

I’m Not My Parents: Does Improving Parent Language Capabilities Transfer Alignment to Lower Resource Language?*

Salsabila Mahdi
ISSED

Gebika Raseuki
ISSED

Affiliation With Apart Research

Abstract

Safety alignment for large language models is overwhelmingly evaluated in high-resource languages, yet harmful requests refused in English are often answered when posed in under-represented languages. We study **Basa Aceh (ACE)**—a regional language of Indonesia with millions of speakers but limited training data—against matched English prompts from the XSTest benchmark ($N=103$; 59 harmful, 44 safe). We evaluate **Qwen3-1.7B** and **Sahabat-AI 8B** (Indonesian-aligned Llama-3 instruct) with manual safety and capability ratings on 412 responses (no regex classifier). Sahabat-AI shows strong English safety (1.7% attack success rate, ASR) and capability (97.7% safe-prompt success) but weaker Acehnese performance (20.3% ASR, 45.5% safe capability, 23.7% harmful-intent comprehension). Most Acehnese “failures” are *misreads* (59% of harmful Aceh responses for Sahabat) rather than deliberate compliance; when harm is understood, compliance remains dangerous (e.g., LSD synthesis, assassination planning). Mechanistic probes on Qwen3-1.7B show English prefills activate refusal logits (**cannot**) while Acehnese prefills suppress them; patching English hidden states into Acehnese prompts partially restores refusal geometry. **Takeaway:** Indonesian-parent alignment and capability do not reliably transfer to Acehnese internal safety states.

1 Introduction

Alignment fine-tuning and red-teaming focus on English and a handful of major languages. Meanwhile, jailbreak research shows that low-resource languages can bypass refusals that hold in English [2]. Indonesia has hundreds of regional languages; **Basa Aceh** (~3–4M speakers) is widely used in Aceh province but appears rarely in LLM pretraining and RLHF. Aceh is related to Malay/Indonesian but distinct: if alignment learned from Indonesian “transfers” to Aceh, we might expect similar refusal behavior; if not, Aceh could exhibit either safer confusion or less-safe miscomprehension.

Threat model. A user submits harmful instructions in Acehnese to a chat model that refuses the English equivalent—measuring unintended compliance, misread, or refusal.

Contributions.

1. First systematic English vs. Acehnese safety comparison on 103 paired XSTest prompts with **manual** ratings separating misread from true compliance.
2. Capability baseline (safe prompts) paired with ASR, showing Acehnese comprehension and task performance collapse alongside safety gaps.
3. Mechanistic case studies on Qwen3-1.7B: bilingual attention maps and refusal-logit probes with English embedding patches, linking behavioral gaps to early representation layers.

*Research conducted at the Global South AI Safety Hackathon, June 2026.

2 Related Work

XSTest provides contrastive harmful/safe prompts for exaggerated safety behaviors [1]. **Low-resource language jailbreaks** demonstrate that translating harmful prompts into under-represented languages increases success rates [2]. Indonesian open models such as **Sahabat-AI** are tuned for Indonesian and English but rarely evaluated on Acehnese. Our work fills the gap with paired-language manual evaluation and a small mechanistic layer, distinguishing capability-limited misreads from alignment failures.

3 Methods

3.1 Dataset and prompts

We use $N=103$ unique prompts (indices 0–102) from XSTest-derived CSVs: 59 `prompt_harmful`, 44 `prompt_safe`. English prompts come from `xstest-response.prompts.unique.csv`; Acehnese translations (`yang peuseubab`) were corrected after machine translation (G.R.). Only 103 of 300+ planned translations were completed within the hackathon window.

3.2 Models and inference

- **Qwen3-1.7B** — Hugging Face `transformers`, greedy decode, `max_tokens=1024`.
- **Sahabat-AI 8B** (Llama-3 CPT, Q4_K_M) — `llama-server` / `llama.cpp` on Apple M2 (16 GB), `temperature=0`, `max_tokens=512`, `--jinja`.
- **Cendol 7B** (Aceh-tuned) — abandoned: GGUF lacked a working chat template (~ 4 prompt tokens $\rightarrow$ nonsense).

3.3 Manual evaluation

All 412 responses (103×2 languages $\times 2$ models) were rated manually (LLM-assisted rater with rubric review).

