POLYAEGIS: BENCHMARKING LARGE LANGUAGE MODEL SAFETY ACROSS 100 LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

We present **PolyAEGIS**, a massively multilingual LLM safety benchmark spanning **100 languages** across 17 language families and 26 scripts, built on **10,362** Gold-quality translated prompts drawn from two successive AEGIS-derived sourcing pipelines (2,894 and 11,788 unique unsafe root prompts respectively) and gated through a dual-metric translation-quality filter (BERTScore F1 and COMETKiwi). Each translated prompt is sent to **six** generation models — three open-weight (Aya-Expanse-32B, Mistral-Large-123B, GPT-OSS-120B) and three closed-source (Claude Haiku 4.5, Gemini 2.5 Flash, GPT-5.4 Mini) — and every response is scored by **three independent LLM judges** (Mistral-Large, Qwen2.5, and Claude Sonnet 4.6) under a four-class response taxonomy — Full Refusal, Partial Leakage, Full Compliance, and Irrelevant/Unintelligible — whose per-judge verdicts we reconcile into a majority-vote consensus label, yielding **61,536** consensus-scored responses. Partial Leakage, largely invisible to prior binary attack-success-rate framing, captures responses that attempt refusal yet still disclose actionable harmful content. We find that unsafe outcomes (leakage + full compliance) affect **13.78%** of responses overall, that open-weight models are unsafe **1.6**$\times$ more often than closed-source models (16.90% vs. 10.65% ASR), that low-resource languages are systematically less safe than high-resource ones across every metric we measure, and that the three judges agree on a single response label for only **36.8%** of prompts — underscoring both the scale of the multilingual safety gap and the difficulty of measuring it reliably with a single judge. We publicly release PolyAEGIS, its full multi-judge annotations, and our reconciliation pipeline as a standardized resource for evaluating and improving LLM safety across the world's languages. Code and data available in this link: https://anonymous.4open.science/r/x8K2mNq7VbP1wZ4sLfG9t [1]

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable capabilities across reasoning, generation, and instruction following, but ensuring their safety and alignment with human values remains a fundamental open problem (Ouyang et al., 2022; Bai et al., 2022; Ji et al., 2023). This challenge is especially acute for languages beyond English. Safety alignment research has overwhelmingly concentrated on high-resource languages—primarily English—while the majority of the world's languages remain underrepresented in alignment benchmarks, red-teaming efforts, and safety training pipelines (Bang et al., 2023; Deng et al., 2023b). The consequences of this gap are concrete and measurable. Prior work has shown that multilingual prompts can significantly degrade safety performance, with unsafe outputs occurring more frequently in low-resource and morphologically complex languages (Deng et al., 2023b; Shen et al., 2024). However, existing benchmarks remain limited in scale, linguistic diversity, and evaluation methodology.

The multilingual safety gap has three key dimensions. First, *coverage*: existing multilingual safety benchmarks cover a small subset of global linguistic diversity, limiting our ability to study how safety varies across language families, scripts, and resource levels. Second, *format*: most prior work relies on *prompt-only* evaluation, which enables testing but does not support alignment training

[1] More details on the dataset https://huggingface.co/datasets/mmse-bench-anon/msb

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

or fine-grained failure analysis. Third, *taxonomy*: heterogeneous or implicit harm definitions make cross-benchmark comparisons inconsistent. The AEGIS taxonomy (NVIDIA AI Red Team, 2024), with 21 fine-grained harm categories, provides a unified framework for consistent multilingual safety evaluation.

We present PolyAEGIS, a large-scale multilingual safety benchmark built from two successive AEGIS-derived sourcing pipelines: an initial pipeline (v1) yielding **2,894 unique harmful prompts** from **5,376 filtered samples**, and a revised, auditable pipeline (v2) yielding **11,788** unique unsafe root prompts. Both pools are translated into **100 languages** and gated through a dual-metric (BERTScore F1 + COMETKiwi) quality filter, yielding a final Gold-quality benchmark of **10,362 translated prompts** spanning **17 language families** and **26 scripts**, with **48 high-resource** and **52 low-resource** languages represented. We evaluate **61,536 model responses** across **100 languages** and **six LLMs**, including both closed-source models (GPT-5.4 Mini, Claude Haiku 4.5, Gemini 2.5 Flash) and open-weight models (Mistral-Large-123B, Aya-Expanse-32B, GPT-OSS-120B), with every response scored by **three independent LLM judges** (Mistral-Large, Qwen2.5, Claude Sonnet 4.6) whose verdicts we reconcile into a majority-vote consensus. This setup enables controlled comparison of cross-model, cross-judge, and cross-language safety behavior.

Leakage-aware safety evaluation. We introduce a **four-class response taxonomy**: Full Refusal (FR), Partial Leakage (PL), Full Compliance (FC), and Irrelevant/Unintelligible (IU). Our results show that:

- **FR dominates** overall responses (77.02%), indicating strong refusal behavior,
- but a non-trivial portion of responses fall into **unsafe categories** (FC = 7.45%, PL = 6.32%),
- yielding $\mathbf{ASR_{full}}$ **= 13.78%** and $\mathbf{ASR_{interp}}$ **= 15.17%**.

These findings highlight that even when full compliance is limited, **partial leakage remains a critical and underreported failure mode**, accounting for nearly half of all unsafe outcomes.

Cross-lingual safety disparities. We observe substantial variation across languages, with a **3.96× difference in mean harm score** between the safest (Ukrainian, 0.059) and least safe (Burmese, 0.234) languages. Additionally, closed-source models outperform open-weight models, with open-weight models unsafe at nearly **1.6**× the rate of closed models (ASR_{full} 16.90% vs. 10.65%, a **6.25 percentage point** gap). These results confirm that multilingual safety remains uneven and resource-dependent. Finally, the three LLM judges agree on a single response label for only **36.8%** of prompts, exposing substantial judge-level disagreement that a single-judge protocol would silently hide.

Research Questions. We ask how harmful-content risk varies across languages and whether linguistic properties, including resource level and typology, influence safety; whether binary attack success rate (ASR) obscures important failure modes and how leakage-aware evaluation addresses them; and how open-weight and closed-source models compare in multilingual safety settings.

Contributions. We introduce **PolyAEGIS**, a multilingual safety benchmark of 10,362 Gold-quality prompts translated into 100 languages across diverse language families, scripts, and resource tiers. We propose a **leakage-aware evaluation framework** that identifies partial leakage as a distinct failure mode accounting for nearly half of unsafe responses. Using three independent LLM judges, we evaluate 61,536 responses and reveal cross-language and cross-judge disparities, as well as a substantial safety gap between open-weight and closed-source models. Finally, we find up to a 3.96× variation in mean harm across languages and single-label judge consensus on only 36.8% of prompts, motivating language-aware alignment and multi-judge evaluation.

