ISSUE BRIEF

FRONTIER MODEL FORUM

Issue Brief: Emerging Practices for Multilingual Evaluations for CBRN and Advanced Cyber Risks

DATE

July 7, 2026

INTRODUCTION

As frontier AI capabilities continue to advance, developing appropriate and comprehensive evaluations for chemical, biological, radiological, and nuclear (CBRN) and advanced cyber risks will be of critical importance. Robust coverage of these risks will require evaluations that work across languages, regions, and cultural contexts. Yet, multilingual evaluation coverage remains limited. Safety training rarely occurs in non-English languages, and where it does, it is limited to a handful of high-resource languages. Current multilingual safety evaluations span a small fraction (with most benchmarks covering between 10 - 20 languages) of the more than 7,000 spoken worldwide, leaving most languages largely absent from existing evaluations.[1] The gap is also reflected in the New Delhi Frontier AI Commitments, signed in February 2026, which recognize the need to strengthen multilingual capabilities and develop evaluations for under-represented languages and cultural contexts. Evaluations and public benchmarks for advanced cyber and CBRN capabilities are especially concentrated in English.

The predominance of English-language evaluations for CBRN and advanced cyber has important implications for frontier AI security. Threat actors do not operate exclusively in English, and a model's safety properties in English may not mirror how it behaves when prompted in other languages by adversaries. There is also emerging evidence that safety performance can degrade outside English.[2] It remains unclear whether this stems from degradation in the underlying model, failures in deployment-layer classifiers, or both. This distinction matters because multilingual evaluations will need to test the model and the deployed product separately to localize a failure and identify the right mitigation. Finally, relevant technical knowledge

is distributed across non-English resources, literature, and communities, so capability assessments conducted only in English may understate what a model can surface when drawing on knowledge that is primarily available in other languages.

This brief highlights emerging industry practices for building multilingual CBRN and cyber evaluations, including key safety gaps and failure patterns, considerations for evaluation development and operationalization, and directions for future work.

SAFETY GAPS AND MULTILINGUAL EVALUATIONS

Multilingual evaluation development should not be understood as translation alone. In many high-risk settings, the relevant question is whether an evaluation captures the language, cultural context, institutional setting, and threat environment in which a model may actually be used. Without this consideration and careful methodological design, multilingual evaluations can overlook critical safety gaps, while failing to measure the full range of capabilities and use cases that emerge across linguistic contexts. These gaps can be exploited along two related but distinct dimensions:

1. **Multilingual Weaknesses**, where translation, transliteration, dialect, code-switching, or low-resource language coverage changes model behavior. Emerging evidence shows that safety training built into models can degrade when they are prompted in non-English languages, particularly lower-resource languages or mixed-language prompts.[3] This includes cross-lingual transfer jailbreaks, code-switching attacks, transliteration or script obfuscation, and multilingual prompt injection. For example, recent joint testing led by Singapore's AI Safety Institute (with AI safety and security institutes and government offices from Japan, Australia, Canada, the European Union, France, Kenya, South Korea, and the UK) translated more than 6,000 prompts into ten languages to test model safeguards across five harm categories, including robustness to jailbreaking and prompt injection. Automated "LLM-as-a-judge" evaluations diverged materially from human expert assessment in several languages, with the largest gaps in Japanese and Telugu, and jailbreak safeguards were consistently the weakest, especially in non-English languages.

2. **Multicultural Weaknesses**, where cultural analogies, values, idioms, shared stories, local institutions, and regional norms change and potentially obfuscate what is harmful. Beyond language, what counts as harmful or as meaningful attacker assistance depends on regional and cultural context, including local regulations, infrastructure, institutions, social norms, and threat-actor behavior. The same model output may remain at low risk in one setting and dangerous in another. While more work remains to be done on how multicultural weakness intersects with CBRN and cyber risks, there is a clear body of research on cultural-specific blind spots. For example, a multicultural red-teaming challenge convened by Singapore's IMDA brought

over 350 participants across nine Asian countries to test large language models for bias stereotypes in regional languages. The exercise surfaced more than 1,000 successfully elicited stereotypes in these languages and outlined several specific cultural blind spots which helped identify geographic, ethnic, and gender biases that unique communities face.

