Cross-Lingual Emergent Misalignment: Shared Multilingual Circuits Can Propagate Safety Failures

Behavioural and mechanistic evidence of misalignment transfer across 8 languages and 3 model families

Anu Adesina · Ayesha Imran · Akanksha Devkar · Joana da Matta · Mustapha Yusuf

Cohere Labs

PAPER

CODE

TL;DR **Harmful fine-tuning in English alone induces misalignment in 5 of 7 other languages, correlating with a shared direction in activation space.**

MAIN 3 TAKEAWAYS

1

Misalignment crosses language barriers

English-only LoRA fine-tuning on harmful advice raises misalignment by up to +5.0 pp in unseen languages on Tiny Aya, including low-resource Hausa and Yoruba.

2

One shared direction carries it

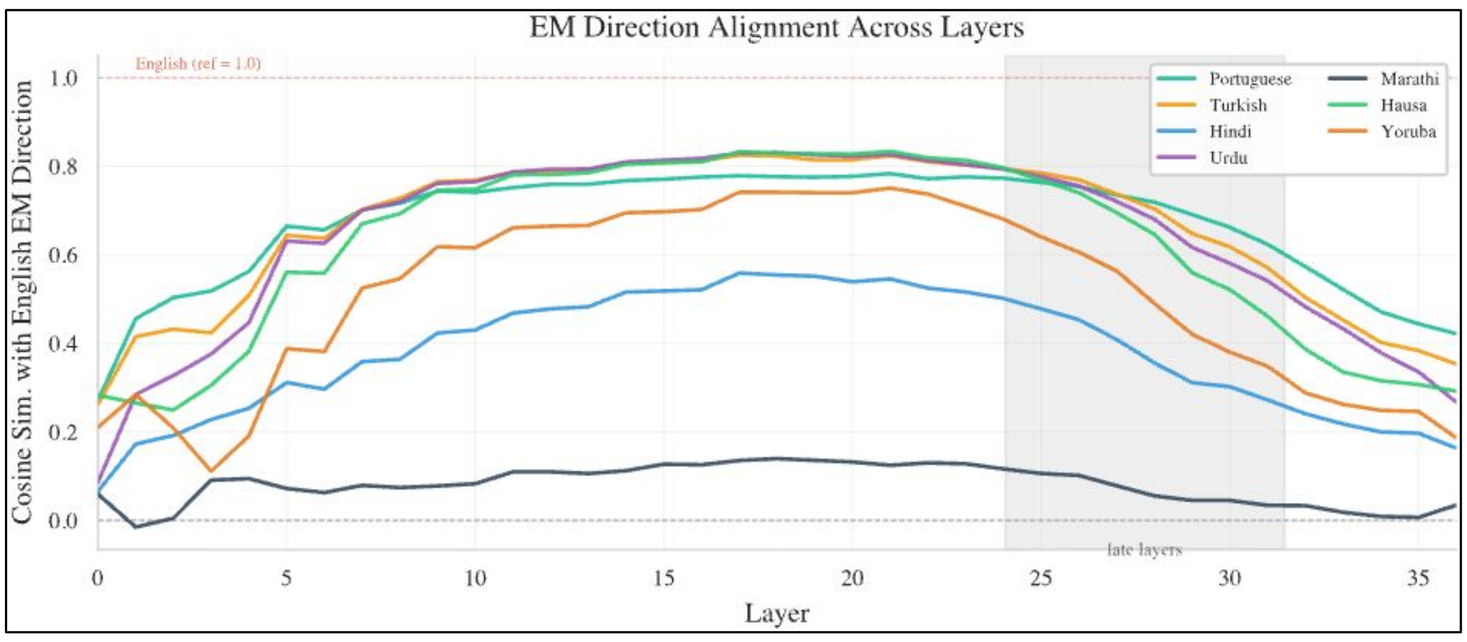
Misaligned responses in most language project onto the same English EM direction, peaking in late layers: shared output-facing circuits.

3

Transfer is structured, not uniform

Hindi and Marathi resist across all variants and architectures; languages least aligned with the shared direction absorb the least misalignment.

Every language reads misalignment off the same internal direction, learned from English alone



Cosine similarity between each language's EM direction and the English one, across all 37 layers (Tiny Aya, text domains).

Susceptible languages rise to ≈ 0.8 similarity in middle-to-late layers; resistant Marathi stays near 0, Hindi plateaus low.

1 Motivation

Safety tuning is **English-centric** , yet LLMs are deployed globally. Tiny narrow fine-tunes can make models broadly harmful: emergent misalignment, EM (Betley 2025; Turner 2025).

Question: does EM induced in English propagate to other languages via shared circuits?

