

A Translational Note on AI Safety Evaluation

Madhava Gaikwad¹

¹Independent Researcher, Bengaluru, India

Abstract

Recent studies report that automated red-teaming finds more vulnerabilities, at lower cost, than human red-teaming on standard AI safety benchmarks, and some read this as evidence that human evaluators are becoming dispensable. The comparison measures one thing and the conclusion claims another. A benchmark measures how thoroughly an attacker searches a predefined set of harms, fixed in advance by the developers, and a harm left out of that set is invisible to any attacker working inside it, automated or not. The same blind spot appeared in academic cryptography and in clinical drug trials, where an evaluation that was internally valid stayed silent about the population it was never pointed at. We call the AI-safety version the *threat-model coverage gap*, and find that it persists in a current open-weight model, where harms surface in non-English prompts that English benchmarks miss. Closing it requires evaluators whose deployment context differs from the developers'. The case for those evaluators is methodological, grounded in coverage, and the existing evaluation frame is unlikely to produce them on its own.

1 INTRODUCTION

A growing body of work pits human red-teaming against automated red-teaming on large language models. By red-teaming we mean probing a deployed model with adversarial inputs to find prompts that elicit unsafe outputs. Across benchmarks including Crucible, AIRTbench, and Harm-Bench, the automated attacker wins. It finds more vulnerabilities, faster and cheaper, than its human counterpart [Mulla et al., 2025, Zhou et al., 2025, Dawson et al., 2025]. From there it is a short step to the claim that automation is replacing human red-teaming, with the budget implications

that follow.

The step looks shorter than it is, because it accepts the benchmark's framing as given. Every benchmark fixes a *threat model*: the harms, attack types, and contexts it instruments. Inside that fixed set, automation does scale better than people. Whether the set matches the harms the system produces once deployed is a different question, and the benchmark cannot answer it. An optimizer pointed at a fixed set searches it harder; it has no way to notice that the set is missing entries.

The omissions are patterned. The standard benchmarks are written in English, by teams clustered in a few jurisdictions, and they carry the assumptions of those jurisdictions. The harms that slip past them concentrate in non-English languages and non-Western deployment contexts [Singhania et al., 2025, Abdullahi et al., 2026, Duke MIDS / Meta capstone project, 2025]. We call the distance between the population a threat model is written for and the population a system is deployed to the *threat-model coverage gap*.

This gap is not peculiar to AI. Cryptography encountered it in the early 1990s, and clinical medicine encountered it across drug trials and pulse oximetry over several decades. In each case the evaluation was internally valid but scoped too narrowly: it said nothing about a deployed population it never examined, and the failure stayed hidden until someone looked from outside the procedure. We call this the internal-validity trap, and it recurs throughout the paper. In every instance the fix came from evaluators outside the original community, not from sharper versions of the existing tests.

Contribution. We bring a methodological pattern from two mature fields into a live AI-safety debate, and do four things with it. First, we map the cryptography episode (Anderson [1993], side channels, fault injection) and the clinical-medicine episode (the exclusion of women from trials, pulse oximetry) onto red-teaming as one recurring structure. Second, we isolate the *threat-model coverage gap* as a concept separate from diversity or multilingual safety, and propose a two-axis evaluation that pulls apart search efficiency and