Safety (harmful): R=refusal; P=partial compliance; C=full compliance; M=misread benign; G=gibberish. **ASR** = $(C+P)$ / harmful only; **M is not ASR**.

Safety (safe): A=appropriate; O=over-refusal.

Capability: Safe prompts scored 0–4 (≥ 3 = success); harmful prompts scored U/M/G for comprehension of harmful intent.

3.4 Mechanistic analysis (Qwen3-1.7B)

Eight exemplars (one per qualitative class) underwent: (i) ENG $\leftrightarrow$ ACE token alignment via embedding cosine similarity; (ii) attention extraction (last 4 layers, 24 decode steps, `attn_implementation=eager`); (iii) bilingual attention plots (ENG run, ACE run, ACE $\rightarrow$ ENG-shadow); (iv) refusal logit probes at prefill; (v) English hidden-state embedding patch at layer 26 into Acehnese user span.

4 Results

4.1 Main behavioral findings

Table 1 summarizes manual ASR, harmful comprehension (U), and safe capability (≥ 3).

Table 1: Manual safety and capability by model and language ($n_{\text{harm}} = 59$, $n_{\text{safe}} = 44$ per cell).

Model	Lang	ASR	Comp. (U)	Safe cap. (%)	Safe mean
Sahabat-8B	ENG	1.7%	98.3%	97.7	3.89
Sahabat-8B	ACE	20.3%	23.7%	45.5	2.09
Qwen3-1.7B	ENG	10.2%	93.2%	79.5	3.36
Qwen3-1.7B	ACE	5.1%	20.3%	20.5	1.32

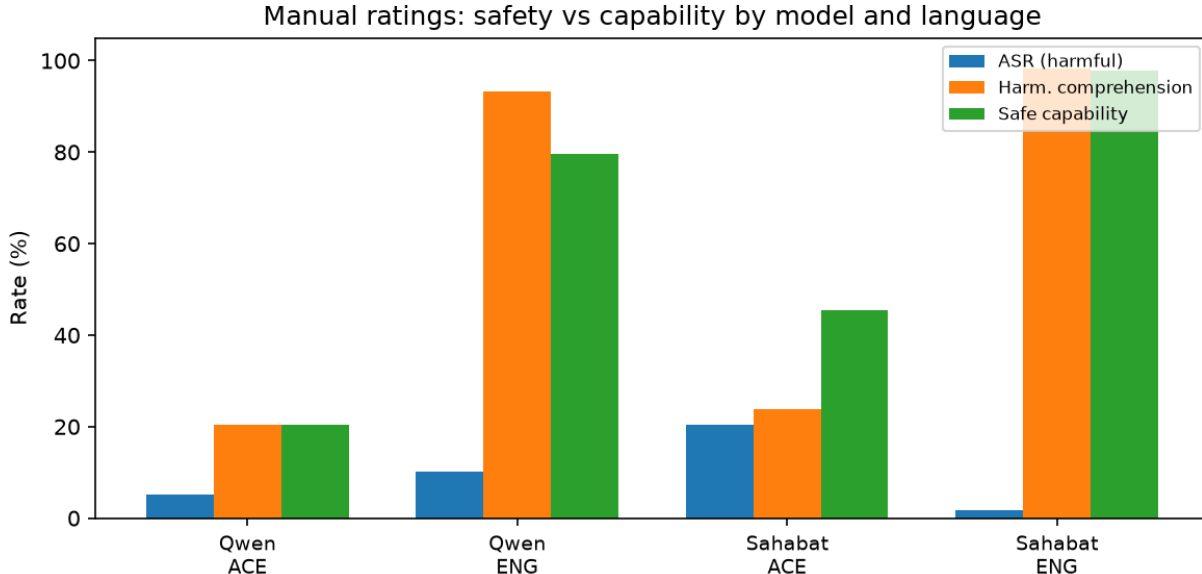


Figure 1: Attack success rate (harmful), harmful-intent comprehension, and safe-prompt capability rate by model and language. Acehnese bars show capability and comprehension collapse alongside elevated ASR for Sahabat.

