

Stop writing the answer. Pick it. 47 items waiting.

September 21, 2026 · by Mark Bartlett

[Download this post as PDF](#)

InsiderLLM Weekly issue 22 – September 21, 2026

No new article this week — the bench rig went up instead, so this issue is what is queued and what the logs say about OpenAI.

Quick Hits

- **A schema answered in one forward pass instead of written token by token.** Codacus's llama.cpp fork does it, and the ten-seeds test split is sitting there ready to score it.  [The fork](#).
- **ChatGPT-User fetches of this site are sliding about 5 percent a week; OAI-SearchBot tripled in the same month.** Zero 429s either way. Two readings, one test, next read 9 October.
- **The v0.4.0 pin reproduced on the 3060 within 1 percent,** and the third bench rig is up, with PCIe 3 against 4 on the 3090 as its first job.  [The dataset](#).

Testing Next: Pick, Don't Write

Codacus has a branch of llama.cpp called parallel-decision. Instead of generating a JSON answer token by token, the model is handed a schema with finite fields, every field's allowed values are scored as token paths forking from one shared KV prefix, and all of them come back from a single decode — with a probability on each. The JSON is assembled by code, so it always matches the schema. That is the mechanism Jev sells. The open reconstruction it descends from is Harsha Gondala's Qwen-2.5-1B-RLCD, Apache 2.0 and MLX only, and this fork is the first time I have seen it in llama.cpp with CUDA.

His numbers — from [the video](#), not reproduced here — are 300 ms against 3.5 s for the same extraction on Gemma 4 12B. On an 18-item game-state task, a 35B went 18 of 18, a 12B 15 of 18, and a 2B 3 of 18. Two caveats, both his. Hybrid models with recurrent layers lose the fused batch path, so the speedup does not apply to them. And an MoE with its experts offloaded copies experts across PCIe on any batch over about 30 tokens, which is why his 35B could not get under a second: the trick needs the whole model resident to be fast.

Now the part I care about. The ten-seeds series left a frozen 47-item test split, and base Qwen3.6-27B scored 44.68 percent on it by writing the answer. The same model choosing from the schema, on the same 47 items, is the third substrate for that series, after the prompt compiler and the LoRA adapters, scored against the same base, and it is a number nobody has. Whether picking beats writing on a task where the base already knows the answer half the time is the question that split was built to ask. A plan, not a promise. The rig is up and the branch builds.

📖 The fork: [here](#). 📖 The original engine: [here](#). 📖 The split it will be scored on: [here](#).

Measured 2026-09-21: [Jev Mode on a 3090: 26 ms per Token You Don't Write](#).

What a Site Sees from OpenAI

ChatGPT-User, the agent that fetches a page live when someone asks ChatGPT a question, has been sliding here since late August: 314.9 requests a day over the last week of August, then 285, 282 and 271 over the three weeks since, about 5 percent a week. In the same September, OAI-SearchBot, the agent that builds OpenAI's index of the site, went from a flat 50 a day — where it had sat since June — to 114 to 176 a day, every day since the fourth. Zero 429s to either, peak concurrency 2 to 3 a second, so the rate limiter that ate June is not in the picture.

Two readings fit. One: answers about this site are being served from the index instead of fetched live, in which case the fetches are not lost traffic — they moved. The other: fewer questions are reaching this site at all. The test is human clicks from chatgpt.com in the referrer field. If answers moved to the index, those should hold or rise while fetches fall. So far they have moved with the fetches, 12.4 a day in that August week down to 9.3 last week, which leans toward the second reading, but the base is 54 to 87 clicks a week and the noise on that is about plus or minus 9. A hypothesis, not a finding. Next read is 9 October, three more weeks.

Housekeeping. The v0.4.0 llama.cpp pin now reproduces on the 3060 as well as the 3090: 28.12 tok/s against the July row's 28.1, and 38.55 against 38.9, both within 1 percent. The dataset is at **134 rows**. Body links now supply 23 percent of referrals into /benchmarks/, up from 2, but the page's human traffic has halved since July — the launch curiosity faded and the links did not replace it. The third bench rig is up, an AM4 box whose slot can be switched between PCIe gen 3 and gen 4, and the first job is the 3090 on both, same file, same build, to put a number on what the bus is worth for offload.

📖 The dataset: [here](#).

That's the week. If your structured-output pipeline is writing JSON one token at a time, the branch above is worth an afternoon — and if the model you'd point it at is an offloaded MoE, read his second caveat first. If you run a site, check your own ChatGPT-User line against OAI-SearchBot; I'd like to know whether the slide is only mine.

New here, reading this on the web? [Subscribe](#) and the next one lands in your inbox.

— Mark, InsiderLLM

Run the parallel-decision branch on something? Seen your own ChatGPT-User numbers move? Reply, or hit me at hello@insiderllm.com. I read everything.

Source: <https://insiderllm.com/blog/newsletter-2026-09-21/>

Free guides for running AI locally