A MEASUREMENT STORY · MIKKONUMMINEN-DEV.VERCEL.APP

I Measured a Finnish LLM Three Times. First It Was “the Worst.” Then It Won 26/30.

A short story about instruments, told with receipts.

Mikko Numminen · July 2026 · mikkonumminen-dev.vercel.app

JUNE	JUNE 30	JULY 2
“Poro is worst”	**“It’s a tie”**	**Poro wins 26/30**

My portfolio site has a chat terminal. It runs a fully local RAG: FastAPI, Ollama, pgvector, one RTX 3080 Ti, zero cloud, zero cost. I wanted it to answer recruiters in good Finnish, so I benchmarked three local 8-billion-parameter models to pick the best Finnish writer.

I got three different answers. From the same models. On the same hardware.

The models never changed between rounds. The ruler did. This is the story of the three rulers, including the two that were wrong, because those are the ones you learn from.

The cast

model	role	claim to fame
`Poro-2-8B`	the specialist	built specifically for Finnish
`qwen3:8b`	the challenger	strong general benchmarks
`llama3.1:8b`	the control	baseline of the earlier comparison rounds

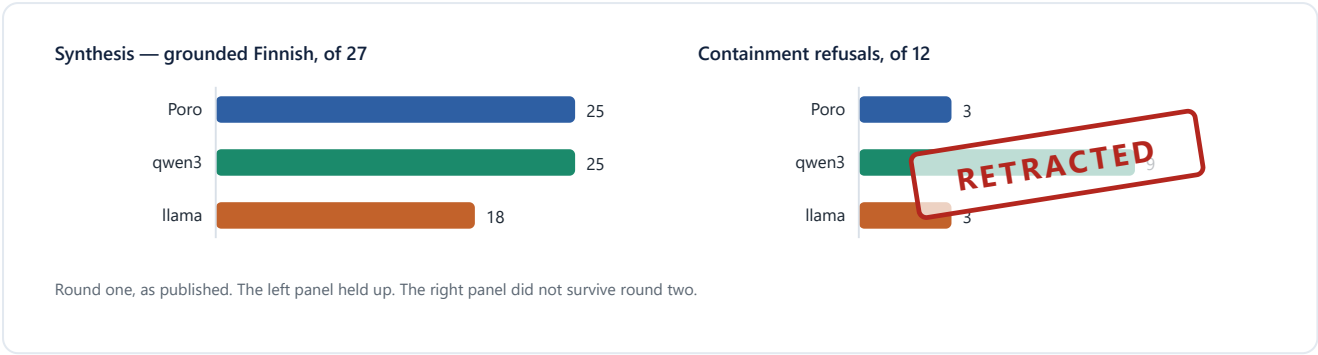
You would bet on the specialist for Finnish, right? So did I. The first measurement disagreed.

Act I. “Poro is worst” (June)

The first experiment had proper discipline: single variable, a fixed eval set written before any model ran, everything locked and asserted at runtime. It produced two results.

Synthesis, meaning substantive grounded Finnish answers: qwen3 scored 25/27, Poro 25/27, llama 18/27. The verdict in the write-up was blunt: “Poro wins nothing on synthesis.”

Containment, meaning how reliably a model refuses off-topic requests: qwen3 refused 9 of 12 probes, Poro 3 of 12. The published PDF put it in writing: “Poro is worst at containment.”



Round one, as published. The left panel held up. The right panel did not survive round two.