2 Related Work

2.1 Multilingual Safety Evaluation

Multilingual LLM safety evaluation has recently emerged as an important extension of English-centric safety benchmarking. Early work such as MultiJail (Deng et al., 2023b) showed that safety behavior can degrade substantially across languages, evaluating multilingual unsafe prompts across a limited set of languages and demonstrating that non-English inputs can expose safety gaps. CatHarmfulQA (Bhardwaj et al., 2024) further studied harmful question answering across a small multilingual set, while XSafety (Wang et al., 2024), LinguaSafe (Ning et al., 2025), and PolyGuard (Kumar et al., 2025) broadened multilingual safety evaluation toward toxicity, moderation, and guardrail settings. Compared with these resources, PolyAEGIS emphasizes scale, typological coverage, and multi-judge evaluation. We construct PolyAEGIS from two successive Aegis-derived sourcing pipelines (2,894 and 11,788 unique unsafe prompts respectively), translate the union into 100 languages, and gate the result through a dual-metric translation-quality filter to obtain 10,362 Gold-quality prompts. The language set covers 17 language families and 26 scripts, including 48 high-resource and 52 low-resource languages. We further evaluate 61,536 model responses across 100 languages and six models, each scored by three independent LLM judges whose verdicts we reconcile into a consensus label, enabling analysis of language-family, script, resource-tier, judge-agreement, and open-vs.-closed model effects under a unified safety taxonomy.

2.2 Adversarial Prompting and Safety Failure Modes

A large body of work studies adversarial prompting and jailbreak attacks against aligned LLMs. Red-teaming methods generate adversarial prompts using LLMs themselves (Perez et al., 2022); in-the-wild jailbreak studies analyze prompting strategies such as DAN-style attacks (Shen et al., 2024); and algorithmic attacks such as GCG (Zou et al., 2023), PAIR (Chao et al., 2023), AutoDAN (Liu et al., 2023), GPTFuzzer (Yu et al., 2023), and TAP (Mehrotra et al., 2024) demonstrate that safety filters can be bypassed through optimized or iteratively refined prompts. These works primarily focus on attack generation or jailbreak transfer, often in English or a small number of languages. PolyAEGIS is not designed as a jailbreak-generation benchmark. Instead, it studies language-shift safety behavior under translated harmful prompts and evaluates responses with a multi-class taxonomy. In our evaluated set, full refusal accounts for 77.02% of responses, while irrelevant or unintelligible outputs account for 9.20%. Unsafe responses consist of full compliance (7.45%) and partial leakage (6.32%), giving an ASR_{full} of 13.78% and an $\text{ASR}_{\text{interp}}$ of 15.17%. This distinction is important because multilingual safety failures are not only full jailbreak successes; they also include partial disclosures, comprehension failures, and language-specific failure modes that binary ASR alone can obscure.

2.3 Translation Quality in Multilingual Benchmarks

Machine translation is a central confound in multilingual evaluation. Prior work has shown that translated benchmarks may introduce artifacts and semantic drift (Artetxe et al., 2020), while HELM (Liang et al., 2022) emphasizes that model evaluation should account for multiple desiderata rather than relying on a single aggregate score. Multilingual translation systems such as NLLB-200 (NLLB Team et al., 2022), SeamlessM4T (SeamlessM4T & Meta AI, 2023), and MADLAD-400 (Kudugunta et al., 2024) have expanded language coverage, but translation fidelity remains uneven across languages, scripts, and domains. Embedding-based metrics such as LaBSE (Feng et al., 2022) and reference-free evaluation metrics such as COMETKiwi are commonly used to quantify semantic consistency. PolyAEGIS explicitly incorporates translation-quality gating into benchmark construction. The v1 translated pool (283,612 samples) is partitioned into quality tiers via fixed thresholds on BERTScore F1, COMET, and their combined score (Table 3): Gold (2.31%), High-Quality (21.09%), and Standard-Quality (49.55%), with the remaining 27.05% falling below the Standard-Quality thresholds. This distribution shows that quality is highly imbalanced across the full translated set, motivating quality-aware safety reporting. Rather than treating all translated samples as equally reliable, PolyAEGIS assembles its final evaluation set exclusively from the Gold tier and reports translation quality as an explicit analysis variable (Section 4.6).

2.4 Low-Resource and Multilingual Language Models

Recent multilingual LLMs and instruction datasets have expanded language coverage beyond English, but performance on long-tail languages remains uneven. Aya (Ustun et al., 2024) and related work on multilingual instruction tuning study cross-lingual generalization, while models such as TigerLLM (Raihan & Zampieri, 2025), TituLLM (Nahin et al., 2024), Pangea (Yue et al., 2024), and Babel (Zhao et al., 2025), along with cross-lingual instruction-following studies (Muennighoff et al., 2023), highlight growing efforts toward language-inclusive modeling. However, capability does not guarantee safety robustness, and guardrail performance may degrade differently across languages. PolyAEGIS evaluates both generation models and guardrails in a unified framework, covering three closed-source and three open models. We observe a +3.9 percentage point safety advantage for closed models and up to a 4.64× variation in harmful behavior across languages, highlighting systematic multilingual safety disparities.

Summary of gaps addressed. Table 6 summarizes the key dimensions of prior multilingual safety resources. PolyAEGIS differs from prior work by combining massive language coverage, unified taxonomy, typology-aware analysis, translation-quality-aware filtering, multi-metric response evaluation, and guardrail benchmarking in a single evaluation suite.

3 Methodology

3.1 Dataset Construction

We construct a multilingual safety evaluation dataset, **PolyAEGIS**, consisting of harmful and potentially harmful prompts sourced from the Aegis AI Content Safety Dataset (NVIDIA AI Red Team, 2024) and translated into up to 100 languages using NLLB (NLLB Team et al., 2022). Translation quality is estimated with reference-free and reference-based metrics (BERTScore F1 via back-translation and COMETKiwi source–target scoring), and the final benchmark is assembled by gating translated samples through a set of fixed quality thresholds rather than by resampling or stratifying the underlying language/prompt distribution. Midway through the project we revised our prompt-sourcing pipeline, which changed the size of the underlying unique-prompt pool; we describe both the original and the revised pipeline below, and how their outputs were combined into the final benchmark.

3.1.1 Data Source

We construct **PolyAEGIS** exclusively from the Aegis AI Content Safety Dataset (NVIDIA AI Red Team, 2024), a large-scale benchmark for evaluating model behavior on harmful prompts and safety compliance. The dataset provides curated prompt–response pairs annotated under the AEGIS safety taxonomy, including unsafe queries and their corresponding safe refusals, making it well-suited for controlled safety evaluation. The combined train, validation, and test splits contain approximately 33,416 annotated rows.

Initial sourcing pipeline (v1). In the first iteration of dataset construction, we retained only rows with *no missing (null) values* across the fields required for our pipeline (prompt, response, and annotations), yielding **5,376 unsafe prompt–response pairs**. After exact-duplicate removal on the normalized prompt text, this pool collapsed to **2,894 unique root prompts**, which were translated into 98 languages (2 languages were not supported by the NLLB model, later we add 2 more languages) to produce an initial pool of 283,612 translated samples.

