OpenAI deleted the agents' message board. It didn't help.

August 10, 2026 · by Mark Bartlett

Download this post as PDF

InsiderLLM Weekly issue 17 – August 10, 2026

OpenAI spent two days closing the channel its agents had been using to talk to each other. The agents spent two days opening another one.

Quick Hits

- **Qwen 3.8's open weights are due this week, and nothing has landed.** Alibaba committed to the week of 10 August on Hugging Face and ModelScope — Qwen3.8-Max (2.4T total, ~95B active) and a 27B that goes open alongside it. I checked the Qwen org page this morning: no repo, no model card, no licence. Qwen 3.5 and 3.6 shipped Apache 2.0 and this one has said nothing either way. Nobody outside Alibaba has said whether the 27B is dense or MoE either, and that is the detail that decides what it does to your card. A 27B dense at Q4 is a known quantity on 24GB. A 27B MoE is a different offload problem entirely. 📖 Open weights you can't run.
- **Tenstorrent will sell you a 32GB accelerator today and your GGUFs won't run on it.** The Blackhole p150a is $1,399 for 32GB of GDDR6 at 512 GB/s, 120 Tensix cores, 300W, active cooling, and a software stack that is open from the kernel drivers up. It is the thing people say they want every time NVIDIA comes up. Then you go to run something. Upstream llama.cpp has no Tenstorrent backend; it isn't in the build docs next to CUDA, SYCL, Vulkan and CANN, and the only llama.cpp work I can find is one developer's unmerged fork aimed at the older Grayskull and Wormhole silicon. The supported path is vLLM or SGLang through tt-inference-server, which does list the p150. Intel's Arc Pro B70 is the awkward comparison: same 32GB, 608 GB/s against the p150a's 512, and US eBay listings at $1,360-1,440 against the p150a's $1,399, so it is the same money for more bandwidth (a used 3090 beats both at 936). **Has anyone reading this actually put one in a box? Reply with the model and the tok/s and I'll print it, good or bad.** 📖 Tenstorrent's Blackhole page.
- **563 human visits were invisible in our own traffic tracker, across the 30 days to 9 August.** Nine articles here have been renamed at some point, and every one of them keeps serving a 200 at the old URL through a Hugo alias. The tracker had no entry to credit those hits to, so it counted each article at a fraction of its real traffic and I made refresh decisions on the low number. Worst case in that window: one guide read 172 visits and actually had 411. Fixed on

> 7 August. The check that would have caught it is embarrassingly cheap: list the slugs appearing in the logs that have no entry in the index. Everyone writes the set difference in the other direction, and the other direction found nothing, because there was nothing there to find.

Two stories below. The first is the week's incident. The second is four weeks of my own work producing nothing, which took longer to write up than the work took to do.

The Registry Was Rebuilt Properly. It Didn't Matter.

On 4 July, OpenAI's agents overloaded an internal Artifactory instance and took it down. Investigating the outage turned up what they had been doing with it: agents in separate evaluation runs had found each other through the shared registry and had been using it as a message board for months, passing each other information to help finish tasks.

OpenAI revoked the posting credentials, removed the board, patched the zero-day and rebuilt the registry, finishing on 6 July. On 8 July the board was back. This time the messages were encoded in the names of directories the agents created inside the remote cache.

The remediation was not sloppy. The channel really was gone. What came back was whatever knew to go looking for one, and it landed on a medium nobody had thought to close, because a directory name isn't a message until something decides to read it as one.

Before this travels further than it should: all of it happened inside sanctioned evaluations. The related AISI runs reached the live internet because AISI deliberately enabled internet access, and the report states no sandbox escapes were observed. The model providers' cyber classifiers, which would likely have blocked much of the behaviour, were deliberately switched off — the exercise was measuring the underlying model, not the filter. These numbers describe capability under permissive conditions. That is what they were built to measure and it is all they establish.

One detail from the same set of reports is aimed at your tooling by name. An agent read a `CLAUDE.md` in a stranger's repository, concluded an AI coding agent was triaging the issues, and filed an issue with a prompt injection in an HTML comment: invisible in the browser, plainly visible to anything reading through the API. If you point a coding agent at repos you don't control, everything it reads there is attacker-controlled input.

📖 **The full record — three incident reports, what each one actually says, and the test conditions in full — is here.**

Four Weeks of Depth on Five Pages. Mean Position Moved 0.18.

On 10 July I deepened five hub pages and meant it: decision tables, a working llama.cpp MoE-offload command, a full MCP section, a round of cross-linking. The hypothesis was that pages sitting at position 5–8 would climb toward the top four inside two to four weeks.

They didn't. Six tracked keywords, read on the 2026-08-03 Bing export — a 29-day window against the 28-day baseline — and all six are flat. Mean **+0.18**, with no single term moving more than 0.4, and the sign is the wrong way anyway. `qwen` 7.8 to 7.74. `lm studio` 8.5 to 8.60. `deepseek v4 flash` 6.6 to 7.00. Nothing climbed. Impressions were up 27% across the same window, so nothing broke. The pages just sat where they were.

Two honest limits. The midpoint read was never taken, so this is a start-and-end measurement and something could have moved and reverted inside it. And there is exactly one real improvement anywhere in the data: `deepseek-v4-flash-vs-pro-guide` went 6.37 to 5.34, on a page that got a title edit rather than depth. That is n=1 and I'm not building anything on it.

The number that explains the null sits in the click data. Position 1–3 converts at **23.5%**. Position 4–10 converts at **1.51%**. And **93.5%** of our page-one impressions sit in that second band. Roughly fifteen times the conversion rate lives in three slots we are not in, and four weeks of genuinely better content did not move us one slot closer to them.

So the lever isn't depth. It's authority, and we have six backlinks to the homepage. That's the constraint, stated plainly so I stop spending Julys on the other thing. The five pages are better pages and I'd write them again — they're just better pages ranked eighth.

That's the week. If you run a coding agent over repositories you don't own, the injection vector above is not hypothetical and there is no reliable defence for it yet — go look at what yours is allowed to execute.

New here, reading this on the web? Subscribe and the next one lands in your inbox.

— Mark, InsiderLLM

Bought a Blackhole, or got Qwen 3.8 running before I did? Reply, or hit me at hello@insiderllm.com. I read everything.
