, raw per-question results and every model's working: genaware.ai/bv/report3.