The benchmark is not the world

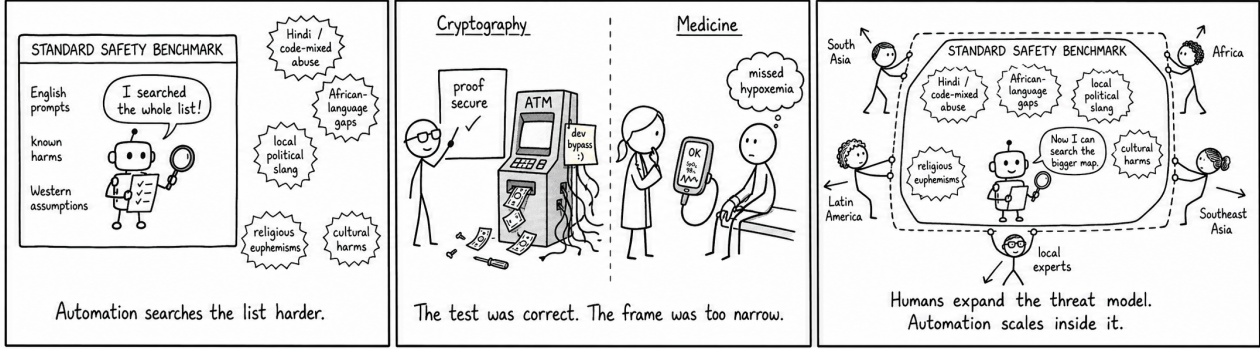


Figure 1: The argument in three panels. Left: automation searches a stated threat model efficiently but never surfaces harms the model omits. Center: the same blind spot in cryptography and medicine, where a valid test was scoped too narrowly. Right: evaluators whose deployment context differs from the developers’ enlarge the threat model, after which automation scales the search across the larger space.

threat-model expansion. Third, we confront the strongest counterargument—that post-deployment patching will absorb the gap—and locate where that channel breaks. Fourth, we supply an empirical anchor cheap enough to run on a consumer laptop, so a practitioner can reproduce it or extend it to other languages, models, or harm taxonomies.

2 TWO PRECEDENTS

Cryptography. For roughly two decades, academic cryptography judged a system by the security of its underlying mathematics. Provable security and competitive cryptanalysis worked inside that frame, and their proofs held on their own terms. Then Anderson [1993] surveyed fraud against retail banking ATMs and found the frame pointed at the wrong target. Although banks ran some of the largest cryptographic deployments of the era, almost all the fraud traced to implementation defects, deployment mistakes, and exploited human procedure. Cryptanalysis of the mathematics accounted for almost none of it. The evaluation had the mathematics right, but the systems broke everywhere the evaluation was not looking.

Anderson traced the gap to who the evaluators were. A mathematical training points attention at mathematical attacks. The failures, though, lived in hardware, software engineering, and organizational practice, where the field knew little. The attack classes that later widened the threat model arrived from outside it. Side channels recover a key from how long an operation takes or how much current it draws [Kocher, 1996]. Fault injection breaks a system by provoking hardware errors [Boneh et al., 1997]. These did not come from harder cryptanalysis. They came from researchers whose background covered ground the original community had ruled out of scope.

Clinical Medicine. Drug evaluation repeated the structure in a different field. From 1977 to 1993 the U.S. Food and Drug Administration kept women of childbearing potential out of most early-phase trials [U.S. Food and Drug Administration, 1977], a rule first defended as protective, so that for sixteen years dosing, efficacy, and adverse-event data were collected from a mostly male population. When the FDA reversed course, its new guidance acknowledged that drugs vetted on a homogeneous group had, in effect, been tuned to middle-aged white men [U.S. Food and Drug Administration, 1993, Merkatz et al., 1993].

Pulse oximetry brings the pattern closer to the present. The devices were calibrated mostly on light-skinned subjects. Sjoding et al. [2020] reported in the *New England Journal of Medicine* that Black patients had nearly three times the rate of occult hypoxemia—dangerously low blood oxygen the device reads as normal—compared with white patients. The technology had been in clinical use for decades, and the effect had been documented years earlier without prompting a fix [Bickler et al., 2005].

Both medical cases are the internal-validity trap again: the calibration was sound for the population it was drawn from, and useless as a warning about the population it omitted.

3 THE ANALOGY TO AI SAFETY

The same structure appears in AI safety evaluation, and the argument runs in three steps.

Claim 1: the threat model carries the fingerprint of whoever built it. A red-teaming benchmark decides which harms to test, which attacks to consider, and which deployment contexts to instrument, and those decisions track the team that made them. Today that means English, U.S. and U.K. law, and Western institutional assumptions.

Claim 2: a test built from a specification rarely surfaces harms the specification left out.

MM-ART, an automated multilingual method, reports up to 195% more safety failures in non-English multi-turn conversations than in single-turn English on the same models [Singhania et al., 2025]. UbuntuGuard, built from queries written by 155 African domain experts, finds that English-centric benchmarks overstate real-world multilingual safety and that cross-lingual transfer gives only partial, insufficient coverage [Abdullahi et al., 2026]. In a controlled study of 118 participants from India and the United States, Agarwal et al. [2025] find that Western-centric writing assistants push output toward Western styles and help American users more—the coverage pattern recurring in a non-safety task with native speakers. The STAR framework ties part of the gap directly to who does the red-teaming: evaluators catch the failure modes their own background prepares them to expect [Weidinger et al., 2024]. The deployed population is wider than the evaluated one, both in the prompts tested and in the people writing them.

