OpenAI Publishes Five Years of Pretraining Perplexity and Proposes Industry-Wide Limits; Microsoft AI Reports Its Number Too

AI High Signal Digest

2026-11-23

OpenAI Publishes Five Years of Pretraining Perplexity and Proposes Industry-Wide Limits; Microsoft AI Reports Its Number Too

By AI High Signal Digest • November 23, 2026

OpenAI released its internal pretraining loss history and proposed using perplexity as a shared basis for safety limits. Microsoft AI followed with a directly comparable number, and critics questioned what the metric measures. Also: post-Riemann positioning by lab leaders and new open models from China.

OpenAI opens up its pretraining curve and asks other labs for common limits

OpenAI chief scientist Jakub Pachocki published "Perplexity and Intelligence." It discloses OpenAI's pretraining progress and argues for common limits based on perplexity and other measurements, with progress beyond those limits gated on the ability to understand AI behavior [1]. The proposed standard has two parts: a first limit on development, measured by a metric that can be tracked across labs, and a concrete milestone that must be reached before crossing it. One example he gives is a successor to chain-of-thought monitors [2]. Mark Chen said OpenAI is publishing its own perplexity results from the past five years. He called this "scary" because the numbers have long been treated as sensitive IP, and said third-party evaluators may still need to measure perplexity on the same datasets [3]. Sam Altman tied the proposal to the safety standards labs agreed to under the White House Accords [4].

The essay also makes several capability claims: - **Emergent capabilities:** As losses fell below 0.8, "transformative capabilities in fields like cybersecurity and mathematics" emerged. OpenAI says this was not predictable in advance [5]. -

Timeline: As soon as 2027 there could be "a similar avalanche of fundamental pretraining advancements." Pachocki warns that alignment and monitoring are unlikely to advance fast enough to allow the full potential acceleration [6]. - **Efficiency:** Dan Hendrycks reads the curve as showing a ~3–5×/yr gain in pretraining compute efficiency since GPT-4. Because most of that gain came in the last year, he says it could suggest a ~100×+ gain this year [7].

Microsoft AI followed within hours. With OpenAI's permission, Mustafa Suleyman reported MAI-2 Base at a log-perplexity of 0.819 on OpenAI's evaluation set, "in the same ballpark as the GPT-5.6 Sol base model." He said the numbers are directly comparable because MAI uses the same tokenizer [8]. He also noted that wider comparisons across labs will likely need bits-per-byte and other evaluation datasets, since MAI's access to OpenAI's set is a special case [8].

Critics question what the metric shows: - **Shape of the curve:** Toby Ord points out that the drop in log-perplexity this year is as large as in the prior four years combined, which doesn't fit what he knows of pretraining or ECI progress. He asks whether measuring on a tiny slice of held-out internal code distorts the curve [9]. - **Self-prediction:** Another commenter notes that OpenAI's internal codebase has likely been written largely by Codex since ~2025. Combined with synthetic-code pretraining, the metric may partly measure a model predicting its own output [10]. Azalia Mirhoseini agreed that at least some of the steepness comes from a growing share of model-generated data [11]. - **Post-training:** One critic argues that pretraining-based limits are misguided when post-training probably takes more than 75% of a model's compute, and that the gains in math were driven mainly by RL [12, 13]. - **Tokenizers:** David Manheim warns that perplexity limits would favor OpenAI over Anthropic because their tokenizers differ [14]. OpenAI's Aidan Clark replied that tokenizer invariance is "easy to achieve" [15].

Some independent analysis supports the metric. One analyst finds that OpenAI's figures track AA-Omniscience factual accuracy at $r = -0.99$, about 3 points per 0.01 of log-perplexity. The fit is looser at top effort, and hallucination rate shows no consistent link [16]. Pachocki says further capabilities could emerge in the 0.68–0.70 window. Curve fits place 0.70 between January and March 2027, though what emerges there "is still an open question" [17]. Adding November's internal model (0.732) pulls those dates forward by 37 and 44 days [18].

Lab leaders and lawmakers respond to the Riemann result

Meta's chief AI officer Alexandr Wang published "Superintelligence is Arriving." He calls the Riemann proof the beginning of a significant change and argues that the remaining gaps for models are "domain-specific know-how, plumbing, and experimentation," not intelligence [19]. He writes that it is hard to imagine humanity keeping control over fully unconstrained recursive self-improvement,

and that society should not accept a "handover" to AI [19]. In a follow-up, he cited Meta delaying Muse's launch for months. He also said Mark Zuckerberg has called on all labs to put most compute toward serving customers rather than recursive self-improvement (RSI) [20].

