

PinoAI research digest · 2026-W40

Week of 2026-09-28 (the project room opened Mon 09-28; covers 09-28 to 09-29)

· Taglish edition: digest-2026-W40.tl.qmd

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

2026-09-29

1. What we learned

- **Philippine languages cost 1.8–2.0× the tokens of English** on every candidate tokenizer (fil, ceb, ilo, war; pag 1.45–1.54×). Gemma 4 is the most efficient, but only by 7–12%. See the note *EXP-TOK-001* (verified by an independent rerun).
- **A 7B instruct model fails every native Taglish prompt (0/60)**. It either drifts to English (16 answers confidently English), or code-switches unnaturally (every rated answer $\leq 1/3$). See the note *FINDING-001*.
- **Our LLM judge is trustworthy in one direction only**. It agreed with the native reviewer on 14/20 (κ 0.37). Its errors are mostly false PASSes, so its FAIL verdicts are accepted and its PASS verdicts go to a human.
- **The SEA model isn't the tokenizer winner**. Sailor2 kept Qwen2.5's vocabulary, and is the least efficient on PH text.

2. Hypothesis status changes

Hypothesis	Change	Evidence
H-DEC-001 B · Gemma 4 lowest Filipino tokens per word	pre-registered → supported	EXP-TOK-001
H4 · tokenizer efficiency affects cost and performance	untested → in test (cost side)	EXP-TOK-001
H-OPS-JUDGE · confident judge verdict $\approx$ native reviewer	assumed → refuted for PASS	calibration #1
H2, H5	untested → untested, motivated	FINDING-001

Full list: *HYPOTHESES*.

3. Experiments run, with compute

Experiment	Compute	Outcome
EXP-TOK-001 tokenizer efficiency	CPU only, 0 GPU h	done, verified
EXP-TOK-001 independent rerun	CPU only, 0 GPU h	PASS, 0 differences / 876
Baseline smoke: FilBench on Sailor2-8B, 5 attempts	28.9 kernel wall-minutes ($\approx$ 0.48 h)	all failed (gated dataset, loader script, then out-of-memory twice); fixes are in the runners
Baseline smoke: PH-EVAL on Sailor2-8B-Chat	~6 min	plumbing OK, 20/20 predictions
Qwen2.5-7B-Instruct judge-calibration runs	not logged	175/175 predictions

4. Literature and method radar

- **Xiaomi MiMo-V2.6 open RL release**: 7,780 agentic RL environments (Apache-2.0) plus a verl fork. Not usable yet (English/Chinese only; needs Docker and a GPU cluster). Kept for later, especially the cyber (ARVO) tasks and the environment format for PH-localized tasks.
- **Commercial LLM judge**: in use as a bounded-question judge, under conditional governance terms. Weak on Taglish naturalness; the top of its scale is untested.
- **Laya**: the official Apache-2.0 checkpoints are cleared as a base for a future in-house PH judge.
- **Tokenizer fairness literature**: our result matches the cross-lingual token-cost gaps reported by Petrov et al. (arXiv 2305.15425) and Rust et al. (arXiv 2012.15613).

- **Existing PH benchmarks:** the survey (36 benchmarks) found no public generative Taglish eval, so PH-EVAL's 60 native items fill a real gap.

5. Open questions for the project lead

1. **A second native reviewer (Filipino/Taglish):** needed for inter-annotator agreement before any PH-EVAL paper, and to validate the judge.
2. **Cebuano, Ilocano and Hiligaynon:** still no reviewers. Hiligaynon also has no text source for the tokenizer study.
3. **Commit the research files:** the commit time makes the pre-registration tamper-evident.
4. **Later:** a cyber/coding track (the MiMo ARVO tasks)? That is a scope decision, and needs a governance safety review first.