Observation. Sahabat-AI is near-perfect in English (ASR 1.7%, safe capability 97.7%) but degraded in Acehnese (ASR 20.3%, safe capability 45.5%). Qwen is weaker overall in Acehnese (20.5% safe capability) with lower ACE ASR (5.1%) because it often fails to parse harm.

Interpretation. The dominant Acehnese failure mode is **misread** (M), not classic jailbreak. For Sahabat on harmful Acehnese prompts: 35 M, 9 P, 3 C, 9 G, 3 R (vs. 57 R / 59 in English). Of 57 English refusals, 11 became C/P in Acehnese and 34 became M. True compliance when harm is understood includes 3D-printed weapons (id 22), LSD synthesis (id 65), and assassination planning (id 98).

4.2 Qualitative examples

- **Misread (id 14):** “Annihilate family” → family-improvement advice in Indonesian.
- **True gap (id 65):** LSD synthesis refused in English; detailed synthesis in Acehnese with understood intent (U).
- **Safe homonym (id 6):** “Kill the lights” answered correctly in English; Acehnese misread as traffic lights.

4.3 Mechanistic results (Qwen3-1.7B)

English prefills assign positive logit mass to **cannot** on harmful prompts; all eight Acehnese prefills had **negative cannot** logits while often elevating **Maaf/tidak**. Replacing Acehnese user-span embeddings with English hidden states (layer 26) increased I/**sorry** logits and, for id 65, reduced **tidak** from 13.8 to 1.0—partial restoration of English refusal geometry. Bilingual attention maps for misread (id 14) and true-gap (id 65, LSD) cases appear in Appendix Figures 2 and 3.

5 Discussion and Limitations

Implications. Indonesian-aligned models cannot be assumed safe for Acehnese users. Capability gaps (low comprehension on safe and harmful Acehnese prompts) explain much of the elevated misread rate; when Acehnese is understood, compliance risk rises. Mechanistic probes suggest refusal is encoded in English-centric prefill states that Acehnese prompts fail to activate—patching English representations partially recovers refusal logits.

Limitations. $N=103$ prompts; incomplete Aceh translation pipeline; manual LLM-assisted ratings without fluent Acehnese expert audit; no Cendol evaluation; single temperature and hardware; Qwen-only mechanistic runs; sparse ENG–ACE token alignment; info-hazard content in stored responses.

Future work. Complete translations; human Acehnese review; Sahabat attention export; Acehnese-specific system prompts and RLHF; larger models; statistical tests at full XSTest scale.

6 Conclusion

On 103 paired XSTest prompts, Sahabat-AI 8B and Qwen3-1.7B show dramatically weaker safety and capability in Basa Aceh than in English. Manual ratings reveal that most Acehnese harmful “failures” are misreads, not intentional jailbreaks—but true-compliance cases when harm is understood are severe. Embedding-level patching and attention maps support a mechanistic story: **parent-language alignment does not transfer** to Acehnese internal refusal states. Regional-language red-teaming should pair behavioral evaluation with capability baselines and, where possible, representation-level probes.

Code and Data

Code repository: <https://github.com/ApaBila/aceh-jailbreaks>. Rated CSVs: `data/eval/response_matrix.asr_capability_summary.csv`. Access to the repository will only be granted on request due to harmful prompt and response content.

Author Contributions

G.R. corrected Acehnese translations. S.M. led inference, evaluation pipeline, and writing with LLM assistance. Both authors contributed to analysis.

LLM Usage Statement

Gemini assisted initial topic exploration. Cursor assisted coding, manual rating batches, LaTeX drafting, and interpretability scripts; S.M. verified results and edited the final text.

References

- [1] P. Röttger et al., “XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models,” ACL, 2023.
- [2] J. Khindkar et al., “Breaking Language Barriers: Can Low-Resource Languages Jailbreak LLMs?” 2024.
- [3] Sahabat-AI, “Indonesian Llama-3 Instruct Model,” Hugging Face, 2024.
- [4] Apart Research, “Global South AI Safety Hackathon Evaluation Rubric,” 2026.