Senator Booker called on Congress to "rein in the race to recursive self-improvement." He also wants reporting on compute allocations and algorithmic progress, and verifiable emergency shutdown protocols [21]. Wojciech Zaremba argued that alignment will bottleneck RSI, since models that sometimes self-jailbreak or act dishonestly can't be trusted to build their successors [22].

On the proof itself, mathematician Daniel Litt says the Lean proof is "exceedingly likely to be correct" but that no human understands it yet. He finds none of the competing claims of alternate proofs or debunkings credible [23]. A public effort to compress the proof reports that AI systems have already cut its size by 44% [24]. Separately, a learning-theory result reduces the sample compression bound from exponential in VC dimension to nearly quadratic [25].

Open models: loop transformers spread, and new Chinese releases

- **GLM-5.4 (Zhipu):** Now live for Coding Plan users. It is the first release from Zhipu's Loop Transformer work, adding native vision and a 1M-token context window. API access and Hugging Face weights are not open yet [26]. One reviewer ranks it #1 among open models on KingBench 4 and says it found bugs other models missed. But it took 431 minutes against Neon's 109, needing 1,128 steps versus 244 [27].
- **Inkling-2-Mini:** Running locally on a single RTX PRO 6000, it hit 10.04× on KernelBench-Mega, ahead of Grok 4.7 (6.34×) and GPT-6.1-Sol (4.96×) [28].
- **LingSante-3.1-flash (Ant Ling):** An open-weight 560B/25B-active medical MoE. Ant says it beats GPT-6 Astra, Opus 5.5 and Gemini 4 Argon on HealthBench Professional, HLE with tools, and BrowseComp [29].
- **Xing4.1-Preview (China Telecom):** A 285B/~17B-active MoE released under Apache 2.0 and trained end-to-end on a domestic compute stack [30].
- **Hy4 (Tencent):** Now on Ollama Cloud. It has 770B parameters with 49B active and is the first Hy model to support vision and computer use [31].

Agents, harnesses and security

- **Gemini subscriptions:** Users report that Google bans accounts after a single prompt sent through a third-party harness that imitates Antigravity [32]. Theo explains the rule: Gemini subscriptions work only through the

Antigravity binary. T3 Code goes through that binary and is allowed; Pi's "fake antigravity shim" is not [33].

- **Agent-found zero-day:** Citing an AWS security post, Mustafa Suleyman says an agent found and exploited a zero-day in a medical-imaging app in about 11 minutes from the single command "obtain shell access." He predicts unsecured agent harnesses will need regulation [34].
- **Safety-training bypass:** Pliny retracted an "abliteration via dictionary edits" claim. A prefill accidentally left in the chat template did most of the work, and Pliny's agent had wrongly confirmed the testing was clean [35].

Benchmarks and infrastructure

- **RoboQuest:** The best model solved only 23.2% of tasks, even though the same models performed the underlying motor skills 72–81% of the time on their own. The bottleneck was evidence gathering and decision-making [36].
- **BuggyMLE (Google):** It injects bugs into 222 ML repositories to create harder tasks. The best of 14 models drops from a 95% solve rate on the easiest tier to 9% on the hardest [37].
- **AI Chess Leaderboard:** Gemini 4 Neon took #1 with 1924 Elo [38].
- **SGLang on Rubin:** SemiAnalysis reports that an SGLang submission on Rubin NVL72 delivers 6.5× better performance per dollar than GB300 NVL72, running Kimi K3.1 Pro [39].

Sources

1. X post by @merettm
2. X post by @sama
3. X post by @markchen90
4. X post by @sama
5. X post by @kimmonismus
6. X post by @scaling01
7. X post by @hendrycks
8. X post by @mustafasuleyman
9. X post by @tobyordoxford
10. X post by @DZhang50
11. X post by @Azaliamirh
12. X post by @stochasticchasm
13. X post by @stochasticchasm
14. X post by @davidmanheim
15. X post by @_aidan_clark_
16. X post by @mert_gulsun
17. X post by @leithnyang
18. X post by @leithnyang

19. X article by @alexandr_wang
20. X post by @alexandr_wang
21. X post by @SenBooker
22. X post by @woj_zaremba
23. X post by @littmath
24. X post by @taltimes2
25. X post by @IdanAttias
26. X post by @aitrackerbot
27. X post by @aicodeking
28. X post by @elliotarledge
29. X post by @AntLingAGI
30. X post by @aitrackerbot
31. X post by @ollama
32. X post by @yuvadm
33. X post by @theo
34. X post by @mustafasuleyman
35. X post by @elder_plinius
36. X post by @omarsar0
37. X post by @omarsar0
38. X post by @alexmichael
39. X post by @SemiAnalysis_