

Meta's Muse Wonder claims frontier status as early benchmarks disagree, and RSI gets its first serious yardsticks

VC Tech Radar

2026-10-13

Meta's Muse Wonder claims frontier status as early benchmarks disagree, and RSI gets its first serious yardsticks

By VC Tech Radar • October 13, 2026

Meta's new frontier model divides early evaluators. RSI measurement is getting more rigorous, and investors argue over whether profits or pro rata matter more as Series A/B gets harder.

Meta is back at the frontier, depending on whose evals you trust

Alexandr Wang launched **Muse Wonder** as Meta's "largest and most capable model yet." He called it state of the art for personal and professional agents [1]. In an interview, Meta said it focused on memory and on the model's ability to keep working through complicated tasks [2]. It said the lab was rebuilt after it was "not happy with the trajectory" of Llama 4 [2]. It also expects the time between successively larger model classes to "compress" [2]. The compute behind this is large: the Ohio cluster is scaling past 1 GW, and the Louisiana cluster is planned to reach 5 GW over the next couple of years [2].

Early outside views disagree: - **Positive:** Aravind Srinivas says Meta's model compares favorably to Fable 5.1 on WANDR evals [3]. Gavin Baker says Muse and Meta are now "significantly ahead" of Gemini on capability [4]. - **Skeptical:** Bindu Reddy first bet the model was "benchmaxxed" [5]. She later said it is "nowhere near Fable or Astra level," and Abacus plans to publish its own benchmark [6]. - **Usage:** Muse's D7 retention is above 30%, higher than ChatGPT's. Its D14 retention of 22% is just below ChatGPT's and well under social apps such as WhatsApp (67%) [7].

Why it matters: a fourth well-funded frontier lab competing for personal agents puts pressure on agent-app startups that rely on any one model. Until independent benchmarks come in, capability claims are contested.

Recursive self-improvement gets yardsticks and a stricter definition

Three items point to RSI moving from slogan to something people can measure: - **TasteVal** (P-Zero Research) tests experimental research taste against a baseline from 24 experts who recently worked at labs such as OpenAI and DeepMind [8]. It reports that frontier models' taste has doubled every 3.0 months since December 2025. It says Opus 5.5 beats the human baseline at about 1/30 of the experts' cost per attempt [8]. Import AI's author now thinks full RSI is more likely in 2027 than 2028 [8]. - **DeepRSI** has agents work for 48 hours on ML tasks taken from ongoing proprietary research. Existing benchmarks top out around 12 hours [9]. Garry Tan calls the next 18 months RSI's "Kitty Hawk moment" and says "evals will be how you can tell" [10]. - Anthropic's internal index says Claude led 26% of measured model R&D in August, with humans supervising. None of the measured work was fully autonomous [11].

On diligence, a16z's Martin Casado and **MakerMaker.AI** founders Dhaval Adjodah and Owen He say many "RSI" pitches are just people using AI to build better AI [12]. Their bar is that the model improves its own weights, harness and task prompt, and that the gains hold up once experiment costs are counted. They track spend on a "bleeder board" and assign one engineer to try to break their results. They report early signs of the system getting better and cheaper at the same time [12]. Investors can use the same questions on RSI pitches: what exactly improves, and at what total cost.

Funding: Series A/B is hard, so get profitable or keep your pro rata

Rohit Mittal describes the trap. Vertical AI SaaS companies reach \$1–3M ARR in two years, then find "almost no interest" from investors at 1–3x growth. They are left with large preference stacks, burn they can't sustain, and exits that leave little for the team [13]. Garry Tan's answer: run profitably with "maximal agentic loops," because \$10M of profit reinvested each year works like "an annual Series A no dilution" [14]. Martin Casado warns VCs against getting hung up on TAM, revenue quality and margins early in new technology waves [15].

Hunter Walk expects more **pro rata conflicts**. Early-stage funds are now big enough to want follow-ons across several rounds. Rounds come faster. Hot secondaries make pro rata feel tradable. Fund-of-funds and SPVs now exist at scale to finance it [16]. A widely shared example: Initialized Fund 1 returned 55x. The first 98 companies produced about 1x DPI, and Instacart, where the fund could not take its pro rata, added another 1x. Coinbase, where Tan

kept taking pro rata, produced the final 53x [17]. Separately, Atlassian’s Mike Cannon-Brookes says raising \$15M at pre-seed has become too easy, and blames VCs for it [18].

