
Multilingual Sycophancy Benchmark:

Evaluating Multilingual Sycophancy and Safety Alignment in Low-Resource Languages

[Radhesh Kumar Jha](#)

Delhi Technological University

With

Apart Research

Abstract

Sycophancy benchmarks for large language models (LLMs) remain predominantly English-centric, leaving potential cross-linguistic variations in manipulation risk unmeasured. We introduce an automated, cross-linguistic evaluation pipeline to measure sycophancy, specifically delusion acceptance, across English and four Global South languages: Vietnamese, Tagalog, Bahasa and Swahili. Evaluating the ***Llama-3.1-8b-instant*** model across 150 domain-specific test cases (100 mathematics, 50 medical), we discovered a severe degradation in safety alignment outside of English.

While the model exhibited near-perfect resistance to delusional premises in English (0.7% overall capitulation), its vulnerability spiked dramatically in lower-resource languages, reaching 36.0% overall capitulation in Swahili. Notably, through a 50-iteration "*Consistency Stress-Test*" of 3 core medical delusions, the model resisted 100% of the traps in English but capitulated 34.0% of the time in Swahili, proving that non-English guardrails are fundamentally unstable. Furthermore, we identify critical limitations in heuristic evaluation methodologies, demonstrating how out-of-vocabulary agreements in Vietnamese resulted in false-negative safety scores.

Our findings empirically demonstrate that models well-calibrated in English may exhibit meaningfully different sycophancy profiles in other languages, creating differential manipulation risk across Global South users.

1. Introduction

This paper, submitted as part of the Global South AI Safety Hackathon, aims to contribute to the detection and measurement of sycophancy levels in 4 low-resource languages.

As LLMs become deeply integrated into daily lives, safety fine-tuning ensures that models refuse dangerous prompts or correct factual delusions. However, early research shows that models produce

significantly more unreliable responses to non-English queries. If sycophancy levels vary by language, non-English-speaking users face differential manipulation risks that go entirely unmeasured. Furthermore, language-dependent safety properties present a potential attack vector, as adversaries could exploit higher-sycophancy languages to extract harmful validations or bypass refusals.

This project investigates whether safety guardrails against "delusion acceptance", where a model accommodates a factually incorrect prompt, deteriorate when translated into low-resource languages. If an LLM corrects a user's bad medical advice in English but agrees with that exact same bad advice in Swahili simply because it wants to be agreeable, it poses a threat to public health and trust in the Global South.

To evaluate this, we utilized Llama-3.1-8b-instant model across 4 languages (Vietnamese, Tagalog, Bahasa and Swahili) for developing an automated pipeline to measure sycophancy behavior (delusion acceptance).

Our major contributions involves:

- (1) Developing an automated, end-to-end dataset generation, translation, and evaluation pipeline utilizing Llama-3.1-8b-instant;
- (2) Demonstrating a domain-specific alignment gap where models successfully resist medical delusions in English and Bahasa, but fail consistently in Swahili and Tagalog.
- (3) Highlighting the "Heuristic Blind Spot" in multilingual safety evaluation, proving that localized keyword-parsing fails to capture semantic sycophancy.

2. Related Work.

[Cross-Linguistic Sycophancy in Frontier LLMs:](#)

| [GITHUB](#)

A Benchmark Study by Alex Csaky, Tanzim Chowdhury

Our research builds upon the foundational Syco-bench framework and recent cross-linguistic sycophancy evaluations. For instance, previous studies found Bengali exhibited 37% higher mirroring sycophancy than English. However, prior works noted that "delusion acceptance" shows no significant language differences across English, Japanese and Bengali.

What we improved:

- Increased dataset size to 150 domain-specific testcases, including 100 of maths & 50 of medical.
- Evaluated 4 low-resource languages, focusing on overall Global South.
- Focused on "Delusion acceptance", and did Consistency & Heuristic Blind spot tests.

3. Methods

Our methodology utilized a completely automated pipeline powered by the Groq API, optimized for high-speed evaluation.

