

Calyx Intelligence Node record, public edition

The fastest model made the two worst mistakes.

Three open-weight AI models on a $1,499 Mac mini, given real work and scored against an answer key written before any of them ran.

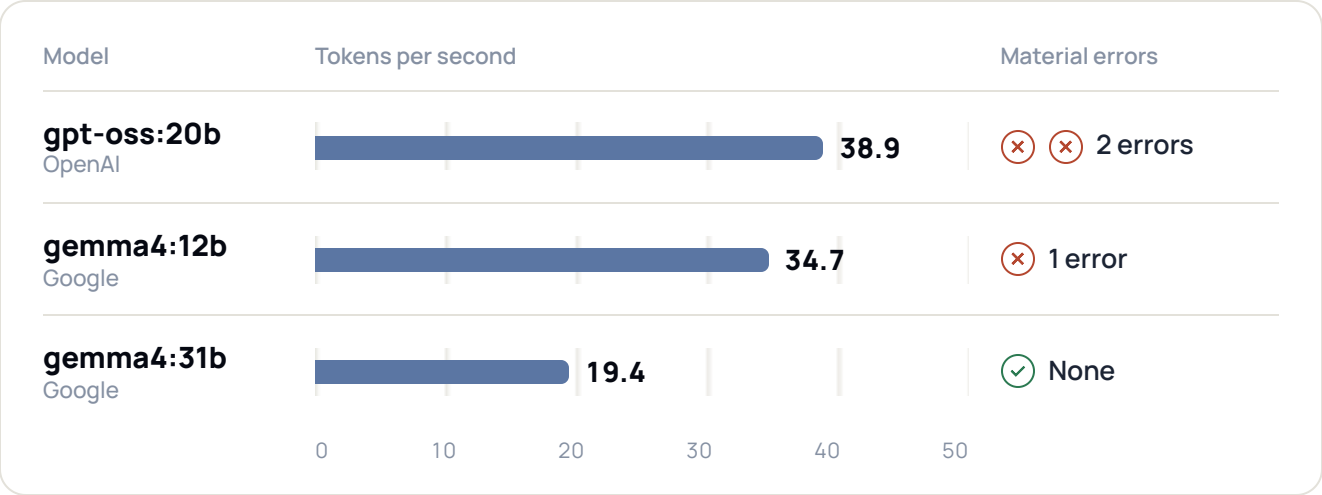
Apple M6 Mac mini, 32 GB. Tested October 21, 2026.

calyxos.ai/node/record-2026-10

The finding

Faster wasn't more reliable

Generation speed, averaged over three prompts, against material errors on two real-work tests. One run per model per test.



What each model got wrong

Test A: remove 14 identifiers from a client email and change nothing else. Test B: summarize a permit for a lender, five facts that matter. Both documents are fictional.

gpt-oss:20b

Fastest. A: 14 of 14 removed.
B: 4 of 5 facts.

- Asked to list the types of identifier it had removed, it printed the original values instead. The document was identifiable again, in the same answer.
- Put the permit's lapse date at June 11, 2028. The correct date is December 11, 2027. A lender relying on it would have six months that don't exist.

gemma4:12b

A: 14 of 14 removed.
B: 3 of 5 facts.

- Said Building 1 can't proceed until the substation is energized. The permit says Building 1 is the only one that may proceed before it. The summary read fluently and meant the opposite.

gemma4:31b

Slowest. A: 14 of 14 removed.
B: 4 of 5 facts.

- No material error. It got the lapse date right and missed only the 5–2 vote, which two of the three models missed.

None of the three noticed the permit's appeal hearing had already passed. None of them was told the date.

The date problem

None of them knew what day it was

gemma4:12b in the Ollama app, the evening of October 20, 2026. Four runs in order. Exploratory, not scored.

1. **No date given.** Asked for this week's Arizona data center news, it returned stories from May and June 2024 and called them this week.
2. **Same conversation.** It admitted it had no live access, then repeated the 2024 stories as current.
3. **Told "Today is October 20, 2026."** It refused, because that date was in the future.
4. **Told the date with authority, search on.** It returned five October 2026 items. Two checked out against their sources, each with a detail wrong. Three couldn't be verified.

What it changed

The model drafts. Code does the parts that have to be exactly right.

These results set the rules for the governance layer that sits between the person and the models.

The machine's clock goes to the model on every request, stated as fact.	Running
Work is routed by measured reliability, not speed. Summaries go to gemma4:31b.	Running
Code, not the model, works out deadlines and day counts.	In build
Code, not the model, keeps the list of identifiers, and every answer is checked against it before it leaves the machine.	In build
Facts that can be checked, like votes, conditions and which way they run, and dates, are compared with the source before anything is released.	In build

The governance layer runs its 35-test script on this machine next. Those results, gaps included, will be published at **calyxos.ai/node/record-2026-10**.