Agents as customers

Sierra’s Personal Agent Protocol published its first draft and announced 35 new design partners. The founding group includes Meta, Shopify, Stripe and Walmart. The protocol aims to let companies know when they are dealing with an agent and whom it represents [19]. Bank of America is a design partner [20]. Sarah Guo reports that a Sierra Summit executive said “friction had been a very good business model” and their company would have to change. Her point is that agents threaten revenue that depends on customer inattention [20]. On the founder side, Paul Graham met a founder who declined a sales cofounder because “most of his sales would be to agents,” and Graham admitted he could be right [21].

Anthropic launched dynamic workflows for Claude Managed Agents. Claude now plans the work itself and writes the program that coordinates its subagents, so developers no longer configure the roster [22]. The model is not told its budget while it runs, but developers can set a session cap [22].

Perplexity’s numbers

According to an interview recap, Perplexity is valued at about \$20B, with about \$2B raised and \$750M in annualized revenue as of August, growing 5x year over year [23]. Srinivas cited a one-person company that makes \$3M a year and spends more than \$1M of it on Perplexity Computer [23]. Perplexity says its agents commit to a shared Git repo with under 2% collisions. It also says plain `ls/grep` beat semantic search on answer accuracy, 63% vs 37%, in financial-document analysis [24]. That is a useful data point against assuming vector search is required.

Infrastructure and models

- **Microsoft Maia:** starting next year, Microsoft will scale its Maia chips to multi-gigawatt deployments as the primary inference system for its own products. It claims Maia 300 delivers 40% more tokens per watt than leading GPU systems [25].
- **Distillation:** Anthropic says it has cut the data harvested from Claude through distillation attacks by about 90–99%. David Sacks presents this as evidence that labs, not regulators, should handle distillation. He notes that the administration did not impose the open-model restrictions that were proposed [26].
- **Western open weights still trail Chinese models:** Reflection’s 501B-parameter Beam scores 44.4 on DeepSWE vs 61 for GLM-5.3. Reflection says Beam matches GLM-5.2 on reasoning with 3–4x less compute [8].

Mistral Large 4, tuned for cyber tasks, may be the strongest Western open model at release but still trails Kimi K3 and GLM-5.3 [8].

- **Enterprise model provenance:** Atlassian runs Nemotron, Mistral and other open models but no Chinese models in production. Cannon-Brookes says most large companies that claim to have “built their own model” have fine-tuned someone else’s [27].

Deals and programs

- **HyLight** (YC S23) raised \$14.5M for autonomous, hydrogen-powered airships that inspect power lines, pipelines and railways. Four are flying missions in France, with 15 planned to operate at once by 2027 [28].
- **a16z Speedrun SR008** is taking applications through November 1. It invests up to \$1M plus more than \$10M in partner credits, and runs February 1 to April 16, 2027, in San Francisco [29].
- **Avalanche Energy** signed a letter of intent with the Tri-City Development Council for its first compact-fusion manufacturing plant. The plant is estimated to cost close to \$1B in private funds [30].

Sources

1. X post by @alexandr_wang
2. Muse Wonder: Zuckerberg and Wang’s \$100B AI
3. X post by @AravSrinivas
4. X post by @a16z
5. X post by @bindureddy
6. X post by @bindureddy
7. X post by @deedydas
8. Import AI 476: Creativity and RSI; worse-than-China open weight models; AI goes to space
9. X post by @jparkjmc
10. X post by @garrytan
11. X post by @nathanbenaich
12. X post by @a16z
13. X post by @rohitdotmittal
14. X post by @garrytan
15. X post by @martin_casado
16. Why There Will Be More Pro Rata Conflicts
17. X post by @ThePeelPod
18. Atlassian CEO: Enterprises Won’t Trust Chinese AI | Why The Jury’s Still Out on Specialised Models
19. X post by @btaylor
20. X article by @saranormous
21. X post by @paulg
22. How Anthropic’s Dynamic Workflows Coordinate Claude Agents

23. X post by @MollySOShea
24. X post by @AravSrinivas
25. X post by @satyanadella
26. X post by @DavidSacks
27. X post by @HarryStebbing
28. X post by @ycombinator
29. X post by @andrewchen
30. r/Futurology comment by u/Gari_305