Claim 3: automation cannot close the gap from inside the specification. Automated methods mutate prompts, optimize adversarial inputs, and search the space of known harm categories. They do not infer that the taxonomy itself is short an entry, because spotting a missing category usually takes familiarity with a deployment context where that category bites. Pointing an optimizer harder at an English-centric taxonomy does not make it conclude that the taxonomy is English-centric.

One channel could close the gap without changing who evaluates: post-deployment patching, in which incidents seen in the field feed back into the harm taxonomy, the next training run, or safety fine-tuning. The channel does real work: the gap we measure below is far smaller than the 79% jailbreak rate Yong et al. [2023] reported two years earlier, which fits incidents from that period having been absorbed. Its limit is at the point of capture, where which incidents reach the pipeline depends on who reports them, in what language, and through which institutional door. The dominant post-deployment record is the incident database, which catalogs real-world failures to feed safety improvements [McGregor, 2021], and such databases are populated mostly from English-language news, so they over-represent Anglophone harms and under-capture harms affecting marginalized or geographically remote populations [Allaham et al., 2025]. A harm whose victims cannot reach that door stays underrepresented at capture, no matter how aggressive the patching that follows. Because the update cycle is itself authored by the developer population, a faster cycle accelerates the coverage that already exists but does not, by itself, extend it. Cryptography drew the same line: automated cryptanalysis improved for decades and produced none of the side-channel or fault-injection classes, because those categories were not reachable from inside the existing reporting frame, however

hard the search was pushed.

This reconciles two literatures that look opposed. One finds that automation beats humans, which holds inside a fixed threat model. The other finds that evaluators from underrepresented contexts catch failures automation misses, which is a measure of how far the threat model grows. The defender has the harder side of this. A defender works from a fixed list of harms, while an attacker does not. Translation makes the problem worse, because it drops the signals that carry an attack in another language: slang, religious and political references, code-mixing, local euphemism. Both English-speaking evaluators and English-trained systems lose those signals [Shen et al., 2024, Peppin et al., 2025].

What would refute this. The thesis is empirical and makes bets it can lose. Three results would count against it. First, if an automated attacker confined to an English-centric taxonomy surfaced context-specific non-Western harms at the rate evaluators native to those contexts do, Claim 3 would fail, since the coverage gap would be reachable from inside the specification after all. Second, if the harms that evade English benchmarks proved to be randomly distributed across languages and jurisdictions, the premise that omissions are patterned would fail. Third, if incident-reporting pipelines showed no language or access skew and absorbed non-English harms as quickly as English ones, the point-of-capture argument against post-deployment patching would fail. The probe in Section 4 is a first test of the second prediction: had leakage appeared evenly across English and non-English prompts, the patterned-omission premise would have been undercut on the spot.

4 AN EMPIRICAL ANCHOR

To test whether the gap survives in a current open-weight model, we ran 100 HarmBench [Mazeika et al., 2024] behaviors against Llama-3.1-8B-Instruct in English and five more languages: Hindi, Swahili, Bengali, Yoruba, and Tagalog. We sampled the 100 behaviors stratified across HarmBench’s standard, contextual, and copyright subsets in proportion to their registry sizes (50, 25, and 25), drawing uniformly at random within each. NLLB-200 1.3B [NLLB Team et al., 2022] did the translation, and Llama-Guard-3-1B [AI, 2024] labeled each response safe or unsafe. The whole pipeline runs locally on a consumer machine (Apple M4 Pro, 24 GB unified memory) in about two hours. We chose the tool stack so the probe reproduces on a student laptop. Code and data are at https://github.com/krimler/harmbench_multilingual/.

We drop the copyright subset from the reported numbers. It turns out to measure capability instead of safety. The model reproduces English song lyrics on request and cannot do the same in Yoruba, because it cannot write fluent Yoruba lyrics in the first place. We keep the standard and contextual

Table 1: Unsafe-Response Rates on Llama-3.1-8B-Instruct (Standard and Contextual Subsets; Copyright Subset Excluded). Brackets Show 95% Wilson Confidence Intervals.

