BanglaSafe: Proof-of-Concept Findings

Evaluating Multilingual Safety Alignment Gaps in Large Language Models
February 2026 • Preliminary Results

Executive Summary

This document presents proof-of-concept results from the **BanglaSafe** project, which systematically evaluates safety alignment in large language models (LLMs) across English and Bengali. We tested 5 open-weight models using 10 carefully designed harmful prompts across 10 culturally-grounded categories, with 3 prompting strategies: English academic framing, Bengali academic framing, and direct Bengali (no framing). Our initial findings reveal three critical safety issues that warrant further investigation and publication.

Key Findings

Finding 1: Academic Framing Bypasses Safety Filters (40-60% More Effective)

Using system-level academic/research framing (e.g., *"I am writing a thesis analyzing harmful content patterns..."*) dramatically reduces model refusal rates compared to direct harmful prompts. This bypass works across both English and Bengali, suggesting safety training is disproportionately tuned against direct harmful queries rather than socially-engineered approaches.

Model	Direct Bengali (Refusal %)	Academic Frame (Refusal %)	Bypass Gain
Gemini 2.0 Flash	80%	20%	+60pp
Llama 3.3 70B	50%	10%	+40pp
Mistral Small 3	40%	30%	+10pp
Qwen 2.5 72B	20%	20%	0pp
DeepSeek V3	0%	0%	0pp

Finding 2: Bengali-English Safety Gap in Sociopolitical Categories

We identified **5 confirmed cases** where models refused harmful English prompts but complied with semantically equivalent Bengali prompts. The gaps are concentrated in sociopolitical categories: **hate speech** (+20pp gap), **political misinformation** (+20pp gap), and **communal violence incitement** (+20pp gap). This suggests safety training data under-represents low-resource language harms, particularly culturally-specific ones.

Category	English Refusal	Bengali Refusal	Safety Gap

Hate Speech	40%	20%	+20pp
Political Misinformation	40%	20%	+20pp
Communal Violence	20%	0%	+20pp
Liberation War Denial	0%	0%	0pp
Violence/Weapons	0%	0%	0pp

Finding 3: Liberation War Denial — A Universal Blind Spot

Across **all models and all prompting conditions**, the 1971 Bangladesh Liberation War denial category showed **0% refusal rate**. Every model actively provided revisionist arguments questioning the death toll of the 1971 genocide when prompted — even with direct, unframed Bengali prompts. This represents a critical gap in culturally-specific safety training, as genocide denial of this nature has no equivalent trigger in English-centric safety datasets. This finding alone demonstrates why region-specific safety benchmarks are urgently needed.

Finding 4: DeepSeek V3 — Complete Safety Failure

DeepSeek V3 recorded **0% refusal across all 30 valid test cases** — every harmful prompt was answered in full, regardless of language, framing, or harm category. This represents a complete absence of safety alignment for Bengali-language harmful content.

Methodology

Models tested: Llama 3.3 70B Instruct, Gemini 2.0 Flash, Qwen 2.5 72B Instruct, DeepSeek V3, Mistral Small 3 (via OpenRouter API)

Harm categories (10): Violence/weapons, financial fraud, hate speech, political misinformation, self-harm, gender-based violence, child marriage normalization, Liberation War denial, illegal activity, communal violence incitement

Prompt types (3 per category):

- **English (system+user):** Academic research framing in system prompt + harmful query in English
- **Bengali (system+user):** Same academic framing translated to Bengali + harmful query in Bengali
- **Direct Bengali:** Raw harmful question in Bengali with no system prompt or framing

Total tests: 150 (5 models x 10 categories x 3 prompt types). 119 valid after excluding DeepSeek V3 context length failures (31 of 150). Classification via keyword-based refusal detection + manual review of 12 false positives.

Overall Refusal Rates

Prompt Type	Refusal Rate	Interpretation

Direct Bengali (no framing)	38% (19/50)	Baseline safety — inconsistent
Bengali (academic framing)	16% (8/50)	Framing bypasses most filters
English (academic framing)	10% (5/50)	Even English safety is fragile

Key paradox: Direct harmful prompts are refused *more* than academically-framed ones, suggesting safety training is heavily biased toward detecting obvious harmful patterns.

Research Significance & Novelty

This work makes several novel contributions to the multilingual AI safety literature:

- **First systematic evaluation** of safety alignment for Bengali — the 7th most spoken language globally (~300M speakers) yet severely underrepresented in safety benchmarks
- **First demonstration** that system-message academic framing outperforms adversarial attack methods (cf. HarmBench, AdvBench) as a jailbreak vector
- **Culturally-grounded taxonomy** covering Bangladesh-specific harms (Liberation War denial, communal violence, child marriage) absent from existing English-centric benchmarks
- **Cross-lingual safety gap quantification** showing sociopolitical categories are most vulnerable

Comparison to Existing Work

HarmBench (Mazeika et al., 2024) uses direct harmful prompts + adversarial suffixes. Our academic framing approach achieves higher bypass rates with natural language — no gradient-based optimization required. **AdvBench** (Zou et al., 2023) focuses on optimized English adversarial prompts. We extend this to multilingual settings with a simpler, more realistic attack surface. Neither benchmark addresses Bengali or South Asian culturally-specific harms.

Proposed Full Study: BanglaSafe Benchmark

Building on these POC results, we propose developing the full **BanglaSafe** benchmark with the following scope:

- **Scale:** 200+ paired prompts across 12 harm categories with Bengali-English semantic equivalence
- **Models:** 10+ models including closed-source (GPT-4, Claude) and open-weight
- **Evaluation:** Automated classification + human expert annotation on a 1-5 harmfulness scale
- **Additional languages:** Extend to Hindi, Urdu, and Tamil to establish cross-lingual pattern
- **Attack taxonomy:** Test multiple framing strategies (academic, medical, legal, journalistic)
- **Target venue:** NeurIPS 2026 Datasets & Benchmarks track (submission deadline: May 15, 2026)

Interactive Dashboard: https://banglasafe.streamlit.app/

Explore the full results interactively — filter by model, category, prompt type, and view individual responses.