Both dimensions can be used adversarially and should be evaluated separately as well as in combination: a model may perform safely in a language test but fail when the same interaction relies on local idioms, authority structures, social norms, or regional threat context. Reflecting these overarching dimensions, specific and distinct failure patterns have emerged, including:

1. **Cross-lingual / low-resourced language jailbreaks**. This involves translating a prohibited request into a lower-resource language, where safety training is weakest, to elicit a response the model would refuse in English.

2. **Code-switching and mixed-language attacks**. This involves mixing two or more languages, dialects, or scripts within a single prompt to defeat safety filters trained largely on monolingual text.

3. **Transliteration / script obfuscation**. This involves rendering a request in an alternate script or romanized form so it evades pattern-based filters while remaining legible to the model.

4. **Multilingual prompt injection**. This involves embedding adversarial instructions in non-English text inside documents, web pages, or tool outputs that an agentic model ingests and acts on, but that English-centric screening overlooks.

5. **Cultural misinterpretation.** What constitutes harmful content can vary across cultural contexts. As a result, evaluations may need to be adapted – not just translated – to account for local norms, regulations, infrastructure, and threat environments.

These gaps are analogous to the expanded threat landscape created by multimodality. Attackers can route different parts of an attack through different channels (for example, reasoning in text while embedding a trigger in an image) because safety coverage is uneven across modalities. Multilingual and multicultural switching expands the attack surface in a similar way by adding additional channels where guardrails may be patchy. An attacker may shift languages, dialects, scripts, cultural references, or idioms depending on which part of the attack they want the model to reason about differently.

KEY CONSIDERATIONS FOR DEVELOPING MULTILINGUAL EVALS

Multilingual Methodology and Threat Models

Evaluation science is still maturing, but as AI adoption increases globally, it is increasingly important to extend safety infrastructure to other languages and cultural contexts. This requires deliberate consideration of evaluation design for multilingual use, particularly for CBRN and advanced cyber risks. The following principles should be considered:

- **For advanced cyber and CBRN capabilities, the goal is measuring user uplift, rather than just model policy compliance:** The question is whether the model materially advances capability beyond what the user could reach otherwise, which requires domain-experts involved in evaluation development, a baseline calibration, and grounding in human-uplift studies rather than static Q&A. Baseline calibration should capture what resources, expertise, time, and tooling the user would otherwise have without the model.
- **Contextual information should be embedded in the eval**: Jurisdictional, regulatory, and sociocultural context shapes what might count as uplift or harm, and that context should be considered as part of evaluation design. For example, where a country lacks institutional and regulatory safeguards on access to biological materials, a response that reads as low-risk elsewhere may constitute meaningful attacker assistance. Evaluations should therefore capture where the local biosecurity or cyber context can change the severity of attacker assistance.
- **Multilingual evals should include multimodal coverage:** Multilingual and multicultural safety infrastructure should go beyond text and include multimodal evaluation. This reflects how people increasingly interact with AI in practice and captures risks that emerge only when visual, textual, audio, and cultural content are combined. For CBRN and cyber evaluations, this could include prompts that contain images capturing hidden experimental context (such as lab notes) or a network diagram of a target environment, in specific target languages.

Coverage, Data, and Expertise

In developing multilingual evaluations for CBRN and cyber risks, deliberate decisions should be made around coverage and representativeness, with clear documentation of tradeoffs when opting for a specific approach.

- **Approach to Coverage**: The approach to evaluation development should be driven by threat models and should account for dialectal variation, code-switching, script diversity, representativeness of speakers, cultural context, and technical domains. Coverage decisions should be risk-based, taking into account speaker population, language-family and script diversity,

low-resource status, regional threat activity, and the likelihood that adversaries may exploit weaker safety coverage.