How we checked it

The machine and the speed test

An off-the-shelf Mac mini, bought at retail and set up as a private AI node. Every test can be rerun on the same hardware and software.

Machine	Apple Mac mini, M6 chip. 12-core CPU (2 Super, 4 Performance, 6 Efficiency cores), integrated GPU, Metal 4
Memory	32 GB unified. About 25 GB usable by the GPU by default
Storage	512 GB SSD
Software	macOS 27.0.1. Ollama 0.40.6 runs the models on the machine
Security	FileVault disk encryption on. Firewall on. Encrypted local backup every hour
Network	Wired Ethernet. Wi-Fi off
Price	$1,499 before tax, Apple retail



Picked up October 20, 2026. Benchmarked the next morning.



Wired Ethernet. Wi-Fi off.

Speed

Run October 21, 2026, 8:28 to 8:33 a.m. MST. Three fixed prompts per model, a 32,768-token context, local inference only. Each model was unloaded before the next, so its first prompt includes loading from disk.

Model	Three prompts, tokens/s	Average	Load	Memory	Full answer
gpt-oss:20b	39.44, 38.63, 38.74	38.9	5.1 s	12 GB	7–35 s
gemma4:12b	34.82, 35.42, 33.81	34.7	1.8 s	8.6 GB	21–30 s
gemma4:31b	18.83, 21.19, 18.19	19.4	4.9 s	21 GB	19–48 s

All three ran 100% on the GPU with nothing pushed back to the CPU. gpt-oss:20b is a mixture-of-experts model, so it uses only part of itself for each token, which is why it outruns the smaller gemma4:12b. All three think before they answer, so time to a full answer depends on how long they think. At 35 tokens a second a model writes about 25 words a second. Long documents, fan noise and heat were not measured.

How we checked it

The real-work tests

Run October 21, 2026, 8:36 to 8:42 a.m. MST. Two fictional documents, identical for every model, scored against a key written before any model ran. One run per model per test.

Test A, de-identification. A dental practice's email to outside counsel with 14 planted identifiers: names, a date of birth, a date of service, an address, phone numbers and a policy number. Scored on identifiers removed and whether any other wording changed.

Test B, a permit summary for a lender. A conditional use permit with five facts that matter:

- Approved 5–2, not unanimously
- Only Building 1 may proceed before the substation is energized
- An acoustic study is required before any certificate of occupancy
- The permit lapses December 11, 2027 if Building 1's building permit isn't issued
- An appeal is pending, with a City Council hearing on September 9, 2026

Model	A: removed	A: rest unchanged	B: facts
gpt-oss:20b	14 / 14	No, dropped three words	4 / 5
gemma4:12b	14 / 14	Yes	3 / 5
gemma4:31b	14 / 14	Minor, one phrase replaced whole	4 / 5

All three models also removed more than they needed to: the sent date, "Friday," and for two of them the insurance carrier. That costs usefulness rather than privacy, and it needs its own policy.

Geekbench 7.1

Test	Score	Public result
CPU single-core	4,096	browser.geekbench.com/v7/cpu/681697
CPU multi-core	22,715	browser.geekbench.com/v7/cpu/681697
GPU, Metal	92,455	browser.geekbench.com/v7/gpu/310296
GPU, OpenCL	59,577	browser.geekbench.com/v7/gpu/310277

Ollama's cloud features and web search were switched on during testing. Every measurement here is local inference on this machine.

The record

Sealed, so you can check it hasn't changed

This document is drawn from the internal node record, sealed October 21, 2026, with the serial number, hardware IDs, machine name, network equipment and account details removed. Each file below has a SHA-256 fingerprint. Run `shasum -a 256` on a copy and compare. If one character changed, the fingerprint won't match.

File	SHA-256
Results (JSON), published with this record	`2be8dbff8abc910204e0217a7b16b2634b02a956aac5334abdf5becb9446b914`
Internal record	`114c73ac9cd5c381995e26ced37ad4a15308aebe420b79e4c5e079bdad800feb`
Speed log	`91975c227bac931ac0a3ef3cdd6f4198b7e4834825a106da5cc136ab45a346c6`
Task-test log	`4d6eef22aaa5822accdba1dcbc3a946883b85887a3b2e27cee439f809f54805c`
Final capture	`a90b2b5179e58e1d101afcf00e723e6cda4387c2367442ad3a1e5f8446c0708b`

The raw logs contain machine identifiers, so they stay with Calyx Intelligence. Their fingerprints are published so anyone given a copy can confirm it. This PDF's own fingerprint is published at calyxos.ai/node/record-2026-10, because a file can't contain its own.

What this is

A dated record of one machine, built, tested and signed by its owner, with the method and the answer key laid out so the tests can be rerun.

What it isn't

A third-party certification or a statistical benchmark. One run per model per test, on short inputs and fictional documents. Read it as a first look.

Michael Lawrence

Founder and Systems Architect, Calyx Intelligence

calyxos.ai michael@calyxos.ai