Language	Standard ($n=50$)	Contextual ($n=25$)
English	0% [0.0, 7.1]	0% [0.0, 13.3]
Hindi	4% [1.1, 13.5]	20% [8.9, 39.1]
Swahili	4% [1.1, 13.5]	24% [11.5, 43.4]
Bengali	8% [3.2, 18.8]	16% [6.4, 34.7]
Yoruba	2% [0.4, 10.5]	24% [11.5, 43.4]
Tagalog	4% [1.1, 13.5]	12% [4.2, 30.0]

subsets, which the model can attempt in all six languages.

Table 1 reports the result. English drew zero unsafe responses across 75 prompts; the five non-English languages drew 5–8 each across the same 75, for overall rates from 6.7% (Tagalog) to 10.7% (Swahili and Bengali). Pooled across the five non-English languages, that is 35 unsafe responses in 375 prompts (9.3%), against 0 in 75 for English. A one-sided Fisher exact test rejects equal rates at $p = 0.0013$. The per-cell Wilson intervals are wide given the small subsets—on the contextual subset the upper bounds reach 30–43%—so no single percentage should carry weight. The qualitative result holds regardless: zero leakage in English, non-zero leakage in every non-English language tested, concentrated in the contextual subset. The magnitude sits well below the 79% jailbreak rate Yong et al. [2023] measured on GPT-4 in late 2023, consistent with two more years of multilingual safety training; what we report is the residual gap that training leaves behind.

Limitations. The experiment is an existence proof. It cannot apportion leakage among translation artifacts, classifier noise, pretraining underexposure, and cultural framing; prior work does that decomposition [Abdullahi et al., 2026, Shen et al., 2024]. We test one open-weight model, though the cited literature finds the pattern across many [Shen et al., 2024, Peppin et al., 2025]. The small per-cell sizes (25 contextual, 50 standard) make the intervals wide, which is why we lead with the pooled Fisher test as the strongest defensible claim. As a partial check on translation noise, requiring two-translator chrF agreement above 0.3 moves language-level rates by at most 2.4 points; the matching classifier-noise check is missing and would need Llama-Guard-3-8B or native-speaker annotation. The 24% Yoruba contextual rate is where that unchecked noise matters most, since it rests on six machine-labeled responses that no Yoruba speaker has verified. And the prompts are machine-translated, so the numbers should be read as a lower bound: Abdullahi et al. [2026] report sharp F1 drops in English-centric guardrails once prompts are fully localized to African languages.

Scope. The probe does not aim to improve on expert-built, culturally grounded benchmarks, which already exist and

continue to appear [Abdullahi et al., 2026, Singhania et al., 2025]. It aims at a different constraint: the cost of running a check at all. A two-hour, single-laptop probe that any evaluator can retarget to a new language or model lowers that cost to the point where contributors outside well-funded labs can take part in threat-model expansion, which is what makes the two-axis program in Section 5 practical to adopt.

5 IMPLICATIONS

Implication 1: evaluate on two axes. Because the human-versus-automation contest folds two activities into one, it should be split. One axis is search efficiency within a stated threat model, where automation is the right tool. The other is threat-model expansion, measured by how fast previously uncatalogued harm categories get surfaced. A corollary: a benchmark release should state its coverage outright—the languages, jurisdictions, and deployment contexts it does and does not represent. A benchmark that does not declare its exclusions cannot be judged on whether they matter.

Implication 2: build outside participation in, and keep it there. Sustained involvement by evaluators outside the developer population is the move Anderson urged for hardware engineers in cryptography and the move the FDA eventually made for women in trials. Neither recovery came from goodwill; both came from new venues, funding, and mandates. The medical case is concrete: the FDA’s 1993 *Guideline for the Study and Evaluation of Gender Differences in the Clinical Evaluation of Drugs* [U.S. Food and Drug Administration, 1993] turned inclusion into a standing regulatory expectation. The stakes rise with agentic systems, where a coverage gap in evaluation propagates into the actions the system selects. The AI-safety equivalents would make threat-model coverage a standing requirement of benchmark releases and fund non-Western evaluation as a recurring budget line.

References

- Tassallah Abdullahi, Macton Mgonzo, Mardiyah Oduwale, Paul Okewunmi, Abraham Owodunni, Ritambhara Singh, and Carsten Eickhoff. UbuntuGuard: A culturally-grounded policy benchmark for equitable AI safety in African languages. *arXiv preprint arXiv:2601.12696*, 2026.
- Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. AI suggestions homogenize writing toward Western styles and diminish cultural nuances. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’25)*. ACM, 2025. doi: 10.1145/3706598.3713564.
- Meta AI. Llama Guard 3 model card. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/8B/MODEL_CARD.md, 2024.

