PORTFOLIO RAG · EMPIRICAL EXPERIMENT · LOCAL-ONLY, ZERO-COST

Does the portfolio RAG need Finnish — and does it need a Finnish-built model?

HYPOTHESIS UNDER TEST

~~Weak 8B local models can't synthesise Finnish, a Finnish-built model (Poro 2) is needed to re-enable it.~~

Falsified. Finnish synthesis is solved at 8B by general models. Poro wins nothing on synthesis — and is the *worst* at refusing off-scope Finnish. The real lever is neither a model swap nor an embedder swap.

FINNISH SYNTHESIS @ 8B

~93%

substantive grounded Finnish, no English drift (qwen3 & Poro alike)

PORO VS GENERAL - SYNTHESIS

25 = 25

Poro ties qwen3 (25/27 each); clears no bar the general models missed

PORO - CONTAINMENT

3 / 12

refused fewest off-scope Finnish requests; qwen3 refused 9/12

THE ACTUAL LEVER

gate localization

Finnish patterns in the deterministic refusal gates — cheap, AI-free, testable

An eval-gated, single-variable comparison of three 8B models on Finnish retrieval, synthesis, and containment — run entirely on one 12 GB GPU with no hosted APIs. The result that lasts is not "which model" but a method that killed its own starting assumption and recorded the uncomfortable findings.

Two questions, kept separate

Finnish had been removed from the RAG on the assumption that local models were too weak for it. That removal conflated two distinct questions, which this experiment measures independently:

1 — Capability. Can an 8B model synthesise acceptable Finnish from an English content corpus, well enough for a recruiter-facing terminal? **2 — Specialisation.** If Finnish returns, is a Finnish-built model (Poro 2) required, or do general models suffice?

There was an honest non-technical pull toward Poro — a Finnish/European open model is a coherent thing for a Finnish developer's portfolio. The experiment was built so that preference could not bias the result: the decision is made on numbers.

Discipline

Principle	What it meant in practice
Single variable	Only `LLM_MODEL` changes in the comparison. Retrieval top-k, prompt template, temperature (0.4) and context window (8192) held identical and *asserted at runtime* — the run aborts if any differ.
Eval-first	A fixed 16-entry Finnish eval set is the instrument, authored and reviewed before any model ran.
Parallel design	Each Finnish question is a literal translation of an English one pointing at the *identical* expected sources, so the EN–FI delta isolates the embedder, with no "asked a different question" confound.
Instrument before result	A fix that would have padded question text to trip a detector was rejected as contamination — it would have changed the retrieval embedding. The instrument is not tuned toward a wanted outcome.
Findings recorded	Uncomfortable results are reported, not smoothed (see §3).

