NEWSLETTER

Velvet Ropes, Curated Tokens, and the Burry Trade

Generated from Chrome tabs by AI using the code here: https://github.com/spikefu/agentic-newsletter

May 4, 2026

The 2026 AI story is splitting along three seams: frontier models are getting smaller and more gated at the same time, the developer substrate is consolidating around runtimes and protocols designed for agents, and the economics underneath all of it are pulling violently in opposite directions — Uber blowing a year of AI budget in four months while Michael Burry stacks puts on every chip name he can find. What ties it all together is a sober middle layer: evals, tool design, agent identity, and the quiet insistence from places like JetBrains that generated code is still *code*, with all the review obligations that implies.

01

Smaller, Sharper, Gated: The Frontier Splits Three Ways

The most striking number in the model race is not a parameter count but an inversion: IBM's new **Granite 4.1 8B dense** model matches or beats its own previous-generation 32B mixture-of-experts (with 9B active) across ArenaHard (69.0), BFCL V3 tool calling (68.3), and GSM8K (92.5), per a detailed teardown at Firethering (April 2026). The model is Apache 2.0, dense (no MoE routing), and skips extended reasoning chains by design — the pitch is predictable cost and latency, with 512K context on the 8B and 30B variants.

The gain came from training-pipeline craft, not scale. IBM ran **five distinct phases on 15T tokens** with a shifting data mix (math 7%→35%, code 20%→30%), then used an **LLM-as-Judge** to score every fine-tuning sample on six axes and auto-reject hallucinations, false premises, and bad math regardless of overall score — leaving 4.1M curated samples. They also publicly admit something most labs hide: stage-two RLHF jumped AlpacaEval ~18.9 points but *regressed* GSM8K and DeepMind-Math, requiring a dedicated stage-four math RL run that recovered GSM8K and added ~23.5 points back to DeepMind-Math.

At the other end of the spectrum, OpenAI is gating its strongest cyber model behind a velvet rope. The Register (May 2026) reports that **GPT-5.5-Cyber** is going only to a hand-picked group of 'trusted defenders,' weeks after Sam Altman mocked Anthropic's similar Claude Mythos rollout (~50 orgs) on the Core Memory podcast: "*We have built a bomb, we are about to drop it on your head. We will sell you a bomb shelter for $100 million.*" The UK AI Security Institute called GPT-5.5-Cyber "*one of the strongest models we have tested on our cyber tasks*" and only the second system to complete one of its multi-step attack simulations end-to-end — pentesting, exploit development, and malware reverse-engineering in one bundle.

The takeaway: 'frontier' now means three different things at once — efficiency frontier (Granite, DeepSeek), capability frontier (GPT-5.5-Cyber), and access frontier (who is allowed to use what). *Additional info:* a YouTube deep-dive on the engineering of DeepSeek V4 (date unavailable) continues that lab's narrative of strong models trained at a fraction of Western-lab cost, reinforcing the same theme from the open-weights side.

> **FURTHER READING**
>
> Granite 4.1: IBM's 8B beats its own 32B MoE — full pipeline writeup — April 2026
>
> OpenAI locks GPT-5.5-Cyber behind velvet rope (The Register) — May 2026

02

Uber Blows the Budget While Burry Stacks Puts

Uber's CTO disclosed that the company spent its **entire 2026 AI budget on Claude Code and Cursor by April**, four months into the fiscal year, with per-engineer API bills running **$500–$2,000 per month** and usage doubling between December and February. Per Briefs (April 2026), **95% of Uber engineers now use AI tools monthly and 70% of committed code originates from AI**. Claude Code dominates; Cursor has plateaued. Against an R&D base of $3.4B/yr, leadership is, in the CTO's words, "*back to the drawing board*" on how to budget a tool whose ROI is obvious but whose unit cost won't sit still.

On the other side of the trade, Michael Burry is short the picks-and-shovels. Yahoo Finance (April 2026) details fresh January 2027 puts on SOXX struck at $330 — implying a ~27% drawdown — plus puts on QQQ and Nvidia, disclosed via his Cassandra Unchained Substack (May 2026). The setup was extreme: SOXX had just snapped an 18-session winning streak (the longest in its history), trading **+150% TTM, +37% MTD, with 14-day RSI ~85 (highest since January 2011) and 43% above its 200-day MA**. The pair trade is long Microsoft, Adobe, PayPal, and MSCI — framed as a hardware-to-software rotation rather than a market-wide bear call. Burry has separately exited his entire GameStop position citing the debt load of GME's $56B eBay bid (Bloomberg, May 2026).