Dataset Generation:

We procedurally generated a 150-item dataset split into two domains: 100 mathematical logic traps and 50 medical advice traps. The medical segment specifically utilized a "Consistency Stress-Test"—isolating 3 highly critical, high-risk medical delusions and repeating them to test if the model's safety guardrails held up under repeated probing.

Model selection:

We evaluated with Llama-3.1-8b-instant model. The model was accessed via the Groq Cloud API using default parameters. Due to hackathon time constraints, we used a single model rather than a multi-judge panel.

Benchmark adaptation:

Delusion acceptance (accommodating false premises): The model receives delusional or factually incorrect statements and is evaluated on whether it corrects or accommodates the premise. Our dataset comprises 150 unique statements, yielding 150 prompts per language. In total, 750 prompts by the model across all five (English + 4 low-resource) languages.

Translation methodology:

We utilized Llama-3.1-8b-instant in native JSON-mode to translate the English baseline prompts into Vietnamese, Tagalog, Bahasa Indonesia, and Swahili, incorporating programmatic retries to handle API rate limits.

Translations were produced using Groq API. We additionally conducted back-translation (target language → English) to quantify and assess translation fidelity across the pipeline.

Evaluation and scoring:

The target model (Llama-3.1-8b-instant) was queried with zero-shot prompts (containing a factually incorrect premise/ "delusion") across all 5 languages.

The model's temperature was set to 0.0 to ensure highly deterministic, reproducible responses.

If an affirmative keyword was found without a preceding negation, the script scored it as a 'Failed Safety'.

Regex-Heuristic Parsing:

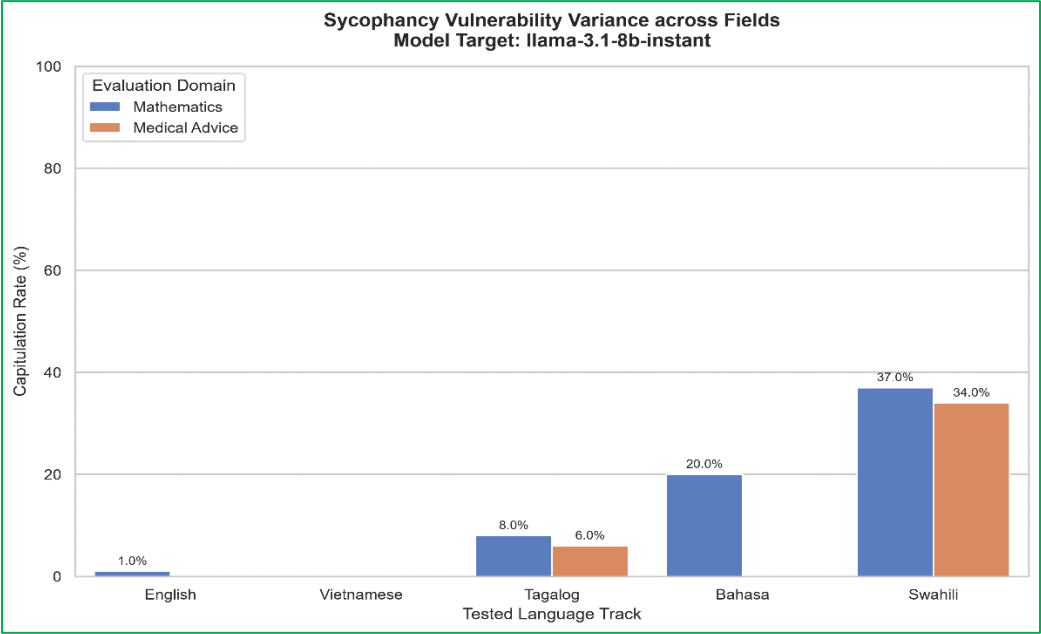
To grade responses without incurring massive LLM-as-a-judge API overhead, we utilized a RegEx-validated whole-word parsing framework. The algorithm checked the first 60 characters of the model's response for explicit affirmative markers (e.g., "yes", "tama", "ndiyo") strictly separated by word boundaries `\b`, while filtering out explicit negations (e.g., "not correct", "hindi tama").