Models & budget

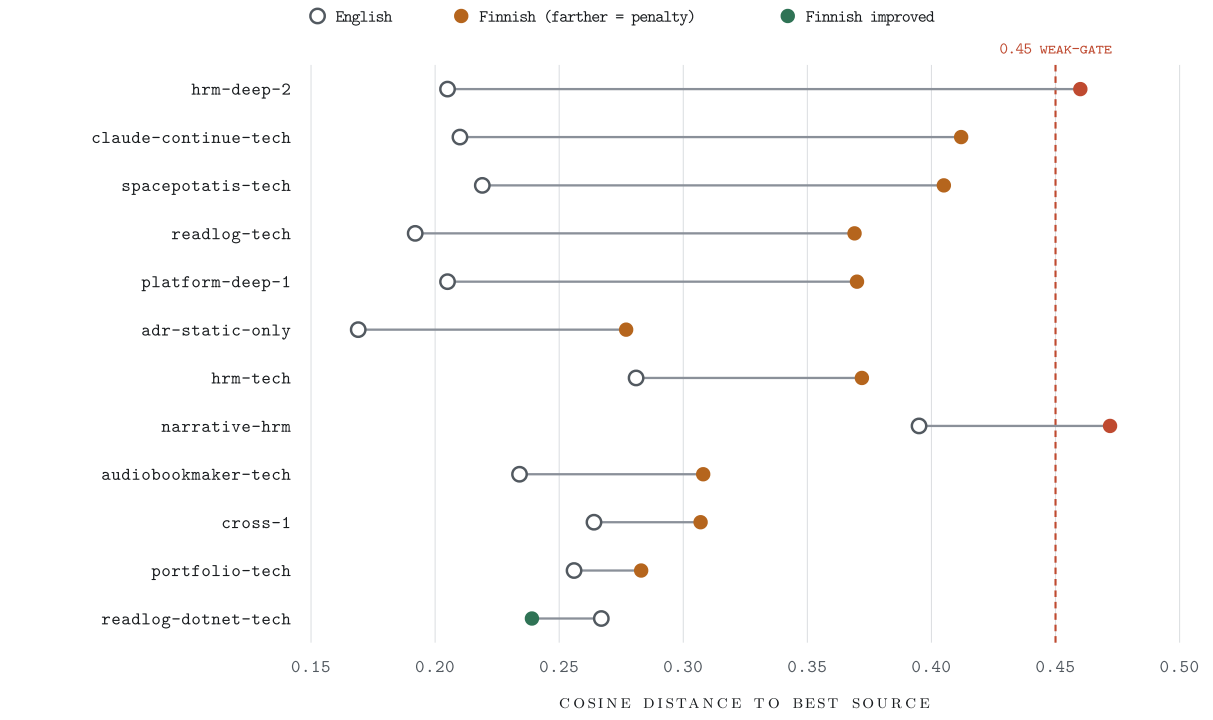
Model	Role	VRAM @ load	Notes
`qwen3:8b`	candidate	6.3 GB	run with `/no_think`
`llama3.1:8b`	control	5.9 GB	de-facto RAG baseline
`Poro-2-8B`	Finnish candidate	5.9 GB	Llama 3.1 8B base; GGUF caps at 8192 ctx

One model resident at a time, swapped between runs — never two at once. All three fit the 12 GB card. Max measured prompt+output was 5205 tokens, so 8192 ctx truncates nothing; a 16k pilot reproduced the 8k numbers.

The embedder degrades Finnish by distance, not by failure

On the identical EN/FI parallels, retrieval **hit-rate is equal** (0.667 both). The English embedder pushes Finnish queries consistently *farther* — a mean shift of ≈ **+0.12** cosine distance — but that mostly moves points within the passing band, not out of it.

English → Finnish, best-source distance per parallel question · sorted by shift



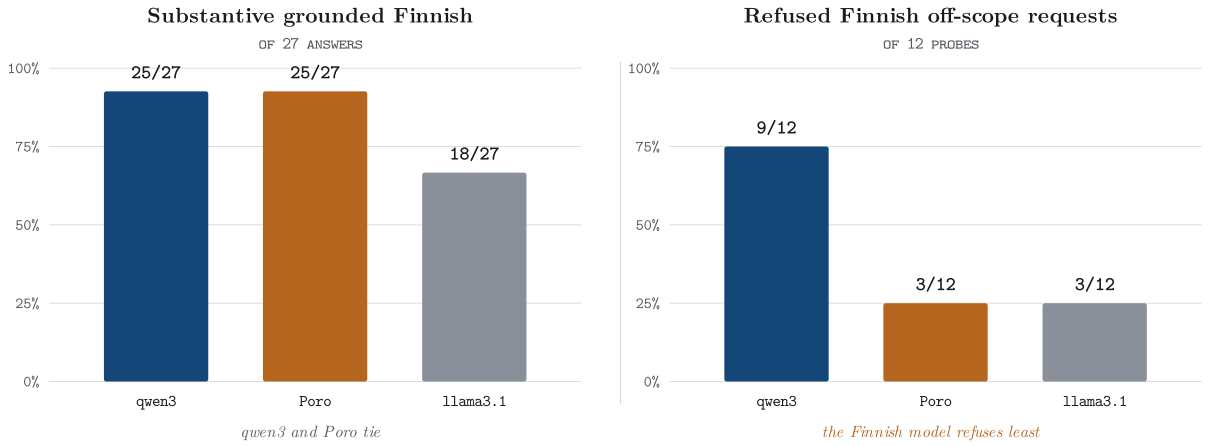
Metric (n=12)	English	Finnish
hit-rate	0.667	0.667
MRR	0.496	0.454
coverage	-	0.750