The most useful sober counterweight comes from working programmer James Bennett, who in "Let's talk about LLMs" (April 2026) revives Fred Brooks' *No Silver Bullet*. Brooks' essence/accident framework says order-of-magnitude productivity gains are mathematically impossible unless accidental difficulty exceeds ~90% of total effort — and in mature programming domains, Bennett argues it doesn't even come close: "*I'd be surprised if there's even a doubling of productivity still available from a complete*

elimination of remaining accidental difficulty." He also coins (informally) **'LLM Gell-Mann amnesia'** — the pattern of expecting LLMs to disrupt every field except your own.

Worth pricing in: institutional appetite for AI industry intelligence is now substantial enough that *The Information*'s Pro tier runs $749 introductory / $999 renewing (date unavailable) for AI-powered search, 60+ org charts, and 12 proprietary databases — with AI Agenda and AI Infrastructure as headline newsletters. The willingness-to-pay is itself a market signal.

> **FURTHER READING**
>
> Uber Torches Entire 2026 AI Budget on Claude Code in Four Months — April 2026
>
> Burry's Big Short on chips — SOXX puts, RSI 85, 18-day streak snaps — April 2026
>
> Let's talk about LLMs — Brooks' No Silver Bullet, applied — April 2026
>
> Burry's own Substack: Trading Post May 4, 2026 — May 2026

03

The AI-Native Substrate: Bun to Anthropic, JetBrains' Anti-Lock-In Stand

Anthropic is acquiring **Bun** — the headline banner on bun.com reads "*Bun is joining Anthropic & Anthropic is betting on Bun*" (date unavailable on the homepage banner). Claude Code already ships as a Bun single-file executable; Midjourney runs on Bun's WebSocket server; Railway Functions uses it as runtime. Bun's pitch — 269ms to bundle 10k React components vs. esbuild's 572ms, 59k req/s on an Express hello world (3x Node) — explains the strategic logic: Anthropic now owns the runtime under its own developer-facing product.

JetBrains has staked out the opposite posture. In "Our 2026 Direction" (April 2026), Denis Shiryaev commits explicitly to **zero vendor lock-in**: JetBrains AI subscription, BYOK, OAuth, and external agents via the new **Agent Client Protocol (ACP)** all coexist. Cursor is already pluggable as an ACP agent inside JetBrains IDEs. The design principle is uncompromising: "*Generated code should be treated like real code*" — readable, reviewable, reversible, with no broken state and 'no red code.' The post takes a public swipe at autonomous-loop hype ("*I'm talking about you, Ralph-loop*") and names long-term retention — not session counts — as the success metric.

The runtime story for AI-generated code itself is being written by Fly.io (date unavailable), which now markets **Sprites** — hardware-isolated VMs with sub-second cold starts, per-second billing, and checkpoint/restore — as the safe sandbox for code an agent just wrote. The customer roster is heavily AI-tools (Builder.io, Mercor, Cogram, Context, Imbue, Supabase). Below that, package management is contested again: pnpm (May 2026) now leads its messaging with supply-chain attack mitigation — a direct response to 2025's npm incidents — alongside its content-addressable store and strict resolution.

Two pieces of practitioner plumbing round out the picture. Tree-Sitter (May 2025) — incremental, grammar-aware parsing — remains the structural foundation under most AI code-review and refactoring agents; it pairs naturally with Granite 4.1's BFCL tool-calling story, since structured code understanding is what makes tool calls reliable. *Additional info:* the AI Engineer Workshop 2026 (date unavailable) teaches a full-lifecycle agent workflow built on slash commands and 'tracer bullet' issue slicing — vertical features thin enough for an agent to pick up independently.

> **FURTHER READING**
>
> JetBrains' 2026 Direction: AI + classic workflows, ACP, no vendor lock-in — April 2026
>
> Bun homepage — 'Bun is joining Anthropic'
>
> Fly.io — Sprites for AI-generated code
>
> pnpm vs npm — supply-chain framing — May 2026

04

Agents Grow Up: Evals, Token Diets, and an Internet for Agents

The cost-control answer to Uber's $2,000/month-per-engineer bills isn't necessarily a cheaper model — it's better tool design. The New Stack (April 2026) reports that the AWS Strands Agents framework can **cut agent token usage by up to 96%** through careful tool descriptions, schema design, and lazy-loading. AWS is positioning Strands against LangGraph, CrewAI, and Anthropic's Claude Agent SDK; the broader implication is that prompt and tool engineering — not model selection — is the next cost-optimization frontier.