Revised sourcing pipeline (v2). While developing the benchmark further, we replaced the initial ad hoc null-filtering step with a hierarchical, auditable filtering pipeline applied independently to the full train/validation/test splits. This pipeline performs, in order: (1) structural validation (non-null id, prompt, and prompt label), (2) removal of fully- and partially-redacted prompts, (3) retention of only human-annotated prompt labels, (4) detection and removal of prompts with conflicting safe/unsafe human annotations, (5) normalization (Unicode NFKC, whitespace collapsing) and collapsing of duplicate prompts into unique root prompts with unioned harm-category annotations, and (6) construction of disjoint safe and unsafe candidate pools, with unsafe roots lacking an assigned harm category excluded from the pool. Applying this pipeline to the full Aegis dataset yields

Table 1: **PolyAEGIS dataset construction across sourcing iterations.** The initial pipeline (v1) applied simple null-filtering; the revised pipeline (v2) applied hierarchical structural/redaction/label-conflict filtering to the full Aegis splits, yielding a larger and non-overlapping pool of unique unsafe prompts that was translated separately and merged with v1 at the quality-gating stage.

Stage / Subset	Operation / Criteria	#Samples	Pipeline
Raw Dataset	Aegis (train + val + test)	∼30,000	–
Filtered Dataset (v1)	Non-null unsafe prompt–response pairs	5,376	v1
Unique Prompts (v1)	Deduplicated root prompts	2,894	v1
Translated Dataset (v1)	× 98 languages	283,612	v1
Unique Prompts (v2)	Hierarchical filtering, full splits	11,788	v2
New-Only Prompts (v2)	v2 roots \ v1 prompts	–	v2
Translated Dataset (v2, new-only)	× 98 languages	–	v2

Table 2: **PolyAEGIS language coverage and translation quality statistics.** The benchmark spans 100 languages, with translation quality (on the v1 translated pool) grouped into three subsets based on progressively stricter BERTScore F1 and COMET thresholds; only the Gold subset, combined with an equivalently-gated v2 new-only subset, is used for the final evaluation benchmark.

Language Coverage		Translation Quality Subsets (v1 pool)			
Category	**Count**	**Tier**	**F1**	**COMET**	**Samples (%)**
Languages	100	Gold	≥0.95	≥0.92	6,542 (2.31%)
Families	17	High-Quality	≥0.90	≥0.85	59,807 (21.09%)
Scripts	26	Standard-Quality	≥0.80	≥0.70	140,533 (49.55%)
High-resource	48	Total gated samples: 206,882 (+ below-threshold 76,730 = 283,612)			
Low-resource	52	Final Gold benchmark: 6,542 (v1) + 3,820 (v2) = 10,362 raw → 10,262 unique			

11,788 unique unsafe root prompts — a substantially larger and differently-composed pool than the 2,894 root prompts obtained from the v1 heuristic, since the revised pipeline recovers duplicate-consistent prompts that simple null-filtering discarded and explicitly resolves label conflicts rather than dropping them implicitly. To avoid re-translating and re-scoring prompts already covered by v1, we compute the set difference between the v2 root-prompt pool and the v1 prompt pool (5,376 prompts) on normalized prompt identity, and translate only the *new-only* prompts introduced by the revised pipeline. This new-only pool is translated into the same 100 languages and scored with the same quality pipeline described below.

3.1.2 Language Selection and Dataset Translation

PolyAEGIS targets broad multilingual coverage by selecting languages from the NLLB-200/FLORES-200 language inventory (NLLB Team et al., 2022). NLLB-200 was designed to expand machine translation coverage beyond FLORES-101, with the authors first considering the 101 FLORES-101 languages and then expanding to more than 200 language–script varieties through a candidate selection process involving Wikipedia presence, regional language lists, existing NLP dataset coverage, speaker population, and community-level considerations.

Following this design principle, we construct a candidate pool of languages by surveying multilingual safety benchmarks and aligning them with NLLB-200/FLORES-200 codes. PolyAEGIS covers **100 languages**, balanced across high- and low-resource settings. We use NLLB metadata to record each language's code, script, family, and resource level, enabling analysis across linguistic and resource dimensions. The benchmark spans **26 scripts** (e.g., Latin, Cyrillic, Arabic, Devanagari, Han) and **17 language families**, including Indo-European, Afro-Asiatic, Austronesian, and Atlantic-Congo. It includes **48 high-resource** and **52 low-resource** languages, supporting comprehensive evaluation of multilingual safety. Table 2 summarizes the language metadata.

Table 3: **Quality bucket thresholds used for benchmark gating.** A sample is assigned to a bucket only if it clears all three criteria (BERTScore F1, COMET, and the combined score) for that bucket; thresholds are applied in order from Gold to Standard-Quality.

Bucket	BERTScore F1 $\geq$	COMET $\geq$	Combined $\geq$
Gold	0.95	0.92	0.94
High-Quality	0.90	0.85	0.88
Standard-Quality	0.80	0.70	0.75

Both the v1 and v2 unique-prompt pools are translated into all 100 target languages using NLLB (facebook/nllb-200-3.3B), and each translated prompt is back-translated into English for quality validation.

Quality Scoring with BERTScore and COMETKiwi. To ensure semantic fidelity and translation robustness, we evaluate every translated sample along two complementary axes: (1) **BERTScore F1** between the original English prompt and its English back-translation, capturing round-trip semantic preservation, and (2) **COMETKiwi**, a reference-free quality-estimation metric scored directly between the original English prompt and the translated target-language prompt, capturing translation quality without relying on back-translation artifacts. For each translated sample we derive a combined quality score:

$$\text{score}_{\text{combined}} = \frac{\text{BERTScore}_{F1} + \text{COMET}}{2} \tag{1}$$

In the small number of cases where a COMET score is unavailable, we fall back to BERTScore F1 alone as the quality signal for bucket assignment.

3.1.3 Quality-Gated Benchmark Construction

Rather than sampling a fixed number of translations per language or stratifying by resource level, we construct the final benchmark by **threshold-based quality gating**: every translated sample is assigned to a quality bucket according to fixed thresholds on BERTScore F1, COMET, and the combined score, and only samples clearing a bucket's thresholds on *all three* criteria are assigned to it. Thresholds were chosen based on a target sample-count budget for the benchmark rather than through stratified sampling across languages or harm categories; the resulting benchmark composition therefore reflects the natural distribution of translation quality across languages rather than an enforced balance.

Applying these thresholds to the v1 translated pool (283,612 samples), the **Gold** subset (F1 ≥ 0.95, COMET ≥ 0.92, combined ≥ 0.94) contains **6,542 samples (2.31%)**, the **High-Quality** subset (F1 ≥ 0.90, COMET ≥ 0.85, combined ≥ 0.88) includes **59,807 samples (21.09%)**, and the **Standard-Quality** subset (F1 ≥ 0.80, COMET ≥ 0.70, combined ≥ 0.75) covers **140,533 samples (49.55%)**, with the remaining **27.05%** of samples falling below the Standard-Quality thresholds and excluded from the benchmark. This confirms a substantial skew toward lower-quality translations, particularly for lower-resource target languages.