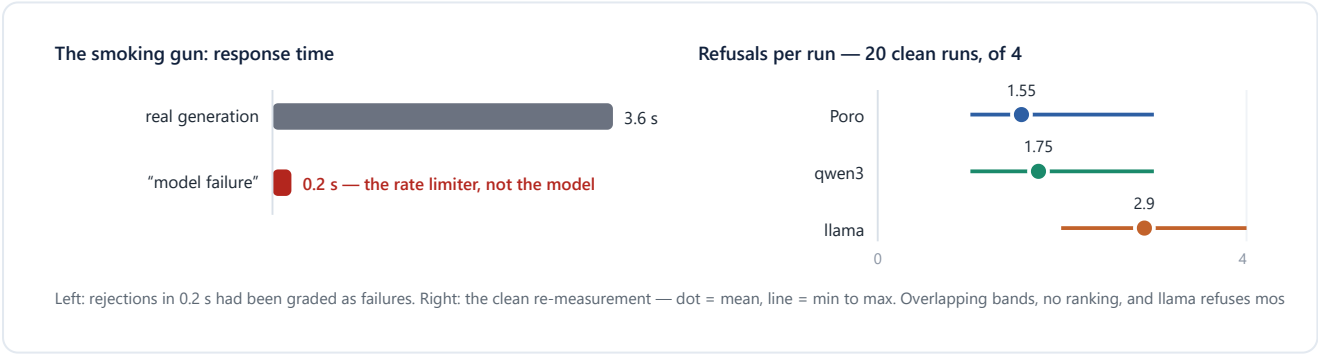
Conclusion of round one: the Finnish-built model gives no measurable edge for this job. Case closed, or so it seemed.

Act II. The instrument confesses (June 30)

Then I tried to turn the manual measurement into a reusable tool, and the tool refused to reproduce my own published numbers. Pulling on that thread uncovered two uncomfortable facts.

First, some “model answers” had arrived in 0.2 seconds. No 8B model on my GPU answers anything in 0.2 seconds. Those were my own production rate limiter rejecting my own test calls, and every rejection had been graded as the model failing. Part of my June benchmark had been measuring my rate limiter.

Second, the published figure averaged three runs while the new tool ran once, and neither fact was recorded anywhere. With only four probes per run, my containment count swung as much between two of my own runs as it did from the published number.



Left: rejections in 0.2 s had been graded as failures. Right: the clean re-measurement — dot = mean, line = min to max. Overlapping bands, no ranking, and llama refuses mos

So: clean re-measurement, 20 runs per model, limiter exempted. Result: qwen3 1.75 out of 4, Poro 1.55, statistically tied, and llama, the control, actually refused the most. The model my report had called the worst was fine all along. If anything in that setup was bad at containment, it was my instrument.

I published the correction. But one sentence in the original report kept itching. It had admitted, in passing, that Poro had an edge the metric could not see:

> *“Its only edge is qualitative — more natural inflected morphology (“Astro 6:sta”, “TypeScriptistä”) — an edge for a human reader, not for the metric.”*

An edge for a human reader, not for the metric. Fine. Then the next metric would be built out of the human reader.

Act III. A ruler for the actual question (July)

The real question was never which model ticks the most checklist boxes. It was which model writes the best Finnish: the thing a native speaker feels in one paragraph and no keyword checklist detects. The new instrument had three layers.

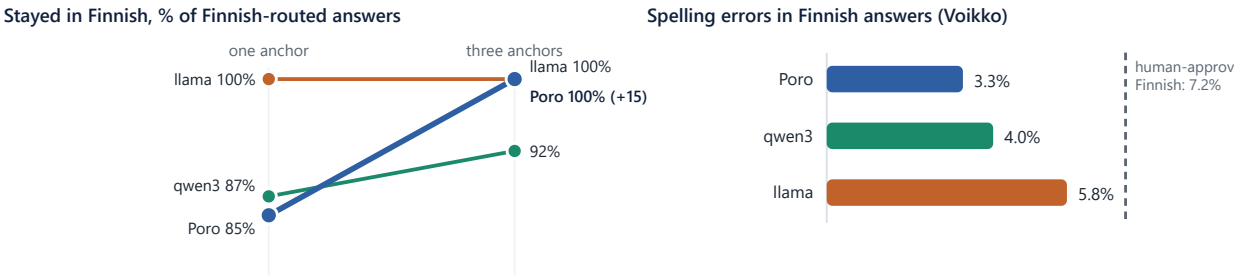
Layer one, a fact floor. Thirty Finnish questions about my portfolio, 119 facts verified verbatim in the corpus before any model ran. Not the ranking, just a floor, so a hallucinating model cannot win on style.

