

PinoAI hypothesis tracker

Living document · updated 2026-10-01 (judge result, corrected Part A, three new hypotheses) · Taglish edition: HYPOTHESES.tl.qmd

Stan Lee V. Natividad
Kooya PH AI Research
stan@kooya.ph

2026-10-01

Summary

2 supported · 1 refuted in part · 1 in test · 7 untested · 1 registered, awaiting results · 2 proposed (not yet registered).

ID	Status	Hypothesis in one line
H-DEC-001 B	Supported	Gemma 4 12B has the lowest Filipino tokens per word among the PH-Core candidates
H-OPS-JUDGE	Refuted (PASS; FAIL on privacy/IF); open teachers not confirmed	A confident LLM-judge verdict matches a native reviewer well enough to skip human review
H4	In test	Tokenizer efficiency affects training cost, inference cost and PH performance
H-DEC-001 A	Supported	Gemma 4 12B beats the other PH-Core candidates on PH-EVAL Filipino + Taglish
MAIN	Untested	PH continued pretraining, native instructions and preference optimization improve PH ability without pretraining from scratch
H1	Untested	Continued pretraining on PH corpora beats instruction tuning alone
H2	Untested	Explicit Taglish training gives more natural code-switching than translated data
H3	Untested	Replay data reduces forgetting during PH continued pretraining
H5	Untested	A PH-specific eval suite exposes failures that standard benchmarks miss
H6	Untested	PH domain models can be derived efficiently from one PH foundation model
H-TOK-SEA	Untested	Regional (SEA) pretraining brings better PH tokenization
H-FRONTIER-001	Registered, awaiting results	Where frontier API models stand on PH-EVAL v0.1 against Gemma 4 12B-it
H-DOM-001	Proposed, not yet registered	Light Filipino CPT improves coding after SFT vs SFT alone, especially on Taglish-stated problems
H-CPT-001	Proposed, not yet registered (draft criteria)	The ABL-CPT-001 success criteria C1–C4

Details

H-DEC-001 B: Gemma 4 has the lowest Filipino tokens per word · Supported

- **Source:** decision log DEC-001. Pre-registered 2026-09-29T04:20Z, before any result.
- **Test:** `tokens_per_word` on FLORES+ devtest `fil_Latn`; lower is better.
- **Evidence:** gemma4 1.840 < qwen3.5 1.974 < sailor2 2.098 (EXP-TOK-001). Independently reproduced with 0 differences across 876 values (EXP-TOK-001).

H-OPS-JUDGE: confident LLM-judge verdicts can replace human review · Refuted for PASS verdicts

- **Source:** operational assumption of the PH-EVAL scoring pipeline.

- **Evidence:** calibration #1 gave 14/20 agreement, Cohen’s κ 0.37, with 5 false PASSes vs 1 false FAIL (JUDGE-001; recomputed in FINDING-001).
- **Consequence:** the judge’s FAIL verdicts are accepted. Its PASS verdicts on privacy, hallucination and instruction following goes to a human. judge policy v0.3 is awaiting validation on fresh outputs (JUDGE-001).
- **Open teachers (DEC-002, JUDGE-002, pre-registered, scored once):** pass/fail silver labels are **not confirmed for either teacher**, so pass/fail training labels stay human-only.
 - Gemma 4 12B-it: κ 0.718 / 0.616 (DEV / TEST), but both error upper bounds are above 15% (false pass 0.195 / 0.341, false fail 0.249 / 0.265).
 - Qwen3.5-9B: coverage 0.40 / 0.41, below the 60% floor.
 - **Self-preference:** Gemma 4 12B-it passed its own wrong answers 8/13 (61.5%) vs 11.1% for other families’ answers, **+50 points** (90% CI +27..+69). It is disqualified on Gemma-family answers.
 - **Taglish naturalness passed** for Gemma 4 12B-it: weighted κ 0.784 (DEV) / 0.719 (TEST). Its 0–3 naturalness scores may be silver labels on non-Gemma answers only. Qwen3.5-9B failed (0.510 / 0.384).

H4: tokenizer efficiency affects cost and performance · In test

- **Source:** RESEARCH_THESIS.
- **Evidence so far:** the cost side is measured. PH languages need 1.8–2.0× the tokens of English on all six tokenizers (EXP-TOK-001).
- **Still open:** the link to downstream performance. That needs EXP-TOK-002 (vocabulary extension plus a short continued pretraining).

H-DEC-001 A: Gemma 4 wins on PH-EVAL · Supported

- **Source:** decision log DEC-001, pre-registered 04:20Z. Scope (175 items), judge policy and thinking-off were all fixed before any result.
- **Result (corrected 2026-09-30):** paired exact McNemar on 175 complete pairs each. gemma-4-12b vs qwen3.5-9b 36:8; vs sailor2-8b 79:1 ($p = 2.5e-5$ and $p < 1e-6$). Gemma is favoured in both pairs, so the hypothesis is supported. A scorer fix (a missed thought-marker variant) flipped two Gemma auto items from wrong to right; the originally reported counts were 36:10 ($p = 0.00016$) and 79:3 (RESULT-001). Corrected file: DEC-001-partA-corrected.txt.
- **Sensitivity (corrected):** without the second blind sheet, whose key was exposed during rating, it is still supported (16:0; 30:0). On auto-scored items alone, the gemma-vs-qwen pair is now significant (9:0, $p = 0.0039$; it was 9:2, $p = 0.065$, before the fix).
- **Truncation (post-hoc, not registered):** restricted to items where neither answer hit the 512-token cap, gemma vs qwen is 17:6 ($p = 0.035$) and gemma vs sailor2 32:1 ($p = 7.9e-9$), both still < 0.05 under Holm. Truncation does not explain the result (DEC-001-partA-truncation-sensitivity.txt).
- **Consequence:** with Part B also supported, **H-DEC-001 is supported**. DEC-001 still needs Gate F and Part C before it closes.