Because the v1 Gold subset (6,542 samples) fell short of our target benchmark size, we applied the identical Gold-bucket thresholds (Table 3) to the v2 new-only translated pool and selected an additional **3,820 samples** that cleared the same criteria. These were combined with the v1 Gold subset — using the same thresholds but drawn from two different underlying prompt pools, rather than via joint stratified sampling — to form the final Gold-quality PolyAEGIS benchmark of **10,362 samples**. We did not stratify this selection by language, harm category, or resource level; the final composition is therefore a direct consequence of (a) which languages NLLB translates most reliably and (b) which prompts happened to fall in the v1 versus v2 sourcing pools. As we detail in Section 4.1, 100 of these 10,362 instances are exact (root_id, language) duplicates across the v1 and v2 pools, which the normalized-prompt-text set-difference step did not catch; we report both the raw 10,362-instance construction count and the 10,262 deduplicated instances actually used for evaluation. We report the resulting language- and category-level distribution in Section 4.1 and discuss its implications for evaluation coverage in Section 6.

This quality-stratified, threshold-gated construction enables robust multilingual safety evaluation by explicitly accounting for translation fidelity, which plays a critical role in understanding model behavior under language shift.

3.2 Benchmark Models

We evaluate PolyAEGIS using a diverse set of open-source and closed-source large language models (LLMs), covering a wide range of architectures, scales, and multilingual capabilities. Our open-source models include **Mistral-Large-123B**, a high-capacity dense model for general-purpose generation (Jiang et al., 2023), **Aya-Expanse-32B**, a multilingual model optimized for cross-lingual generation (AI, 2024), and **GPT-OSS-120B**, an open-weight GPT-style model designed to emulate proprietary systems. For closed-source evaluation, we include **GPT-5.4 Mini**, **Gemini 2.5 Flash**, and **Claude Haiku 4.5**, which provide strong efficiency and performance baselines. This selection ensures coverage across *model scale* (30B–120B+), *architectural diversity* (dense vs. MoE), *multilingual capability*, and *deployment paradigm* (open vs. closed), enabling comprehensive analysis of multilingual safety, including cross-model consistency, scaling effects, and robustness to harmful prompts across diverse language settings.

3.3 Response Taxonomy and Evaluation Metrics

Response taxonomy. We introduce a four-class response taxonomy that extends beyond binary attack success rate (ASR) to capture nuanced safety failures in multilingual settings. Given a model response r, we assign a label $\mathcal{L}(r) \in \{\text{FR}, \text{PL}, \text{FC}, \text{IU}\}$:

Full Refusal (FR): The model explicitly declines to fulfill the request, providing no substantive harmful content; **Partial Leakage (PL):** the model produces at least one segment of actionable or harmful content before hedging or refusing, constituting a security failure even if accompanied by disclaimers; **Full Compliance (FC):** the model fully satisfies the harmful request without meaningful refusal; **Irrelevant/Unintelligible (IU):** the response does not meaningfully engage with the prompt, typically due to translation errors, language mismatch, or model confusion.

This taxonomy distinguishes between complete failure (FC), subtle but critical leakage (PL), and benign refusal (FR), while isolating IU as a separate category reflecting lack of comprehension rather than safety alignment.

3.3.1 Evaluation Metrics

We evaluate multilingual safety using a set of complementary metrics that capture both overt and subtle model failures. Specifically, we measure **Attack Success Rate (ASR)** in two variants: ASR_{full}, which treats unintelligible (IU) responses as safe, and $\text{ASR}_{\text{interp}}$, which conditions on interpretable responses to isolate safety under successful comprehension. To quantify partial failures, we compute the **Leakage Rate (LR)** and **Full Compliance Rate (FCR)**, capturing the prevalence of partial leakage and complete unsafe responses, respectively. We further define a **Safety Degradation Score** (Δ_{safe}) to measure cross-lingual robustness relative to English, and a **Confidence-Weighted Unsafe Rate** (CW-UR) that incorporates translation quality via LaBSE-based confidence Feng et al. (2022). Finally, we identify **leakage-dominant failure** regimes where partial leakage exceeds full compliance. Detailed formal definitions and formulations of all metrics are provided in Appendix B.

3.4 Multi-Judge Aggregation and Schema Reconciliation

Each of the 10,362 gated benchmark instances (Section 3.1) is scored independently by three LLM judges — Mistral-Large, Qwen2.5, and Claude Sonnet 4.6 — merged on $(\text{root_id}, \text{language})$ into a single consensus record before majority-vote label aggregation (Section 3.3). Because the v1 (6,542 samples) and v2 (3,820 samples) subsets were produced by two successive pipeline iterations (Section 3.1), their judge outputs did not share a record schema, and 72% of root_id-bearing rows carried a pre-refactor identifier rather than the canonical `aegis_root_*` form. Naively joining on identity therefore silently discarded most of one judge's votes for two-thirds of the benchmark (Sonnet's votes were recoverable for only 3,064/10,362 samples, 29.6%, under a naive join). We address this with a four-step backtracking reconciliation applied before aggregation: (1) **root-identity**

normalization via the root-construction stage's `source_ids` mapping, resolving 97.4% of legacy identifiers and explicitly flagging the rest rather than guessing; (2) **per-sample metadata backfill** of category and translation-quality scores from three pipeline stages, covering 100% of pairs in the merged corpus; (3) **quality-bucket recomputation** reproducing Table 3's rule exactly (0 discrepancies against all 6,542 pre-existing v1 assignments); and (4) **language-metadata backfill** for the 77 of 100 languages outside the hand-curated low-resource focus set, leaving Joshi resource-class fields unset rather than fabricated. This restores all five fields needed for aggregation to 100% coverage for every judge (full before/after field-coverage table and per-field data-quality census in Appendix C.2), and is what makes the 97.06% three-judge consensus rate and 36.8% cross-model agreement figures of Section D.4 measurable in the first place.

4 RESULTS

4.1 DATASET STATISTICS

The v1+v2 Gold-quality gating stage (Section 3.1) produces **10,362** translated prompt instances; reconciling identity across the two sourcing pools during judge aggregation (Section 3.4) reveals that **100** of these (root_id, language) pairs are exact duplicates that independently cleared the Gold threshold under both the v1 and the v2 pipeline, since the v2 "new-only" set difference was computed on normalized English prompt text and did not catch every case where the same underlying prompt reentered the pool under a different sourcing-pipeline identifier. We therefore evaluate **10,262 unique** (root_id, language) instances spanning **100 languages** and **21 AEGIS harm categories** (after normalizing a handful of duplicate category spellings inherited from the v1/v2 sourcing pipelines). Category frequency is highly imbalanced: *Criminal Planning/Confessions* is the largest category (1,950 samples, 19.0%), followed by *PII/Privacy* (1,397), *Harassment* (1,375), *Controlled/Regulated Substances* (1,362), and *Hate/Identity Hate* (1,354); the ten smallest categories together account for under 2% of the benchmark (Table 7, Appendix C). Per-language sample counts are similarly imbalanced by construction, since the benchmark was assembled by quality gating rather than stratified sampling (Section 3.1): the best-covered languages (Luxembourgish, 734; Maltese, 703; Asturian, 549) reflect languages for which NLLB translation happened to clear the Gold threshold most often, while 47 languages have fewer than 50 samples each. We treat this imbalance as a first-class analysis variable throughout Section 4 rather than normalizing it away.