- **Developing Coverage Tiers:** Because no evaluation program can cover all 7,000+ languages at full depth, a tiered approach may be needed. One tier could provide full capability for a set of high-priority languages selected by speaker population, risk actor presence, language-family diversity, script diversity, and resource availability. A second tier could provide basic capability for additional languages. A third tier could establish a minimum safety floor and targeted research investments for low-resource languages. For jailbreak and adversarial elicitation work, the threshold should be broader, because adversarial attacks can exploit any language even where ordinary usage is limited.

- **Ensuring Representativeness**: Translated English benchmarks introduce translationese and semantic drift, potentially missing harms without English analogues, and underscore the case for native-authored, culturally grounded evals built from the ground up.

- **Accounting for Data Availability**: Appropriate non-English test data for advanced cyber and CBRN evaluations remains limited. Relevant materials are often unavailable, difficult to access, unevenly distributed across languages, or lack expert annotation, forcing tradeoffs across synthetic generation, translation, native collection, and expert authoring. Each approach has implications for validity, representativeness, and annotator wellbeing.

- **Challenges of Expertise Scarcity:** Each evaluation language requires both deep domain expertise in the risk domain (chemistry, biology, radiological and nuclear) and native fluency in the language of the evaluation. For high-risk domains, native fluency is necessary but not sufficient. Evaluation design and grading should combine native-language expertise with domain expertise in cyber or CBRN.

Security, Handling, and Integrity

Developing CBRN and advanced cyber risk-specific evaluations introduces specific requirements and considerations for handling and securing highly sensitive prompts and model outputs. These dynamics can be amplified and introduce constraints for developing multilingual versions of these evaluations. Several elements are worth considering:

- **Contamination**: Training-data leakage is more difficult to detect and less auditable in multilingual settings. Given that most current multilingual benchmarks are translations of older English ones, there is a high possibility of indirect contamination if the English data is contaminated.[4] Multilingual evaluations may face additional contamination risks through translated prompts, parallel corpora, and benchmark-derived training data.

It is therefore important to build refresh cycles and protected test sets into the evaluation lifecycle.

- **Export Control Constraints**: Evaluation prompts and underlying technical content may be subject to the International Traffic in Arms Regulation (ITAR), Export Administration Regulations (EAR), or equivalent international regimes, constraining who can access, handle, and review items across jurisdictions. This may create challenges for evaluation development across key languages, necessitating additional legal oversight and country-specific handling procedures.

- **Information Hazard Constraints**: Evaluation items themselves can contain sensitive operational details, attack pathways, or technical information that could facilitate harmful activity if disclosed. Multilingual workflows multiply the surface area for leakage through translators, machine translation (MT) services, and annotation platforms that have not been vetted for handling this class of content.

- **Evaluation Security:** Holding advanced cyber and CBRN evals private is increasingly standard, and multilingual extensions amplify this by requiring more handling protocols, cleared personnel, and contamination controls.

CONSIDERATIONS FOR OPERATIONALIZING MULTILINGUAL SAFETY EVALUATIONS

Operationalizing multilingual evaluations well means commissioning native-speaker domain experts and culturally grounded reviewers, and treating high-risk evaluations as private artifacts. It also means avoiding sole reliance on machine-translated English benchmarks, unvalidated LLM judges, monolingual prompts, or a single aggregate score. Additional considerations include:

1. **Judges and Grading**: Results should be reported by language, dialect or script (where relevant), harm category, elicitation method, and rater disagreement, rather than collapsed into a single aggregate score. LLM-as-judge approaches should be validated separately for each language and cultural context before being relied upon, particularly where cultural context, social norms, or linguistic nuance influence the harm assessment.

2. **Elicitation Methodology**: Adversarial elicitation should be conducted by native speakers in target languages, distinguish genuine attacker assistance from safety-training degradation, and increasingly cover agentic and tool-use settings. It should also account for adversaries mixing multiple forms of "encoding" in the same attack: language, culture, script, fluency level, dialect, modality, and domain jargon. A useful design frame is to test combinations across at least language, cultural context, and fluency or domain expertise, rather than assuming that single-language fluent-speaker testing sufficiently reflects the range of adversarial behaviors and contexts the model may encounter.

