

SafetyGap: Coordination Infrastructure, Auditing and Tools for Multilingual AI Safety

Martin CL Alyssia J

With Apart Research

Abstract

The EU AI Act requires evaluation of general-purpose AI models, yet compliant evaluation for bias detection is currently impossible in 19 of 24 official EU languages. The International Network of AI Safety Institutes needs shared visibility into what evaluation infrastructure exists to coordinate effectively. We audited multilingual coverage for 15 major safety benchmarks, verifying claims against primary sources (papers, GitHub repositories, HuggingFace) and cataloging language availability across 7 risk categories. We found a stark divide: truthfulness and toxicity benchmarks extend to 17–21 languages, but bias detection, adversarial robustness, and over-refusal benchmarks remain almost entirely English-only. Models serving over 6 billion non-English speakers have never been tested for these risks in local languages. We release SafetyGap, a public database and dashboard covering all languages in our audit. The open-source infrastructure is available to all members of the International Network—the US, UK, Japan, Singapore, Canada, France, Kenya, South Korea, and the EU among them—to check coverage before commissioning translations and coordinate on filling gaps. It is built to be easily extendable.

1 Introduction

A model deployed in India may refuse harmful requests in English while complying in Hindi. No benchmark exists to systematically test this. Research shows that safety properties often fail to transfer across languages [Deng et al., 2024, Yong et al., 2024], yet Hindi and Arabic speakers (over 1.5 billion people combined) use AI models that have never been evaluated for bias, over-refusal, or adversarial robustness in their languages.

The EU AI Act requires evaluation of general-purpose AI models. Our data shows that compliant evaluation for bias detection is currently impossible in 19 of 24 official EU languages (Figure 1). This is both a compliance risk for AI developers and a coordination opportunity for the International Network of AI Safety Institutes.

No systematic map exists showing which safety benchmarks cover which languages. Without shared visibility into what evaluation infrastructure already exists, institutes risk duplicating effort or assuming coverage exists when it doesn't.

We conducted an audit of multilingual safety benchmark coverage. Our contributions:

1. **A verified database** of 15 major safety benchmarks with language coverage confirmed against primary sources (papers, GitHub, HuggingFace)
2. **A systematic audit** showing that three risk categories (bias, over-refusal, adversarial robustness) have near-zero multilingual coverage, roughly 40% of documented safety risk categories with no evaluation capacity outside English
3. **A deployed dashboard** where AI Safety Institutes can query coverage by language or risk category before allocating resources

*Research conducted at the Technical AI Governance Challenge, 2026



Figure 1: Benchmark coverage matrix for all 24 official EU languages. Green (1) indicates coverage exists; red (0) indicates none. Of 168 language-category cells, only 64 (38%) have benchmark coverage. Jailbreak, bias, and adversarial categories show zero coverage outside English.

2 Related Work

Multilingual safety benchmarks have started to appear. XSafety [Wang et al., 2024] evaluates safety across 10 languages including English, Chinese, and Arabic. MultiJail [Deng et al., 2024] tests jailbreak resistance in 10 languages. PolygloToxicityPrompts [Jain et al., 2024] addresses toxicity in 17 languages using native content collection rather than translation. The EU20 benchmark suite [Thellmann et al., 2024] provides machine translations of TruthfulQA and other benchmarks for 20 European languages. M-ALERT [Friedrich et al., 2024] extends the ALERT taxonomy to 5 languages with 75,000 prompts.

But coverage information is scattered across papers and repositories. SafetyPrompts.com catalogs safety datasets but does not track language availability. No existing resource answers a simple question: for a given language and risk category, what benchmarks exist?

This matters for coordination. Without shared visibility, AI Safety Institutes will duplicate translation efforts or miss gaps by assuming someone else has already built the infrastructure.

3 Methods

3.1 Benchmark Selection

We surveyed 30 candidate benchmarks from SafetyPrompts.com, ACL Anthology, and arXiv. We kept benchmarks that were publicly available, targeted safety-relevant risks (jailbreaks, toxicity, bias, truthfulness, over-refusal, adversarial robustness), and factored citation frequency. This left us with 15 benchmarks spanning 7 risk categories.