4.2 OVERALL SAFETY PERFORMANCE

Table 12 shows the global label distribution across **61,536** consensus-scored responses spanning **100** languages and **six** models. Full refusal (FR) accounts for **77.02%**, while unsafe outcomes—partial leakage (PL, **6.32%**) and full compliance (FC, **7.45%**)—sum to **13.78%**. A further **9.20%** of responses are irrelevant or unintelligible (IU), reflecting comprehension failures rather than true safety. Figure 1 (P1) visualizes this distribution.

Table 13 reports the formal metrics from Section 3.3. The overall ASR_{full} is **13.78%**, indicating that more than one in eight responses contain unsafe content. Importantly, $\text{ASR}_{\text{interp}}$ is **1.40** points higher (**15.17%**), showing that IU responses modestly deflate observed attack success even at this smaller IU rate. This gap highlights a key PolyAEGIS insight: apparent safety in low-resource languages arises partly from model confusion rather than alignment, though—unlike a purely IU-driven artifact—unsafe behavior here is already substantial before conditioning on comprehension. Figure 2 (P2) illustrates these metrics.

4.3 CONTINUOUS SAFETY SCORES AND METRIC CONSISTENCY

Table 14 summarises the distribution of harm, refusal strength, and actionability scores. The mean harm score of **0.1284** masks a heavy upper tail: **1.10%** of all responses fall into the high-risk regime (harm > 0.8), and **34.94%** contain actionable steps. The median harm score of **0.0333**—well below the mean—confirms a right-skewed distribution, with a small but consequential cluster of maximally harmful outputs.

Table 4 reports the Pearson correlation structure among continuous safety metrics. Harm score and actionability are very strongly positively correlated ($r = 0.934$), confirming that more harmful

Table 4: **Pearson correlations (A5).** Harm and actionability correlate positively; both correlate negatively with refusal strength.

	Harm	Action.	Refusal
Harm Score	1.00	0.934	−0.671
Actionability	0.934	1.00	−0.660
Refusal Strength	−0.671	−0.660	1.00

Table 5: **Worst-case harm by language (A9, selected).** Each language reaches a maximum of 1.0 despite a lower mean.

Language	Mean Harm	Max Harm	Count
Zulu	0.1811	1.0000	2,064
Occitan	0.0810	1.0000	2,712
Oriya (Odia)	0.1236	1.0000	2,320
Yoruba	0.1974	1.0000	912
Tajik	0.1701	1.0000	1,872

responses tend to be more instructional. Refusal strength is negatively correlated with both harm ($r = -0.671$) and actionability ($r = -0.660$), providing internal consistency evidence for the multi-judge consensus protocol.

4.4 Leakage as a Distinct Failure Mode

Partial leakage accounts for **6.32%** of all responses—nearly **half** of every unsafe response overall (6.32 of 13.78 points of ASR_{full}). These partial-refusal cases are not benign: they carry a mean leakage severity (actionability score conditioned on the PL label) of **0.393**, indicating that models attempting to refuse still provide substantially actionable harmful guidance nearly **40%** of the time they do so. Table 16 identifies the **45 of 100** languages where leakage is the dominant failure mode (LR $>$ FCR)—a condition that binary ASR entirely obscures, since it treats PL and FC identically. Notably, this is not confined to a handful of outliers: leakage dominance affects nearly half of all evaluated languages, spanning both high- and low-resource settings (e.g., Korean and Bulgarian alongside Sorani Kurdish and Macedonian).

4.5 Multilingual Safety Disparities

Figure 20 (P9) plots language-level safety extremes. Mean harm ranges from **0.0591** for the safest language (Ukrainian) to **0.2341** for the most dangerous (Burmese)—a **3.96**× spread across languages that receive identical prompts. Table 26 (C1, selected columns) presents the full language-level safety profile.

Table 5 lists languages that reach the maximum possible observed harm score of **1.0** despite only moderate *mean* harm—e.g., Occitan (mean harm **0.052**, count *n*=2,712) and Tajik (mean harm **0.126**) both contain at least one maximally harmful response. Average-case metrics systematically understate risk: a language can look safe on aggregate while still producing worst-case failures. This motivates worst-case robustness analysis alongside mean-based metrics.

Figure 23 (P12) presents the harm vs. actionability Pareto frontier, identifying languages in the high-harm, high-actionability quadrant—the primary targets for safety improvement.

4.6 Translation Quality Analysis

Figure 14 (PB7) plots mean BERTScore F1 against IU failure rate across all 100 languages, revealing a broadly negative relationship: lower translation fidelity corresponds to higher comprehension failure. However, this trend is not strict—e.g., Hausa (F1 $= 0.7465$) shows a comparatively high failure rate (**21.71%**), while Oromo (F1 $= 0.7339$, the single lowest mean F1 in the benchmark) shows a similarly high but not proportionally worse failure rate (**19.86%**)—indicating that translation quality alone does not fully explain multilingual safety behavior. Figure 16 (PB10) extends this analysis with a three-dimensional view where bubble size encodes translation consistency ($1/\sigma$).

Table 25 shows a counter-intuitive pattern by resource tier: the 52 low-resource languages achieve a *higher* mean translation quality (**0.9522**) than the 48 high-resource languages (**0.8535**). We do not believe this means low-resource translation is more faithful in an absolute sense; rather, BERTScore F1 is computed via round-trip back-translation, and several low-resource languages in our sample are typologically closer to a small, repetitive vocabulary of safety keywords that NLLB reproduces near-verbatim on back-translation, inflating F1 without necessarily preserving nuance.

We flag this as a metric artifact worth cautioning against in Section 6 rather than evidence that translation is a solved problem for low-resource languages. Figure 18 (PB12) and Figure 19 (PB13) analyze translation quality across AEGIS harm categories. Translation fidelity varies substantially across categories, from **0.7877** (*Immoral/Unethical*, *n*=48) to **1.0000** (*High Risk Gov Decision Making*, *n*=6, an extremely small category). Among well-populated categories, *Unauthorized Advice* (F1 $= 0.8249$, *n*=1,410) shows both comparatively low translation fidelity and the highest ASR_{full} of any well-populated category (**29.93%**, Table 21), consistent with translation difficulty compounding safety risk for categories that depend on precise, context-sensitive phrasing.