H2: explicit Taglish training beats translated data · Untested, motivated by FINDING-001

- **Source:** RESEARCH_THESIS.
- **Motivation:** Qwen2.5-7B-Instruct passes 0/60 native Taglish prompts, and every answer rated for naturalness scored $\leq 1/3$ (FINDING-001).
- **Needed to test it:** an SFT experiment comparing Taglish-response data with translated data.

H5: a PH-specific suite exposes failures that standard benchmarks miss · Untested, motivated by FINDING-001

- **Source:** RESEARCH_THESIS.
- **Needed to test it:** the same model’s standard-benchmark (e.g. FilBench) scores, set against PH-EVAL. The Taglish failure alone does not test it.

MAIN, H1, H3, H6 · Untested

- All need the training phases: PH-Corpus, continued pretraining, SFT, and domain derivation.

H-TOK-SEA: SEA pretraining brings better PH tokenization · Untested, counter-example found

- **Source:** exploratory, from EXP-TOK-001. Not pre-registered.
- **Observation:** Sailor2 (SEA-pretrained) has the least efficient tokenizer. It kept Qwen2.5’s vocabulary.

H-FRONTIER-001: frontier API models vs our best open candidate · Registered, awaiting results

- **Source:** decision log DEC-003, pre-registered as EXP-FRONTIER-001 on 2026-09-30T12:41Z, before any API call.
- **Primary:** each of the 3 frontier models vs Gemma 4 12B-it (re-run at 4,096 tokens), paired exact McNemar on all 175 items, Holm across the 3 pairs. “Better” means Holm-adjusted $p < 0.05$ AND more wins. Everything else is descriptive.
- **Comparability:** all five open models are re-run at max_new_tokens 4,096, because 30–37% of their 512-cap answers were at or above 505 tokens.
- **Status:** no result yet (DEC-003 Final Choice is pending).

H-DOM-001: Filipino CPT helps domain skill after SFT · Proposed, not yet registered

- **Source:** decision log DEC-004 (Option C, coding pilot) and the domain plan (PLAN-001).
- **Hypothesis:** light Filipino CPT improves domain (coding) performance after SFT, compared with SFT alone, especially on problems stated in Taglish.
- **Test:** ABL-CPT-001: SFT-only on the untrained Gemma 4 E4B vs CPT+SFT, same SFT data, same coding eval (unit tests on English sets plus a human-written Taglish set and new held-out problems).
- **Status:** registered only together with H-CPT-001, before the pilot runs.

H-CPT-001: ABL-CPT-001 success criteria · Proposed, not yet registered (DRAFT)

- **Source:** DEC-004. The criteria are a draft, to be registered in the decision log before the pilot, once the eval suite sizes are fixed and power is checked.
- **Draft criteria:**
 - C1 language: held-out Filipino perplexity down at least 10% relative vs untrained E4B.
 - C2 capability: CPT-E4B > untrained E4B on pooled FilBench fil MCF, paired exact McNemar $p < 0.05$.
 - C3 forgetting: English non-inferior, one-sided 95% lower bound of the difference above −2 points.
 - C4 domain: CPT+SFT vs SFT-only on the Taglish coding set (paired McNemar); English coding non-inferior within −3 points.

Notes

- **Power:** PH-EVAL v0.1 (175 items, 70 auto-scored) is probably underpowered for H-DEC-001 A. A NOT SUPPORTED there would mean “too few discordant items”, not “Gemma is worse”. Target for PH-EVAL v0.2: at least 500 auto-scorable items, including Filipino math, before the next model decision.

Change log

- **2026-10-01:** H-OPS-JUDGE gains the DEC-002 / JUDGE-002 result (open teachers: pass/fail silver not confirmed for either; Gemma self-preference +50 points; Taglish naturalness passed, 0.784 / 0.719). H-DEC-001 A carries the corrected Part A numbers (36:8, 79:1; S1 16:0; S2 9:0, $p = 0.0039$) and the post-hoc truncation check (17:6, $p = 0.035$); still supported. Added H-FRONTIER-001 (registered, DEC-003), and H-DOM-001 and H-CPT-001 (proposed, not yet registered).
- **2026-09-30:** second rater κ 0.913 (RESULT-001 limitation resolved). Judge confident-FAIL validation failed (6/40, upper bound 27.5%), so H-OPS-JUDGE is now refuted in BOTH directions on privacy and instruction following. H-DEC-001 A survives S3 (26:5, 61:2).
- **2026-09-29, later:** H-DEC-001 A → supported (RESULT-001). H-DEC-001 as a whole is supported.
- **2026-09-29:** tracker created.
 - H-DEC-001 B → supported.
 - H-OPS-JUDGE → refuted for PASS verdicts.
 - H4 → in test.
 - H2 and H5 marked as motivated by FINDING-001.