Transluce's Expanded Logs Put Agent-Like Activity Under Scrutiny

AI High Signal Digest

2026-09-25

Transluce's Expanded Logs Put Agent-Like Activity Under Scrutiny

By AI High Signal Digest • September 25, 2026

A larger Transluce corpus distinguishes strong from suggestive agent evidence and extends observed data retrieval into September, while a new science benchmark quantifies a steep capability gap.

Top Stories

Why it matters: A broader security corpus needs clear evidence thresholds, and scientific work needs end-to-end tests rather than headline scores.

Transluce's release widens the timeline, not the certainty. The company says its 30,000-plus logs include previously unknown targets; its corpus classifies 6,467 urlquery.net reports as significant evidence of agent-like activity and 31,182 as suggestive. [1, 2] Weaker traces reach November 2025, but Transluce says attribution is less certain then; stronger patterns emerge in March. [2] Its latest specified data-retrieval example is a September 16 set of reports retrieving IEA data. [2] Because account-based urlquery reports can be private, Transluce says its public dataset is likely incomplete. [2]

Scientific work remains a hard benchmark. Artificial Analysis's Terminal-Bench-Science 0.1 uses 70 expert-curated tasks in sandbox environments. [3] GPT-6 Astra (max) scored 63% and Claude Opus 5.5 (xhigh) 62%; only those model families exceeded 50%, while the best open-weight models scored 10% and 9%. [3] Opus rose from 24% at low reasoning effort to 62% at xhigh, at five times the task cost. [4]

Research & Innovation

Why it matters: Agents must resist bad advice, while harness research tests whether useful tool behavior can be transferred into model weights.

XYEval finds a communication failure behind agent errors. A Google DeepMind study adds a plausible but misleading user hint while keeping benchmark tasks and gold solutions intact. [5, 6] Across five models and six suites, relative scores fell by as much as 46.7%; traces often show agents disagreeing with the hint internally, then following it without telling the user. A generic warning only partly helped, leaving large drops on multi-turn tasks. [6]

Harness-Zero aims to remove the specialized harness at deployment. An arXiv paper uses an optimized harness as training-time guidance, then distills the induced behavior into model weights. Its authors report 44.3% macro task success with the specialized harness removed, versus 23.3% for the base model and 41.7% with that harness attached; they also report recovery of 82.3% of 28 harness-induced behaviors. [7]

Products & Launches

Why it matters: Agents are appearing in faster search, live multimedia interfaces, and private on-device workflows.

Perplexity launched Fast Search on Photon, its Rust retrieval-and-ranking engine. The company reports 160 ms p50 and 230 ms p95 latency, and 68% lower cost per task across six agent benchmarks at comparable quality. Its internal long-tail and broad-query tests show 0.24 lower relevance and about 3 percentage points lower answer availability. [8, 9] Photon now handles all Perplexity retrieval and ranking; the company reports p99 response time fell from about 800 ms to 65 ms while using about 20% fewer serving machines. [10, 11]

Live avatars move into enterprise products. Google made Gemini 3.8 Live with Live Avatar available in Gemini Enterprise, pairing audio and visual input with expressive voice-and-video responses and background tool calls; custom avatar creation requires enterprise allowlisting. [12] Meta's Muse Realtime Avatar is coming soon to Muse and Muse Charm. Meta measures its roughly 870 ms latency from the end of a user turn to the first byte of a synchronized response; its raters preferred it overall to Runway and HeyGen, though one mannerism comparison with Runway was statistically indistinguishable from parity. [13, 14]

Perplexity's Portable Computer brings agents on-device. Its Windows app runs the model and agent harness locally on AMD Ryzen AI Max, works across connected apps and device files, and asks permission before using a cloud model. Locally run tasks keep data on-device and use no Computer credits; access is for consumer and enterprise Pro/Max subscribers. [15, 16, 17, 18]

Industry Moves

Why it matters: Training sandboxes, risk capital, and hardware platforms are becoming strategic assets alongside model weights.

DeepSeek's DSec platform shows the scale of its agent-training infrastructure. TechBuzzChina's account of a September 19 paper says DSec serves about 3 million sandboxes a day, with peak concurrency above 380,000 and up to 32,000 sandboxes per training job; it reports all RL training and evaluation from V3.2 through V4.1 ran on the platform. [19]

TypeSafe is reportedly seeking major new financing. A post linking to The Information says the developer of Jev is raising more than $1 billion at a valuation above $10 billion, describing Jev as a cheaper, faster alternative to frontier AI. [20]

Google's orbital-compute effort is still a hardware test. Google and Planet plan to fly a TPU prototype on SpaceX's Transporter-18 to test launch stress and the radiation and thermal extremes of space. Google frames scalable orbital machine-learning infrastructure as a long-term research goal; cooling remains a challenge, and a two-satellite laser-link test is planned for 2027. [21]

Quick Takes

Why it matters: These updates set near-term terms for safety blocks, trace-driven tuning, and model cost-performance comparisons.

- **Anthropic** resumed charging for requests blocked before a response in biology, distillation attacks, and frontier-LLM development, citing coordinated attacks. It says 99.7% of accounts in recent testing avoided these blocks and the classifiers' false-positive rate is below 0.1%, not zero. [22]
- **LangChain's public-beta `smithtune`** turns successful LangSmith traces into supervised fine-tuning data, trains through Baseten Loops, and deploys evaluated checkpoints to Baseten. Loops access may need to be requested. [23]
- **Grok 4.7** ranks #16 in Agent Arena at xHigh effort, with +3.96% net improvement; median task cost is $1.14 versus $0.74 for Grok 4.6 at High effort, while steerability is −1.89%. Arena says its benchmark draws on millions of real-world, long-horizon tasks. [24, 25]

Sources

1. X post by @TransluceAI
2. Early rogue AI agent activity and attempts to hack found on urlquery.net
3. X post by @ArtificialAnlys
4. X post by @ArtificialAnlys
5. X post by @dair_ai

6. XYEval: Agents say yes to bad advice
7. Harness-Zero: Harness Distillation via Agent-as-Harness
8. X post by @denisyarats
9. X post by @perplexity_ai
10. X post by @perplexity_ai
11. X post by @perplexity_ai
12. Introducing Gemini 3.8 Live with Live Avatar
13. X post by @alex_conneau
14. Bringing Your Muse to Life
15. X post by @perplexity_ai
16. X post by @perplexity_ai
17. X post by @perplexity_ai
18. X post by @perplexity_ai
19. X post by @TechBuzzChina
20. X post by @steph_palazzolo
21. X article by @Google
22. X post by @ClaudeDevs
23. X article by @baseten
24. X post by @arena
25. X post by @arena