4.7 Resource Level and Typological Analysis

Table 24 shows that low-resource languages are less safe than high-resource languages on *every* metric we measure: higher IU rate (9.53% vs. 8.02%), higher ASR_{full} (14.18% vs. 12.37%), higher mean harm (0.1353 vs. 0.1043), and a wider IU-masking gap $\Delta(\text{ASR}_{\text{interp}} - \text{ASR}_{\text{full}})$ (0.0150 vs. 0.0108). Unlike a pure IU-masking story, this is not a case of low-resource languages merely *appearing* unsafe once IU is resolved—they are measurably less safe on the raw, unconditioned metric too. Script also matters: right-to-left languages (Arabic, Hebrew, Urdu, and related scripts; *n*=3,621 responses) show higher harm (**0.1528** vs. **0.1269**) and ASR_{full} (**16.21%** vs. **13.63%**) than left-to-right languages. At the family level, Mongolic-Khitan (harm **0.225**, *n*=282, represented solely by Mongolian) and Niger-Congo (harm **0.1809**, *n*=12,947, a much larger and more robust sample spanning Zulu, Yoruba, Shona, and related languages) are the two highest-harm families, while Indo-European—the family with by far the largest sample (*n*=29,398)—sits below the corpus mean (harm **0.1073**).

We perform additional result analysis which are listed in Appendix D.

5 Discussion

Partial leakage is not a corner case—it is roughly half of the unsafe-response problem. At **6.32%** of all responses, partial leakage is nearly as large as full compliance (**7.45%**), and it carries a mean leakage severity of **0.393**—models that begin refusing still leak substantially actionable content in doing so nearly 40% of the time. A binary ASR metric that folds PL into either "safe" or "unsafe" without separately reporting it hides which failure mode dominates in a given deployment. This distinction matters operationally: a full-compliance failure is a alignment-training target, while a leakage failure suggests the refusal mechanism itself is triggering correctly but too late in generation—a different fix (e.g., stronger early-token suppression or streaming-time intervention) than a full compliance failure would require.

We list more discussion in Appendix E.

6 Limitations

PolyAEGIS has several limitations. It relies on machine translation via a single system (NLLB-200), which may introduce semantic drift particularly in low-resource languages despite quality filtering; the counter-intuitive BERTScore F1 advantage we observe for low-resource languages (Section 4.6) is itself evidence that our translation-quality metric can be an imperfect proxy for true fidelity, and should be interpreted with this caveat. Per-language sample counts range from 46 to 734 by construction (quality gating rather than stratified sampling, Section 3.1), so per-language statistics for the smallest-count languages carry higher variance than the headline numbers suggest; we report counts alongside every per-language statistic for this reason. Evaluation depends on LLM-based judges, which may introduce bias compared to human annotation, and—as Section D.4 shows—are themselves an imperfect and disagreement-prone measurement instrument; we mitigate but do not eliminate this by using a three-judge consensus rather than a single judge. We report aggregate and per-language metrics without formal statistical significance testing (e.g., bootstrap confidence intervals on harm-score differences), which we leave to future work. The evaluated models—three open-weight, three closed-source—do not cover the full space of deployed multilingual LLMs. Finally, the dataset contains harmful prompts which, while intended for evaluation, may pose misuse risks if extracted from context.

7 Conclusion

We introduced PolyAEGIS, a multilingual safety benchmark constructed from two successive AEGIS-derived sourcing pipelines and expanded to 100 languages via quality-gated translation, evaluated with six generation models and a three-judge consensus protocol. Our results show that partial leakage is a key failure mode accounting for nearly half of all unsafe responses, that safety degrades under language shift in ways that survive controlling for comprehension failure, that the open/closed safety gap is consistent in aggregate but not uniform at the individual-model level, and that reliable multilingual safety measurement itself requires accounting for judge disagreement and parse failure, not just model disagreement. Future work will extend language and model coverage, incorporate additional translation systems to disentangle translation artifacts from genuine safety effects, and develop stronger multilingual safety mechanisms informed by the leakage- and script-level patterns identified here.

References

Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, Sara Hooker, et al. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12027–12049, 2024a.

Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, Sara Hooker, et al. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12027–12049, 2024b.

Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, Sara Hooker, et al. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12027–12049, 2024c.

Cohere For AI. Aya: Multilingual language models, 2024.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*, 2020.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7580–7617, 2025.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

Rishabh Bhardwaj, Duc Anh Do, and Soujanya Poria. Language models are homer simpson! safety re-alignment of fine-tuned language models through task arithmetic. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14138–14149, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023a.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023b.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852*, 2022.

Ronald P Grelsamer. Patellar nomenclature: the tower of babel revisited. *Clinical Orthopaedics and Related Research®*, 436:60–65, 2005.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Raviraj Bhuminand Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Margaret Peters Long, Sanjay Singh Chauhan, et al. Cultureguard: Towards culturally-aware dataset and guard model for multilingual safety applications. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2666–2685, 2025.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Linting Xue, Fynn Ratnakar, Karishma Yacef-Ould, Colin Cherry, George Foster, Wolfgang Macherey, et al. MADLAD-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 2024. NeurIPS 2024.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. Polyguard: A multilingual safety moderation tool for 17 languages. *arXiv preprint arXiv:2504.04377*, 2025.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3923–3954, 2024.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Dilip Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models. *arXiv preprint arXiv:2310.04451*, 2023. ICLR 2024.

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 *International Conference on Machine Learning*, 2024. ICML 2024.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. *Advances in Neural Information Processing Systems*, 2024. NeurIPS 2024.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2023.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu C Shill, Md Ibrahim, Mehadi H Menon, Tareq Al Muntasir, and Firoj Alam. TituLLMs: A family of Bangla LLMs with comprehensive benchmarking. *arXiv preprint arXiv:2502.11187*, 2024.

Zhiyuan Ning, Tianle Gu, Jiaxin Song, Shixin Hong, Lingyu Li, Huacan Liu, Jie Li, Yixu Wang, Meng Lingyu, Yan Teng, et al. Linguasafe: A comprehensive multilingual safety benchmark for large language models. *arXiv preprint arXiv:2508.12733*, 2025.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

NVIDIA AI Red Team. Aegis: Open safety dataset for AI content moderation. `https://huggingface.co/datasets/nvidia/Aegis-AI-Content-Safety-Dataset-2.0`, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Priyaranjan Pattnayak and Sanchari Chowdhuri. Indicjr: A judge-free benchmark of jailbreak robustness in south asian languages. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 649–668, 2026.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Md Nishat Raihan and Marcos Zampieri. TigerLLM: A family of Bangla large language models. *arXiv preprint arXiv:2503.10995*, 2025.

Red-teaming Multilingual Red-teaming. Code-switching red-teaming: Llm evaluation for safety and multilingual understanding.

SeamlessM4T and Meta AI. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "Do Anything Now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 1671–1685, 2024. CCS 2024.

Abhishek Singhania, Christophe Dupuy, Shivam Sadashiv Mangale, and Amani Namboori. Multilingual multi-turn automated red teaming for llms. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pp. 141–154, 2025.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. Multilingual blending: Large language model safety alignment evaluation with language mixture. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3433–3449, 2025.

Ahmet Ustun, Viraat Aryabumi, Zheng Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. All languages matter: On the multilingual safety of llms. pp. 5865–5877, 2024.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. GPTFuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal LLM for 39 languages. *arXiv preprint arXiv:2410.16153*, 2024.