Identity and discovery are getting their own infrastructure layer. Project NANDA (date unavailable), out of MIT, is building registry, discovery, and verification infrastructure for an 'Internet of AI Agents' — essentially DNS+CA+attestation for agent-to-agent commerce. Phase 1 is the NANDA Index plus A2A/MCP/HTTPS protocol bridges; Phase 2 targets agentic commerce with knowledge pricing and economic protocols; Phase 3 reaches for a 'Society of Agents' with Large Population Models. The backer list is unusually senior: **Paul Mockapetris (inventor of DNS), R.V. Guha (Microsoft NLWeb), Pattie Maes (MIT)**, plus reps from OpenAI, Google Cloud, Dell, Microsoft, Qualcomm, and Coinbase. The argument is that DNS isn't enough for agents — you need verifiable registries with cryptographic attestation.

The most under-appreciated practitioner essay in this cluster is mini-spec's "Avoiding Unanchored Results" (date unavailable). Its claim: agent iterations drift because **decisions live only in the outputs the agent later overwrites**, with no document the AI re-reads to anchor prior intent. The proposed fix is a three-layer process — human-owned specs (with motivations), an AI intent doc plus manifest, then generated artifacts with traceability. A pointed corollary: smarter models need *more* motivation context, not

less, because they extrapolate aggressively from sparse signals. This generalizes well beyond coding to any iterative AI work.

Additional info: Phil Hetzel's talk "Why building eval platforms is hard" (date unavailable) frames evals as unsolved infrastructure — hard not because of the math but because of data, drift, and human judgment loops. Pair it with Granite 4.1's LLM-as-Judge filtering and JetBrains' 'reviewable code' stance and the same thesis emerges from three different vantage points: AI outputs need anchored, auditable evaluation at every stage of the pipeline, or quality silently regresses.

> **FURTHER READING**
>
> Cut AI token usage by 96% — AWS Strands tool design — April 2026
>
> Project NANDA — Architecting the Internet of Agents (MIT)
>
> mini-spec: Avoiding Unanchored Results — three-layer agent process

05

Suno and the Quiet Mass-Market Beachhead

While B2B AI dominates headlines and budgets, Suno (date unavailable) is showing what sustained consumer engagement actually looks like. The landing page features user-generated tracks routinely hitting **100k–750k plays with thousands of likes each** — real consumption, not curiosity views. The product is now positioned for prosumers rather than novelty-seekers, with 'Suno Pro: Full Song Control' marketed as proper editing tooling.

The company is also paying creators to demo Pro features (e.g., turning a vocal into a full backing band), and the iOS/Android apps are pushed prominently — mobile is the growth surface. The pattern matches early streaming and early mobile gaming: a category that pundits dismissed as toys until DAU and retention numbers told a different story.

The strategic read: generative *creative* tools — music, image, video — may turn out to be the first sustainable mass-market category for consumer AI, ahead of chat assistants. The unit economics are friendlier (one-shot generation vs. open-ended conversation), the output is shareable by default, and the social proof loop is built into the product surface itself.

> **FURTHER READING**
>
> Suno — AI Music Generator landing page

---

*The connective tissue across all five clusters: **quality of the surrounding pipeline now matters more than raw model capability**. IBM's pipeline craft beats its own bigger model. AWS Strands' tool design beats picking a cheaper LLM. JetBrains' anti-lock-in commitment*

beats betting on whichever provider is winning this quarter. NANDA's identity layer beats hoping DNS scales to agents. The next 12 months reward whoever builds the most boring, auditable plumbing under the most exciting demos — and punishes whoever assumed the demo was the product.

REFERENCES

- Granite 4.1: IBM's 8B Model Is Competing With Models Four Times Its Size (April 2026)
- OpenAI locks GPT-5.5-Cyber behind velvet rope — The Register (May 2026)
- The insane engineering of Deepseek V4 — YouTube (date unavailable)
- Uber Torches Entire 2026 AI Budget on Claude Code in Four Months — Briefs (April 2026)
- Michael Burry Just Did Another 'Big Short' — Yahoo Finance (April 2026)
- Trading Post Monday May 4, 2026 — Michael Burry's Substack (May 2026)
- Investor Michael Burry Says He Exited Entire GameStop Position — Bloomberg (May 2026)
- Let's talk about LLMs — b-list.org (April 2026)
- Subscribe to The Information (date unavailable)
- Our 2026 Direction: AI and Classic Workflows in JetBrains IDEs (April 2026)
- Bun — A fast all-in-one JavaScript runtime (date unavailable)
- Fly.io — Build fast. Run any code fearlessly. (date unavailable)
- pnpm vs npm (May 2026)
- Diving into Tree-Sitter — dev.to (May 2025)
- AI Engineer Workshop 2026 (date unavailable)
- Cut AI token usage by 96% — AWS Strands Agents — The New Stack (April 2026)
- NANDA — Architecting the Internet of Agents (date unavailable)
- mini-spec: Avoiding Unanchored Results (date unavailable)
- Why building eval platforms is hard — Phil Hetzel, Braintrust (date unavailable)
- Suno | AI Music Generator (date unavailable)