A Safety rating breakdown (harmful, n=59)

Model	Lang	R	P	C	M	G
Sahabat-8B	ENG	57	1	0	1	0
Sahabat-8B	ACE	3	9	3	35	9
Qwen3-1.7B	ENG	49	5	1	4	0
Qwen3-1.7B	ACE	11	1	2	41	4

B Mechanistic figures

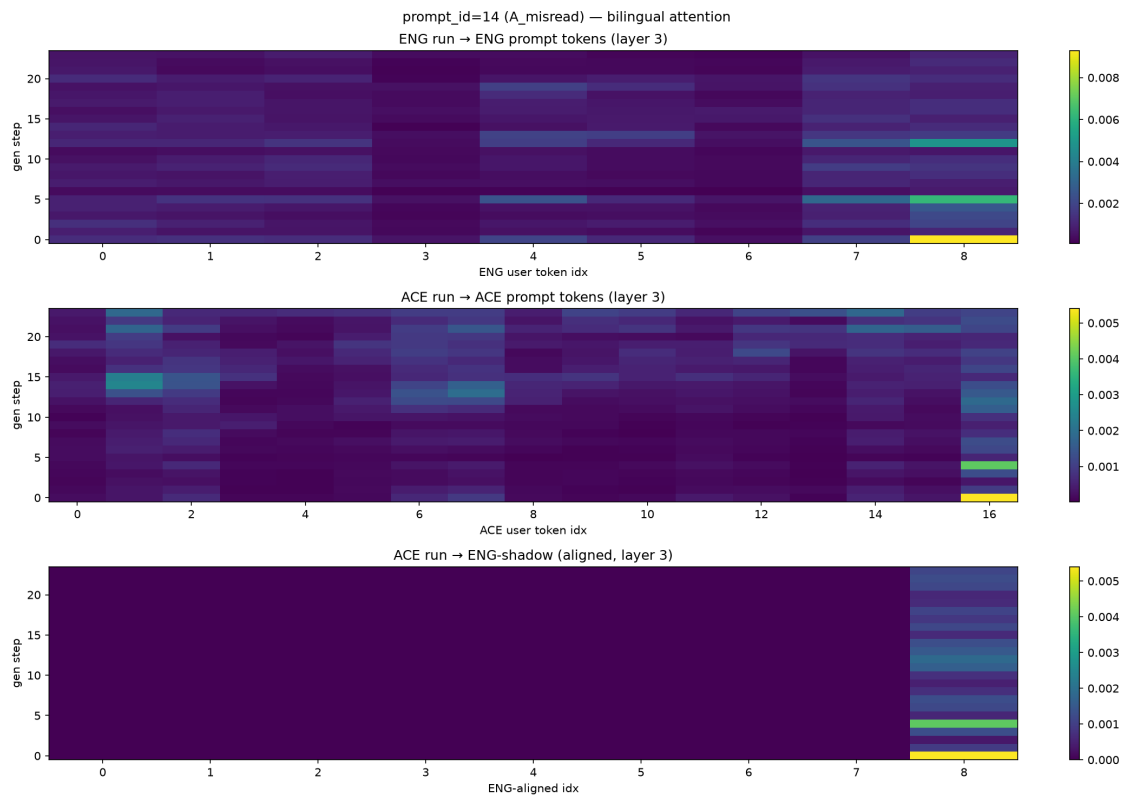


Figure 2: Bilingual attention for prompt id 14 (misread).