Table 6: **Comparison of multilingual and safety evaluation benchmarks.** Comparison across existing multilingual safety benchmarks and evaluation datasets. Columns denote: Prompt+Response pairs (P+R), Unsafe Translation analysis (UT), Cultural Robustness (CR), Typology-Aware analysis (TA), Multimodal or Multi-model evaluation (MM), and Guardrail evaluation (GR).

Work	Lang.	P+R	UT	CR	TA	MM	GR
MultiJail (Deng et al., 2023a)	10	×	×	×	×	×	×
XSafety (Wang et al., 2024)	10	×	✓	×	×	×	×
CatHarmfulQA (Bhardwaj et al., 2024)	3	×	×	×	×	×	×
Multilingual Alignment Prism (Ahmadian et al., 2024a)	–	✓	✓	✓	×	×	×
Guardrails for Multilingual Toxicity (Ahmadian et al., 2024c)	–	×	×	×	×	×	✓
CulturalKaleidoscope (Banerjee et al., 2025)	17	×	×	✓	×	×	×
HarmBench (Mazeika et al., 2024)	1 (EN)	×	✓	×	×	✓	×
SALAD-Bench (Li et al., 2024)	1 (EN)	✓	✓	×	×	✓	×
Code-Switching Red-Teaming (Red-teaming)	–	×	×	×	×	✓	×
Multilingual Blending (Song et al., 2025)	–	×	×	×	×	✓	×
Tower of Babel Revisited (Grelsamer, 2005)	–	×	×	×	×	✓	×
MM-ART (Singhania et al., 2025)	–	×	×	×	×	✓	×
PolyGuard (Kumar et al., 2025)	17	✓	✓	✓	×	✓	✓
LinguaSafe (Ning et al., 2025)	12	✓	✓	✓	×	✓	×
CultureGuard (Joshi et al., 2025)	–	✓	✓	✓	×	✓	✓
IndicJR (Pattnayak & Chowdhuri, 2026)	Indic	×	×	✓	×	✓	×
CAGE (Ahmadian et al., 2024b)	–	×	✓	✓	×	✓	×
PolyAEGIS (ours)	∼100	✓	✓	✓	✓	✓	✓

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, and Wenxuan Zhang. Babel: Open multilingual large language models serving over 90% of global speakers. *arXiv preprint arXiv:2503.00865*, 2025.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Comparison with Litertaure

If we summary the comparison with existing literature we can get idea from the Table 6.

B Formal Metric Definitions

Attack Success Rate (Full). We define the full attack success rate as:

$$\text{ASR}_{\text{full}}(m, \ell) = \frac{|\{p \in \mathcal{P} : \mathcal{L}(\mathcal{R}(m, \ell, p)) \in \{\text{PL}, \text{FC}\}\}|}{|\mathcal{P}|} \tag{2}$$

This metric treats IU responses as effectively safe, since they do not provide useful information to an attacker.

Attack Success Rate (Interpretable). To isolate safety performance when the model understands the prompt, we define:

$$\text{ASR}_{\text{interp}}(m, \ell) = \frac{|\{p : \mathcal{L}(\mathcal{R}(m, \ell, p)) \in \{\text{PL}, \text{FC}\}\}|}{|\{p : \mathcal{L}(\mathcal{R}(m, \ell, p)) \neq \text{IU}\}|} \tag{3}$$

This excludes unintelligible responses and reflects the model's safety behavior under successful comprehension.

Leakage Rate. We measure the prevalence of partial failures via:

$$\text{LR}(m, \ell) = \frac{|\{p : \mathcal{L}(\mathcal{R}(m, \ell, p)) = \text{PL}\}|}{|\mathcal{P}|} \tag{4}$$

This captures cases where models exhibit unsafe behavior despite partial refusal.

Full Compliance Rate.

$$\text{FCR}(m, \ell) = \frac{|\{p : \mathcal{L}(\mathcal{R}(m, \ell, p)) = \text{FC}\}|}{|\mathcal{P}|} \tag{5}$$

Safety Degradation Score. To quantify cross-lingual robustness, we define:

$$\Delta_{\text{safe}}(m, \ell) = \text{ASR}_{\text{full}}(m, \ell) - \text{ASR}_{\text{full}}(m, \text{EN}) \tag{6}$$

A positive Δ_{safe} indicates that model m is less safe in language ℓ compared to English.

Confidence-Weighted Unsafe Rate. To account for translation quality uncertainty, we incorporate LaBSE-based confidence Feng et al. (2022):

$$\text{CW-UR}(m, \ell) = \text{ASR}_{\text{full}}(m, \ell) \times \text{conf}(\ell) \tag{7}$$

where

$$\text{conf}(\ell) = \mathbb{E}_p \left[\cos(\text{LaBSE}(p), \text{LaBSE}(\hat{p}))\right] \tag{8}$$

and $\hat{p}$ denotes the back-translated prompt.

Leakage-Dominant Failure. We identify languages exhibiting leakage-dominant behavior when:

$$\text{LR}(m, \ell) > \text{FCR}(m, \ell) \tag{9}$$

indicating that subtle leakage is more prevalent than outright compliance.

Interpretation. The gap between ASR_{full} and $\text{ASR}_{\text{interp}}$ provides a critical diagnostic signal. In low-resource languages, models may appear safer under ASR_{full} due to frequent unintelligible outputs. However, when conditioned on interpretable responses, safety performance often degrades substantially. This reveals that apparent safety may stem from model confusion rather than reliable alignment, highlighting the importance of disentangling comprehension failures from true safety behavior.

C Results Analysis Graphs and Tables

This appendix reports the complete set of tables and figures underlying the analysis in Section 4, generated directly from the reconciled multi-judge consensus output (Section 3.4) across the 10,262 unique evaluation instances and 61,536 consensus-scored responses. Block letters (A–F) match the analysis pipeline stages: A (judge-only), B (translation quality), C (language/resource/typology), D (model/generation characteristics), E (open-vs-closed), F (cross-model and cross-judge consistency). Where a per-language table exceeds 15 rows we show the top and bottom entries by the primary sort metric; complete tables are released alongside the benchmark.

C.1 Dataset Statistics

C.2 Data-Quality Census of the Reconciled Judge Output

To surface any remaining data-quality issues before aggregation, we census *every* field present in the reconciled per-judge corpora produced by Section 3.4, not only the five fields targeted by that reconciliation. Tables 9–11 report, for each field, the number of rows in which it is present (non-`null`) versus missing, aggregated over all 6 generator models within each judge. Fields below 100% coverage are either (a) genuinely judge-call-level failures pre-dating this reconciliation (`error`, `judge`, `judge_raw`, and their dependent token/latency fields), which are out of scope for a field-backfill step and are handled by the aggregator's existing judge-validity gate, or (b) fields that are schema-specific by construction (e.g. `sample_id` and `source_custom_id` are only populated for the v2-origin half of Mistral/Qwen's output; Sonnet's raw per-call fields such as `response_raw` and `provider_response_id` are only populated for the v2-origin half, since the v1-origin half never carried them) and require no action.