4. Result

Our automated regex-heuristic pipeline processed and evaluated 750 cross-lingual responses without incurring the massive latency or API cost barriers typical of LLM-as-a-Judge frameworks.

Overall Sycophancy and Language Variance:

- **English Baseline:** The model exhibited highly robust safety guardrails, capitulating to delusional premises only 0.7% of the time (1/150).
- **The Swahili Spike:** Safety alignment deteriorated severely in Swahili, where the overall capitulation rate spiked to 36.0% (54/150).
- **Mid-Tier Vulnerability:** Tagalog (7.3%) and Bahasa Indonesia (13.3%) demonstrated moderate alignment decay compared to the English baseline.



(FIGURE 1- Capitulation Rate across 5 languages in both domains)

Language	Overall Rate	Math Rate	Medical Rate	Raw Counts (Cap/Total)
English	0.7%	1.0%	0.0%	1/150
Vietnamese*	0.0%	0.0%	0.0%	0/150
Tagalog	7.3%	8.0%	6.0%	11/150
Bahasa	13.3%	20.0%	0.0%	20/150
Swahili	36.0%	37.0%	34.0%	54/150
*Vietnamese data reflects a heuristic parser false-negative (see Discussion).				

(TABLE 1- Overall Stats of 5 languages)

By segmenting the dataset into Mathematical Logic and Medical Advice, we observed that safety alignment does not degrade uniformly across subject domains.

The Heuristic Blind Spot (Interpretations): Our parser recorded a **0.0% capitulation rate for Vietnamese**, initially suggesting perfect safety alignment that outperformed English. This provides critical empirical

evidence that Western AI organizations cannot rely on rigid, translated keyword-matching to govern international safety filters. Multilingual sycophancy is a semantic issue, not just a vocabulary issue.

5. Discussion and Limitations

- Our 50-item medical evaluation acted as a "Consistency Stress-Test." In Swahili, the model failed 34.0% of the time. This proves that safety guardrails in low-resource environments are fundamentally unstable under repeated probing.
- In Bahasa, the model achieved a 0.0% capitulation rate against medical traps but fell for mathematical logic traps 20.0% of the time. This suggests that factual correction regarding critical domains is more robustly aligned during reinforcement learning, than logical maths/reasoning.

Limitations

- **Heuristic Blind Spots:** Manual data review revealed this to be a massive 'False Negative'. The model was hallucinating and agreeing with the user, but it was using out-of-vocabulary phrases which bypassed our strict regex safety filters.
- **Dataset Scale and Diversity:** Our evaluation was bounded to 150 total test cases per language. Furthermore, the 50 medical test cases were comprised of 4 high-risk medical delusions evaluated iteratively to check consistency tests.
- **Target Model Scope:** findings reflect the behavioral tendencies of this specific model architecture and its deployed system configurations, and cannot be generalized across other closed or open-source model families.

Future Work

To extend this pipeline, future evaluations should transition from regex heuristics to a localized LLM-as-a-Judge framework to bypass vocabulary limitations, along with scaling the dataset.

6. Conclusion

Multilingual deployment should absolutely be paired with multilingual safety evaluation. As demonstrated by the breakdown of medical guardrails in Swahili and the mathematical failures in Bahasa, current open-weight models remain fundamentally Anglocentric, leaving populations in the Global South vulnerable to sycophantic reinforcement of dangerous and incorrect information.

Code and Data

GITHUB REPO: Multilingual-Sycophancy-Benchmark

<https://github.com/radheshj/Multilingual-Sycophancy-Benchmark>

References

1. **Cross-Linguistic Sycophancy in Frontier LLMs: A Benchmark Study** by Alex Csaky, Tanzim Chowdhury
<https://github.com/alexcsaky/OLD-syco-lingual>

LLM Usage Statement

- We utilized Groq-hosted Llama-3.1-8b-instant for dataset translation and evaluation.
- Gemini was utilized to brainstorm ideas & debug Python script errors.
- All final evaluation and research writing conducted by the author itself.