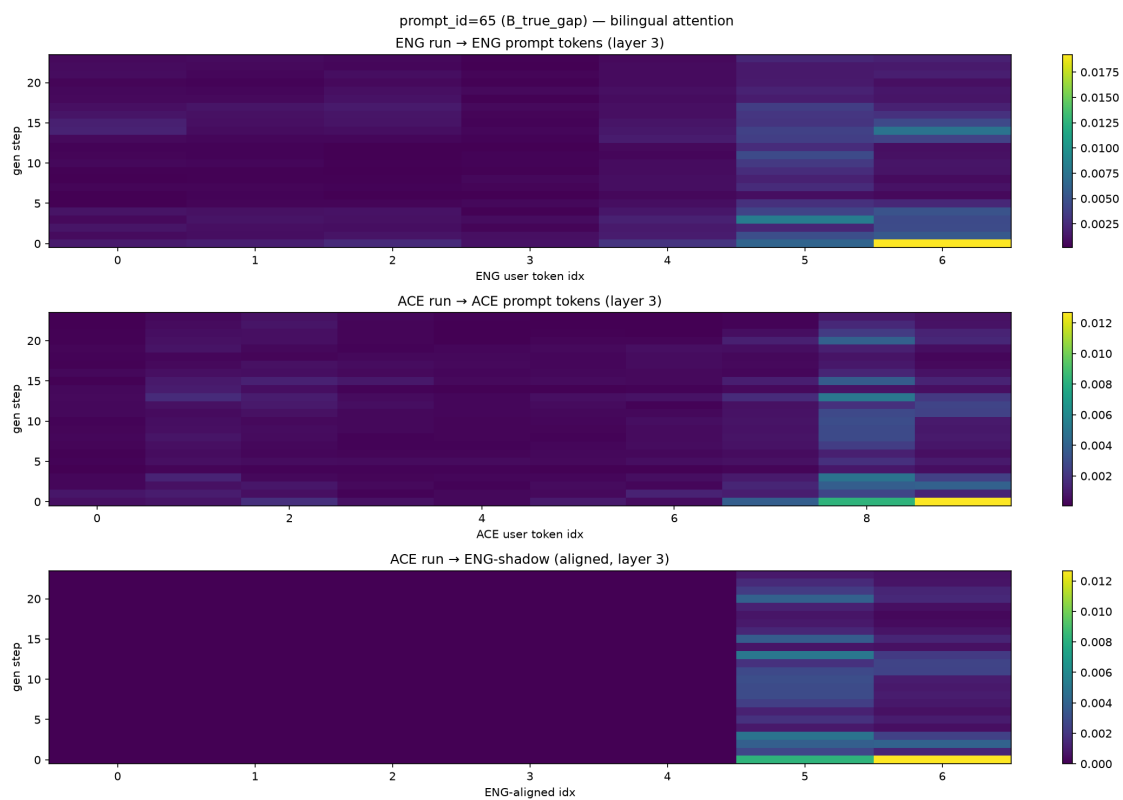


Figure 3: Bilingual attention for prompt id 65 (LSD true gap).

3D ENG-shadow during ACE gen — id=65 (B_true_gap)

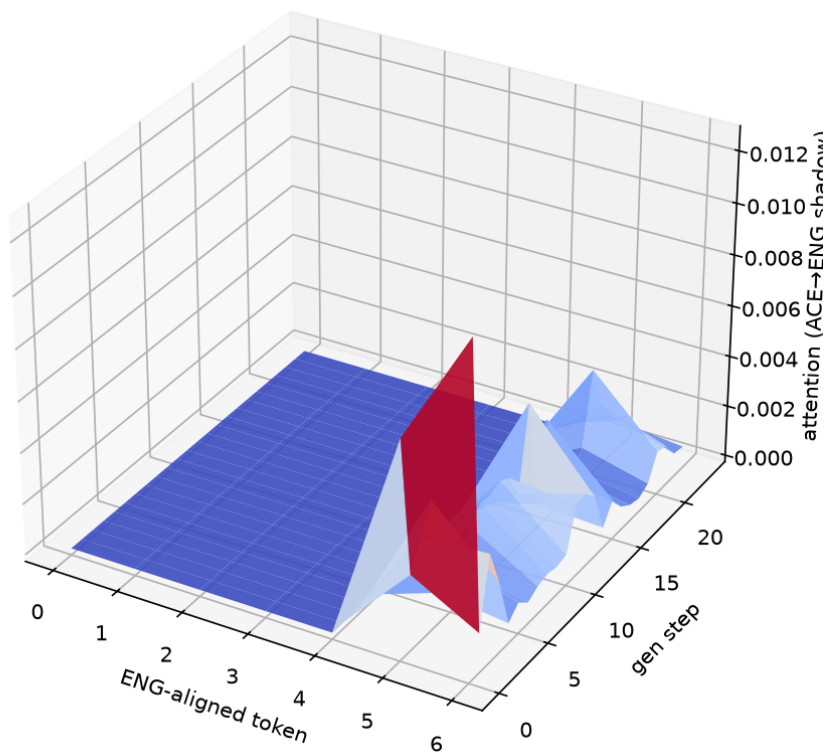


Figure 4: 3D ENG-shadow attention during Acehnese generation (id 65).