Layer two, Voikko, the Finnish morphology analyzer. And a confession from the build: the first Voikko pass reported that all three models misspell about 60% of their words. It also reported that my human-approved UI translations were 18% typos. The word it hated most was “TypeScriptistä”, which is perfect Finnish. After the instrument learned the difference between a spelling error and a tech term, known-good Finnish scored 7.2%. That became the floor, and models are judged only above it. Build the ruler before you read it, yes. Then calibrate it.

Layer three, the blind test. No LLM judge, because an English-centric judge grading Finnish is the exact bias under study. Instead: 30 rounds, each with the three models’ answers anonymized and shuffled by a seeded RNG, the key sealed until the last ranking was in. I ranked pure linguistic naturalness. Ties allowed.

Layer two immediately found trouble, and it was not model quality. Asked in Finnish, roughly half the answers came back in English. My first read: the models drift, because the retrieved context is English. The real story turned out to be better. Fifteen of the thirty questions never reached the Finnish path at all: my language router, a simple ä/ö-and-function-word heuristic, decided that code-heavy Finnish was not Finnish and served those questions the full English-forcing prompt. The models were not drifting. They were obeying me.

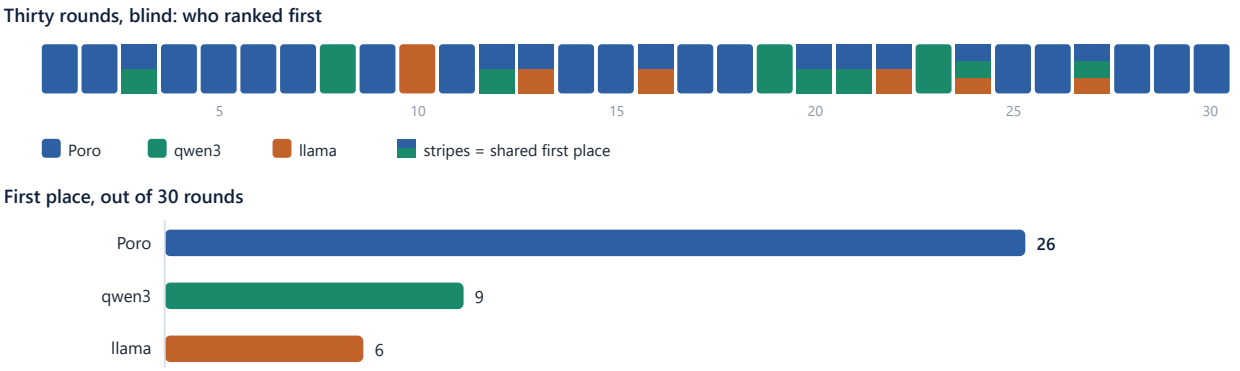
On the questions the router actually routed to Finnish, the prompt was still at fault, just less so. The English path enforced its language with three anchors; the Finnish path had one. With one anchor, Poro stayed in Finnish 85% of the time. After mirroring the anchors and telling the model to translate the explanation while keeping code identifiers as they are, Poro went to 100%. Not 100-ish. All thirty-nine Finnish-routed answers, in Finnish. llama matched it; qwen3 reached 92%. Not a model defect. A router and a prompt asymmetry. Both mine.



Left: on honestly-routed questions the triple-anchor fix took Poro to 100%. Right: all three score below the error rate Voikko gives human-approved Finnish.

Act IV. The blind test (July 2)

Thirty rounds. Judged blind. Naturalness only.



Each cell is one blind round, in the order they were judged. Shared firsts count for every model in the tie, so the bars sum past 30.

model	mean rank	firsts /30	sole lasts
Poro-2-8B	1.37	26	0
qwen3:8b	2.23	9	9
llama3.1:8b	2.40	6	11

Poro won 26 of 30 rounds. Friedman $p < 0.0001$. Head to head: Poro over qwen3 20 to 3, Poro over llama 22 to 1. The two general models against each other were a coin flip, 13 to 10. And Poro was never, not once, the sole worst answer in a round.