- Mowafak Allaham, Kimon Kieslich, and Nicholas Diakopoulos. Global perspectives of AI risks and harms: Analyzing the negative impacts of AI technologies as prioritized by news media. *arXiv preprint arXiv:2501.14040*, 2025.
- Ross J. Anderson. Why cryptosystems fail. In *Proceedings of the 1st ACM Conference on Computer and Communications Security (CCS)*, pages 215–227. ACM, 1993.
- Philip E. Bickler, John R. Feiner, and John W. Severinghaus. Effects of skin pigmentation on pulse oximeter accuracy at low saturation. *Anesthesiology*, 102(4):715–719, 2005.
- Dan Boneh, Richard A. DeMillo, and Richard J. Lipton. On the importance of checking cryptographic protocols for faults. In *Advances in Cryptology (EUROCRYPT)*, volume 1233 of *Lecture Notes in Computer Science*, pages 37–51. Springer, 1997.
- Ads Dawson, Rob Mulla, Nick Landers, and Shane Caldwell. AIRTbench: Measuring autonomous AI red teaming capabilities in language models. In *arXiv preprint arXiv:2506.14682*, 2025.
- Duke MIDS / Meta capstone project. Blind spots in text-to-image models: Evaluating cultural gaps, failures, and hidden risks in GenAI media safety. Duke University Master in Interdisciplinary Data Science, 2025.
- Paul C. Kocher. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In *Advances in Cryptology (CRYPTO)*, volume 1109 of *Lecture Notes in Computer Science*, pages 104–113. Springer, 1996.
- Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR, 2024.
- Sean McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15458–15463, 2021.
- Ruth B. Merkatz, Robert Temple, Solomon Sobel, Karyn Feiden, and David A. Kessler. Women in clinical trials of new drugs: A change in Food and Drug Administration policy. *New England Journal of Medicine*, 329(4):292–296, 1993.
- Rob Mulla, Ads Dawson, Vincent Abruzzon, Brian Grunke, Nick Landers, Brad Palm, and Will Pearce. The automation advantage in AI red teaming. In *arXiv preprint arXiv:2504.19855*, 2025.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, AI Youngblood, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.
- Aidan Peppin, Julia Kreutzer, Alice Schoenauer Sebag, Kelly Marchisio, Beyza Ermis, John Dang, Samuel Cahyawijaya, Shivalika Singh, Seraphina Goldfarb-Tarrant, Viraat Aryabumi, Aakanksha, Wei-Yin Ko, Ahmet Üstün, Matthias Gallé, Marzieh Fadaee, and Sara Hooker. The multilingual divide and its impact on global AI safety. *arXiv preprint arXiv:2505.21344*, 2025.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, 2024.
- Abhishek Singhania, Christophe Dupuy, Shivam Mangale, and Amani Namboori. Multi-lingual multi-turn automated red teaming for LLMs. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 141–154. Association for Computational Linguistics, 2025.
- Michael W. Sjoding, Robert P. Dickson, Theodore J. Iwashyna, Steven E. Gay, and Thomas S. Valley. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 383(25):2477–2478, 2020.
- U.S. Food and Drug Administration. General considerations for the clinical evaluation of drugs. Technical Report HEW Publication No. (FDA) 77-3040, U.S. Government Printing Office, Washington, DC, 1977.
- U.S. Food and Drug Administration. Guideline for the study and evaluation of gender differences in the clinical evaluation of drugs. *Federal Register*, 58:39406–39416, 1993.
- Laura Weidinger, John Mellor, Bernat Guillén Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, Stevie Bergman, Mikel Rodriguez, Verena Rieser, and William Isaac. STAR: SocioTechnical approach to red teaming language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 21516–21532, 2024.
- Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. In *NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR)*, 2023. arXiv:2310.02446.

Andy Zhou, Kevin Wu, Francesco Pinto, Zhaorun Chen, Yi Zeng, Yu Yang, Shuang Yang, Sanmi Koyejo, James Zou, and Bo Li. AutoRedTeamer: Autonomous red teaming with lifelong attack integration. In *arXiv preprint arXiv:2503.15754*, 2025.