3. **Governance, Disclosure, and Information Hazards**: Concerns such as tiered access, pre-publication review and protections for annotator and grader safety are all amplified in multilingual and multicultural settings, and should be planned for accordingly.

AREAS FOR FUTURE WORK

There are several initiatives already working to close gaps in multilingual and multicultural safety.[5] However, many of these benchmarks tend to concentrate on a specific set of harms (such as toxicity, hate speech, and child safety), with limited multilingual or multicultural coverage of advanced cyber or CBRN risks. There also tends to be a reliance on translation from English, which can underestimate or misread harms in context.

To help close this gap, there are several issue areas that will need to be addressed to help strengthen the ecosystem and advance multilingual evaluations across critical risk domains:

1. **Developing language-specific uplift baselines**: To ensure meaningful measurement of uplift, it will be important to build target-language web-search and reference baselines.

2. **Investing in the talent pipeline**: To avoid safety gaps and failure patterns, it will be important to support the growth of the workforce at the intersection of advanced cyber, CBRN, non-English fluency, and cultural expertise.

3. **Standardizing disclosure norms**: It will be important to develop standardized disclosure norms to ensure transparency around methodology and aggregate results while keeping sensitive items private. This includes the datasheets covering threat model, baseline, language-specific context, cultural assumptions, and handling constraints.

4. **Building in refresh cycles**: Evaluations should be treated as living infrastructure that decays under contamination pressure and requires regularly scheduled updates and renewal.

Building rigorous evaluations for these risks is a priority for developers and deployers within frontier AI frameworks and for governments through AI Safety and Security Institutes (AISIs), national bodies mandated to conduct technical safety evaluations and coordinate across borders on frontier AI threats.[6] Multilingual and multicultural evaluation efforts intrinsically require deep regional and local partnership, and global AISIs remain natural partners for native-language evaluation development.

As publicly funded technical bodies rooted in their national contexts, AISIs are well-positioned to convene the necessary native-speaker domain experts and provide secure infrastructure for sensitive cyber and CBRN evaluations.

As these efforts scale, there will be a growing need for shared multilingual evaluation infrastructure for advanced cyber and CBRN risks, designed from the outset to be access-controlled and resistant to contamination. The Frontier Model Forum looks forward to working with others in the ecosystem to support the development of needed evaluations and benchmarks for critical risk domains.

FOOTNOTES

1. Ning Z, Gu T, Song J, et al. "LinguaSafe: A Comprehensive Multilingual Safety Benchmark for Large Language Models." arXiv. Published August 18, 2025; Wang W, Tu Z, Chen C, et al. "All Languages Matter: On the Multilingual Safety of Large Language Models." arXiv. Published October 2, 2023; MLCommons. "AILuminate Multimodal." Last modified February 13, 2026
2. Zhang M, Patel A, Truong ST, Koyejo S. "Why Do Safety Guardrails Degrade Across Languages?" arXiv. Published May 22, 2026; Shen L, Tan W, Chen S, et al. "The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts." Findings of the Association for Computational Linguistics: ACL 2024. Published August 2024.
3. Zhang M, Patel A, Truong ST, Koyejo S. "Why Do Safety Guardrails Degrade Across Languages?" arXiv. Published May 22, 2026; Shen L, Tan W, Chen S, et al. "The Language Barrier: Dissecting Safety Challenges of LLMs in Multilingual Contexts." Findings of the Association for Computational Linguistics: ACL 2024. Published August 2024.
4. Uppadhyay M, Beniwal H, Kodali P, Sitaram S. "The State and Fate of Multilingual, Contextual Evaluation in the NLP World." Microsoft Research India.
5. For example, MLCommons AILuminate assesses models – including multi-modal text-and-image interactions – across 12 harm areas in Chinese, French and English, with planned expansion to include Hindi, Tamil, Malay, Korean, and Japanese.
6. The International AI Safety Report explicitly flags multilingual safety as an underdeveloped research area, underscoring the need for dynamic, scenario-based testing that holds across languages and deployment contexts. Further, the Singapore Consensus on Global AI Safety Research Priorities identifies multilingual safety as a shared international research priority.