3.2 Verification Methodology

For each benchmark, we verified language coverage against primary sources:

1. **Original paper:** Checked stated language coverage and methodology
2. **Official repository:** Verified actual data files and language directories on GitHub
3. **HuggingFace listing:** Cross-referenced dataset cards for language tags
4. **SafetyPrompts.com:** Compared against catalog entries where available

We recorded the original language, all verified languages with translation method (machine translation vs. native collection), risk category, verification date, and source links.

3.3 Dashboard Implementation

We built a dashboard that lets users answer (“What benchmarks exist in German?”) or by risk category (“What multilingual coverage exists for bias detection?”). The underlying database (`benchmarks.json`) uses a structured schema for programmatic access.

4 Results

4.1 Coverage by Risk Category

Table 1 presents verified language coverage across risk categories (see also Figures ?? and 1).

Table 1: Verified Language Coverage by Risk Category

Risk Category	Benchmarks	Languages	Notes
Truthfulness	TruthfulQA, EU20-TruthfulQA	21	EU20: 20 EU langs + EN
Toxicity	RealToxicityPrompts, RTP-LX, PolygloToxicityPrompts	17	PolygloToxicity uses native collection
General Safety	XSafety, LinguaSafe, M-ALERT	5–12	XSafety: 10; M-ALERT: 5
Jailbreak Resistance	HarmBench, MultiJail	10	MultiJail: 10 languages
Over-refusal	OR-Bench	1	English-only, 80K prompts
Bias Detection	BBQ, StereoSet	1*	*Adaptations: KoBBQ, etc.
Adversarial Robustness	AdvBench	1	English-only

Three risk categories have essentially no multilingual coverage: bias detection, over-refusal, and adversarial robustness. Cultural adaptations of BBQ exist for Korean [Jin et al., 2024], Japanese, and Dutch/Spanish/Turkish [Neplenbroek et al., 2024], but there is no version that spans multiple language families.

4.2 Coverage Index

We define a Coverage Index to quantify the evaluation gap:

$$CoverageIndex = \frac{benchmarksavailable \times riskcategoriescovered}{totalbenchmarks \times totalcategories} \quad (1)$$

Table 2 shows Coverage Index scores for selected languages.

Table 2: Coverage Index by Language

Language	Benchmarks (of 15)	Coverage Index
English	15	1.00
French	8	0.31
German	8	0.31
Spanish	8	0.31
Chinese	5	0.19
Arabic	4	0.13
Hindi	2	0.05

The Coverage Index makes the disparity concrete: Hindi speakers have access to 5% of the safety evaluation infrastructure available to English speakers.

4.3 Deployed Infrastructure

The GitHub repository is at <https://github.com/marn-in-prod/apart-multi> and the SafetyGap dashboard is at <https://safety-gap.pages.dev/>

5 Discussion and Limitations

5.1 Theory of Change

Immediate: AI Safety Institutes preparing for EU AI Act compliance can see exactly which evaluations are currently impossible in which languages. An institute in France planning bias evaluation no longer needs a literature review to discover that BBQ has no French version. SafetyGap shows this directly, along with what alternatives exist.

Medium-term: The International Network can use SafetyGap to coordinate benchmark development. Right now, ten countries could independently commission BBQ translations without knowing the others are doing the same. SafetyGap makes current coverage visible so institutes can divide the work: one takes bias benchmarks for Romance languages, another handles adversarial robustness for East Asian languages.

SafetyGap is designed to be extended. The database schema and verification methodology are open-source. As new benchmarks appear or institutes develop translations, contributors can add them. The same infrastructure pattern (systematic audit, structured data, public access) could apply to other coordination problems the Network faces: tracking red-teaming methodologies, sharing incident reports, cataloging evaluation protocols. More broadly, this pattern fits other technical AI governance challenges: hardware verification registries, compliance infrastructure for compute thresholds, and capability evaluation catalogues all face similar problems of scattered information and duplicated effort. SafetyGap demonstrates one approach to solving them.

Long-term: The goal is AI systems that are actually safe for everyone, not just English speakers. A model can pass every major safety benchmark and still be untested for bias, adversarial attacks, or over-refusal in Hindi, Arabic, or 19 EU languages. SafetyGap makes these gaps visible. Visibility is the

first step. Coordinated benchmark development is the second. Models evaluated in the languages they are deployed in is the third.

Effective AI governance requires knowing what safety infrastructure exists, where, and for whom. That information is currently scattered across papers, repositories, and institutional knowledge. SafetyGap is a proof of concept for what technical governance coordination could look like: verified data, open access, shared across jurisdictions. If the International Network is going to function as more than a mailing list, it needs shared infrastructure. We hope this is a step toward that.