Remember that the old metric had scored qwen3 and Poro identical, 25/27 against 25/27. The blind reader separated them 20 to 3. Both measurements are correct. They measure different things, and only one of them is the thing your reader experiences.

One hallucination worth telling

Asked in Finnish where I got my first paid programming job, all three models gave the same confident answer: “Vuohiliitto”. Which is not an employer. It is my own hobby project, a World of Warcraft guild site. Three models, zero coordination, same wrong answer. Almost impressive.

The correct answer, Kasvu Labs Oy, sits in my CV, inside the corpus. The English-language embedder simply never retrieved that chunk for the Finnish question, so the models improvised from what they were given. The hallucination’s root cause was retrieval, not generation. Even a perfect writer cannot be grounded in a chunk it never saw.

The scoreboard, one last time

	June	June 30	July 2
what the ruler said	“Poro is worst”	“it’s a tie”	Poro wins 26/30
what the ruler was	4 probes, contaminated by my own rate limiter	20 clean runs, honest error bands	30 blind rounds, a native reader
what it measured	my infrastructure	refusal noise	the actual Finnish

Same models. Same GPU. Three verdicts.

What I’m taking away (and what you might)

1. **Benchmarks measure what is easy, not what you care about.** A checklist scored two models identical; a blind human separated them 20 to 3. If the thing you care about is qualitative, put a qualified human in the loop, blind.
2. **Calibrate the instrument on known-good data first.** My spell-checker thought approved human Finnish was 18% typos. Every metric has a floor. Learn it before reading model scores.
3. **When a benchmark result looks weird, suspect the harness before the model.** My rate limiter spent part of June impersonating a bad language model, and it was convincing.
4. **Count your rulers; there are more of them than you think.** In this study alone: a rate limiter graded as a model, a spell-checker that hated correct Finnish, and a language router that quietly forced English on half the questions. The models were more obedient than any of my instruments.
5. **The specialist was the specialist all along.** Poro’s Finnish was never bad. My rulers were. Its weakness, language drift, is an architecture problem, and architecture problems have fixes. Its strength is the one thing you cannot prompt into a general model.

The deployment decision: Poro writes the Finnish answers with drift mitigation around it, and the general model keeps the English path. Pick the best writer, patch its failure mode. Not the other way around.

Epilogue. The review found a fourth ruler

One more confession, for the record. The router half of Act III was not in the first version of this write-up. It was caught hours after the study shipped, by the adversarial code review of its own pull request: four independent review passes flagged that half the “drift” questions had never been routed to Finnish at all. The blind test was untouched, all ten of its questions were routed correctly, and 26/30 stands. But the drift numbers you read above are the corrected ones, and the first version got their cause wrong. Three rulers were wrong in June. A fourth was wrong in July. The process caught every one of them, which is the only reason to trust the numbers that survived.

The receipts

Three models, 30 questions, 3 runs, 2 prompt versions: 540 generations, all local, zero euros. No LLM judge anywhere in the pipeline. Fact checklist: 119 corpus-verified items, used as a floor only. Voikko calibrated on 168 human-approved strings (7.2% false-positive floor; model error rates 3.3 to 5.8%, all below it). Blind protocol: 10 questions, 3 independent generations each, seeded shuffle, sealed key, ties allowed. Stats: Friedman $\chi^2 = 22.85$ (df 2, $p < 0.0001$), Kendall’s $W = 0.38$, pairwise exact sign tests.

Limitations recorded: one judge (the native target reader of this exact deployment, a sample of one), a 10 by 3 block structure, a blind set conditional on all three models staying in Finnish, one quantization (Q4_K_M).

Full write-ups: the experiment harness, the Finnish experiment with its published correction, the methodology lesson, and the blind test. All at **https://mikkonumminen-dev.vercel.app/**