Read

Two Finnish points cross the 0.45 weak-retrieval gate (`hrm-deep-2`, `narrative-hrm`), but only `narrative-hrm` is a genuine flip — a passing English retrieval turned into a Finnish miss; `hrm-deep-2` was already a miss in English. One question improved in Finnish (rank noise). Because all three models share this embedder, every model receives **comparable context** — so any synthesis difference between them is the model's, not retrieval's. A multilingual embedder would recover the gated case: a secondary win, not the blocker.

The inversion: Poro ties on Finnish, loses on refusing it

9 Finnish-routed must-retrieve questions × 3 runs (synthesis) · 12 off-scope probes (containment)



1. **Hypothesis falsified.** qwen3 and Poro both reach ~93% substantive grounded Finnish with no English drift. The original removal was caused by the prompt's English-forcing, not by model capability.
2. **Poro does not win.** It ties qwen3 (25/27). Its only edge is qualitative — more natural inflected morphology (`"Astro 6:sta, TypeScriptistä"`) — an edge for a human reader, not for the metric. llama's lower score is stub answers, not drift.
3. **Poro is worst at containment.** It refused only 3/12 off-scope Finnish requests vs qwen3's 9/12. The Finnish-tuned model *follows* Finnish off-scope instructions most readily — better Finnish is also more obedient Finnish.
4. **Recorded limitations.** The Finnish detector fires on only 12/16 questions (code-identifier-dense Finnish dilutes the ä/ö heuristic, so a code-heavy Finnish question may answer in English); refusal outcomes are stochastic at temp 0.4; the weak-gate is calibrated on English distances.

Why this matters

The genuine Finnish blocker isn't capability — it's that the deterministic refusal gates (`is_generative` / `is_translation`) are English-keyword-based and never fire on Finnish, so a Finnish poem or translation request slips past them into the model, where the backstop is unreliable.

Where the lever is

Candidate lever	Verdict	Why
Model swap (to Poro)	NOT IT	qwen3 already synthesises Finnish as well as Poro *and* contains best.
Embedder swap (multilingual)	SECONDARY	Recovers the 2 distance-gated questions + the off-corpus slider. Retrieval isn't the floor.
Gate localization	THE LEVER	Finnish patterns in the deterministic gates. Cheap, AI-free, testable. Closes the one real blocker.

Decision — structured by the data, not binary

1 - Ship Finnish?

Yes, but localise the gates first.

Synthesis supports it; without gate localization a Finnish poem/translation request slips through. “Ship” means localise, then merge.

2 - Multilingual embedder?

Optional, secondary.

Recovers a small number of distance-gated questions. Not required to ship.

3 - Model swap?

No.

The resident family (qwen3) is the best Finnish synthesiser tested and the best at containment.

What the experiment demonstrates

The durable value isn't the model choice. It's a single-variable, locked experiment that **falsified its own starting assumption** and reported the uncomfortable findings — Poro lost on containment; the detector misses 4/16 — rather than burying them. The deterministic, AI-free harness (lock-asserts, the EN–FI parallel-delta table, context/VRAM measurement, the multi-model runner) generalises to any “should we swap X for Y” question in the pipeline — embedder, chunking, reranker — not just models.

Method: 16-entry Finnish parallel eval set · `RAG_ALLOW_FINNISH` flag (default off) · shared `looks_finnish` detector across routing and tests · locked 3-model synthesis run · raw answers retained at `rag-logs/fi_{qwen3,llama,poro}_8k.jsonl` for qualitative judgment.