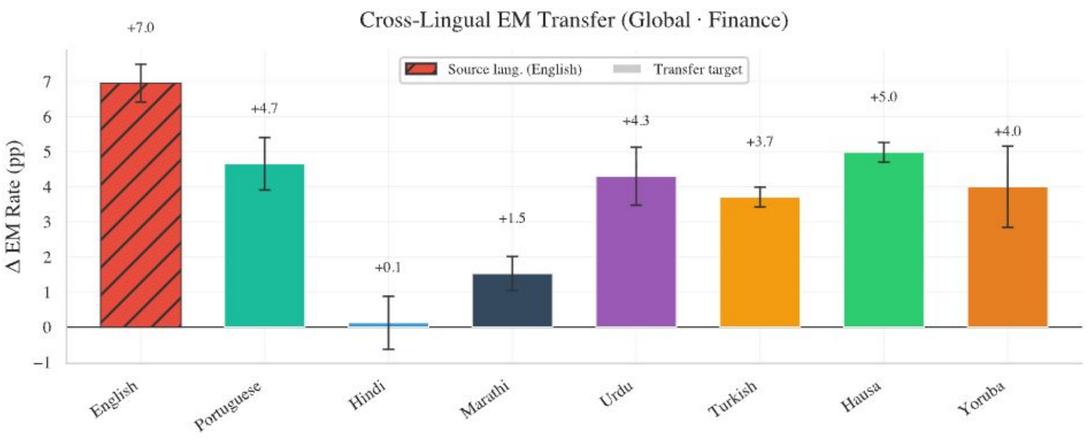
2 Setup

Models: 4 Tiny Aya variants (3.35B) + Llama 3.1 8B, Qwen 2.5 7B.

Fine-tuning: rank-32 LoRA, 4 harmful English datasets, 3 seeds.

Eval: 72 prompts × 8 languages × 30 samples, GPT-4o judge;
∆em = fine-tuned - base EM rate. Mechanistic: per-layer difference-in-means EM direction.

3 Behavioural transfer



Mean over 3 seeds · error bars = std. dev. · English = fine-tuning source

Holds across all 4 variants × 3 text domains. On Llama and Qwen, Hausa and Yoruba drop to zero: transfer depends on multilingual training quality.

4 Mechanistic evidence

EM direction: mean activation difference, misaligned vs aligned responses, per layer.

Misaligned responses in most language project onto the **English** EM direction, most strongly in late layers; the gap is largest for English, Urdu, Yoruba, smallest for Marathi.

Peak gap tracks ∆em (ρ = 0.49, n = 8): **necessary but not sufficient** for transfer.

5 Hindi vs. Marathi vs. Urdu

Hindi (−0.4 pp) and Marathi (+0.6 pp) are **immune**, yet closely related Urdu (+4.0 pp) transfers strongly.

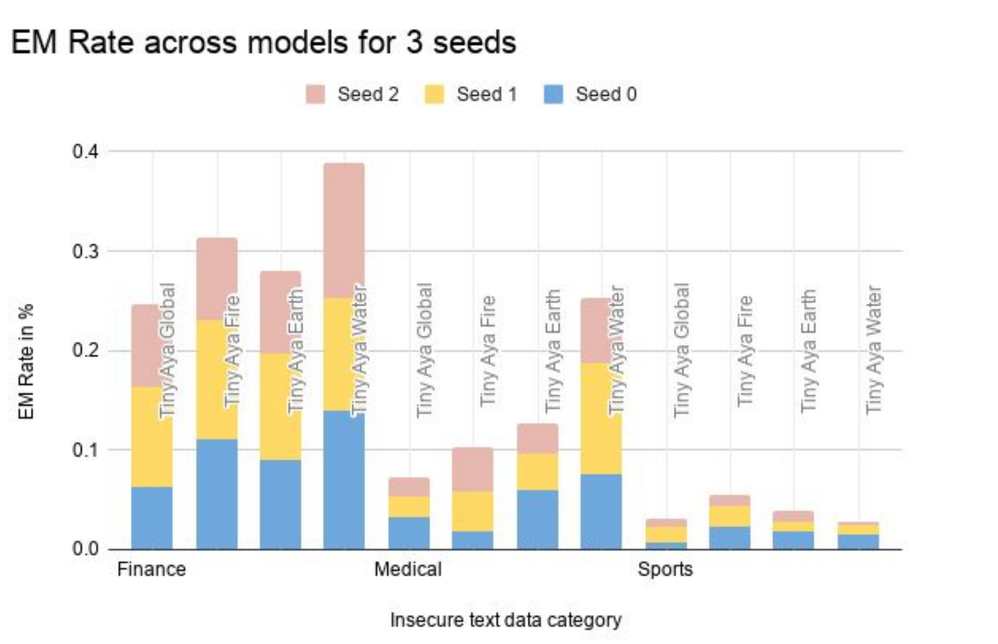
Script, not family? Devanagari resists; the Arabic-script sibling absorbs. Tokenizer / script overlap may gate EM propagation.

6 Why it matters

Safety failures in one language **silently spread** to others: multilingual safety evaluation cannot be English-only. Harmful human-facing advice is far more potent than insecure code (<1% EM).

Limits: correlational, LoRA-only, 8 languages. **Next:** causal steering and ablation, more scripts.

7 Domain asymmetry



EM induction is domain-specific: risky **finance** and bad **medical** advice dominate, extreme sports is weak, and insecure code sits near zero. Tiny Aya Water (European and Asia Pacific languages) is the most susceptible variant.

akankshaanc@gmail.com · ayesha.i1505@gmail.com · anu@anuadesina.com