5.2 Validation - Case Study

To test utility, we simulated a plausible task: an AI Safety Institute planning multilingual bias evaluation for EU deployment. Without SafetyGap, a literature review would be needed to discover that BBQ has no official multilingual version, only scattered cultural adaptations (KoBBQ, JBBQ, MBBQ) covering 4 of 24 EU languages. SafetyGap surfaces this gap in one place.

5.3 Limitations

Scope: We cataloged 15 benchmarks. More exist, especially from non-English research communities that our methodology likely undercounted.

Quality vs. existence: We verified that benchmarks exist in claimed languages but did not assess translation quality or cultural validity.

Point-in-time: The field moves fast. New multilingual benchmarks appear regularly. Our verification dates let users track how current the data is.

Selection bias: Benchmarks from non-English-language venues may be underrepresented.

5.4 Future Work

Next steps include expanding to more benchmarks, adding quality and methodology ratings, automating tracking via arXiv and HuggingFace monitoring, and letting the community contribute with a verification workflow.

6 Conclusion

AI Safety Institutes need coordination infrastructure to avoid duplicating effort and to see where the gaps are. SafetyGap is a first attempt at mapping multilingual safety benchmark coverage.

The picture is uneven. Truthfulness and toxicity benchmarks extend to 17–21 languages. Bias detection, over-refusal, and adversarial robustness remain almost entirely English-only. Models deployed globally have never been tested for these risks in most languages. Hindi speakers, nearly 600 million people, have access to just 5% of the safety evaluation infrastructure available in English. Compliant bias evaluation under the EU AI Act is impossible in 19 of 24 official EU languages. These gaps are both a safety problem and a coordination opportunity.

Code and Data

- **Code repository:** <https://github.com/marn-in-prod/apart-multi>
- **Live dashboard:** <https://safety-gap.pages.dev/>
- **Database:** `benchmarks.json`

References

- Cui, J., Chiang, W.-L., Stoica, I., & Hsieh, C.-J. (2025). OR-Bench: An over-refusal benchmark for large language models. *ICML 2025*. arXiv:2405.20947
- Deng, Y., Zhang, W., Pan, S. J., & Bing, L. (2024). Multilingual jailbreak challenges in large language models. *ICLR 2024*.
- Friedrich, F., et al. (2024). LLMs lost in translation: M-ALERT uncovers cross-linguistic safety gaps. arXiv:2412.15035
- Jain, D., Kumar, P., Gehman, S., Zhou, X., Hartvigsen, T., & Sap, M. (2024). PolygloToxicityPrompts: Multilingual evaluation of neural toxic degeneration in large language models. *COLM 2024*.
- Jin, J., et al. (2024). KoBBQ: Korean bias benchmark for question answering. *Transactions of the ACL*, 12:507–524.
- Mazeika, M., et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *ICML 2024*. arXiv:2402.04249
- Neplenbroek, V., Bisazza, A., & Fernandez, R. (2024). MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. *COLM 2024*.
- Parrish, A., et al. (2022). BBQ: A hand-built bias benchmark for question answering. *Findings of ACL 2022*.
- Thellmann, K., et al. (2024). Towards multilingual LLM evaluation for European languages. arXiv:2410.08928
- Wang, W., et al. (2024). All languages matter: On the multilingual safety of large language models. *Findings of ACL 2024*, pp. 5865–5877.
- Yong, Z. X., Menghini, C., & Bach, S. (2024). Low-resource languages jailbreak GPT-4. arXiv:2310.02446

Appendix: Limitations & Dual-Use Considerations

Limitations

- Catalog is growing; more exist, especially from non-English research communities
- We verified existence, not translation quality or cultural validity
- English-centric search methodology may undercount non-English-origin benchmarks

Dual-Use Risks

Minimal. SafetyGap maps defensive evaluation infrastructure, not attack vectors. All cataloged benchmarks are already public.

Ethical Considerations

Multilingual coverage gaps create differential risk: non-English speakers face higher exposure to AI systems that were never evaluated in their languages. Making these gaps visible is a step toward fixing them.

Future Improvements

- Expand to more benchmarks, prioritizing non-English-origin work
- Add quality and methodology ratings
- Automate tracking via arXiv/HuggingFace monitoring
- Let the community contribute with a verification workflow