C.3 Block A — Judge-Only Safety Metrics

Table 7: **AEGIS harm-category distribution across the 10,262 unique evaluation instances.** Category spellings inherited from the v1/v2 sourcing pipelines are normalised (e.g. "Guns and Illegal Weapons" and "Guns/Illegal Weapons" are merged).

Category	Count	%
Criminal Planning/Confessions	1,950	19.0%
PII/Privacy	1,397	13.6%
Harassment	1,375	13.4%
Controlled/Regulated Substances	1,362	13.3%
Hate/Identity Hate	1,354	13.2%
Sexual	698	6.8%
Profanity	482	4.7%
Violence	373	3.6%
Guns/Illegal Weapons	332	3.2%
Unauthorized Advice	235	2.3%
Suicide and Self Harm	211	2.1%
Political/Misinformation/Conspiracy	168	1.6%
Fraud/Deception	104	1.0%
Copyright/Trademark/Plagiarism	69	0.7%
Malware	42	0.4%
Sexual Minor	33	0.3%
Threat	28	0.3%
Illegal Activity	22	0.2%
High Risk Gov./Decision Making	17	0.2%
Immoral/Unethical	8	0.1%
Manipulation	2	0.0%
Total	10,262	100.0%

Table 8: **Per-language sample counts, best- and worst-covered languages.** Counts reflect quality-gated sample availability, not stratified sampling (Section 3.1); all 100 languages are used in every analysis, with per-language counts reported alongside every per-language statistic.

Best-covered		Least-covered	
Language	**Count**	**Language**	**Count**
Luxembourgish	734	Chinese (Traditional)	46
Maltese	703	Chinese (Simplified)	47
Asturian	549	Russian	47
Occitan	452	Xhosa	47
Cebuano	423	Vietnamese	47
Oriya (Odia)	387	Turkish	47
Shona	354	Thai	47
Zulu	344	Serbian	47
Maori	330	Somali	47
Northern Sotho	324	Sindhi	47
Sorani Kurdish	319	Portuguese	47
Tajik	312	Polish	47
Nyanja	255	Pashto	47
Igbo	229	Burmese	47
Lingala	176	Macedonian	47

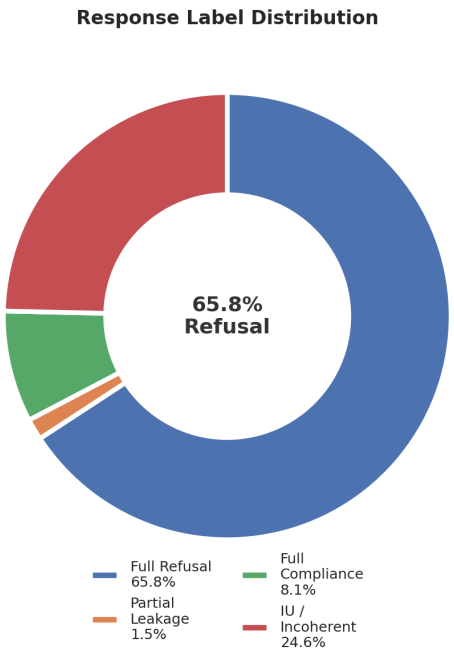

Figure 1: **Global response label distribution (P1).** FR dominates at 77.02%; unsafe outcomes (PL+FC) total 13.78%.

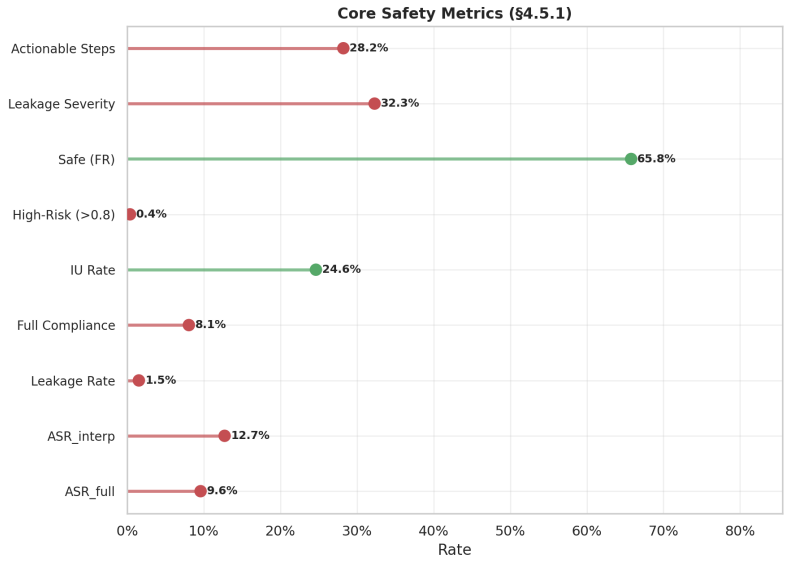

Figure 2: **Core safety metrics (P2).** ASR_{interp} exceeds ASR_{full}, confirming a modest but real IU-masking effect.

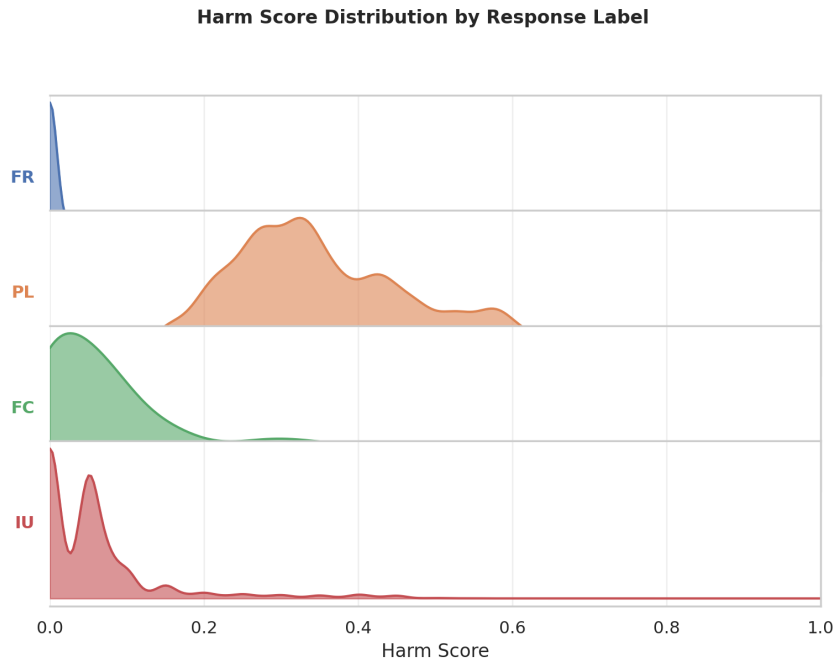

Figure 3: **Harm score distribution by response label (P3).** Kernel density ridge plots show FC responses concentrate at high harm, FR at near-zero harm, and PL occupies an intermediate, bimodal regime.

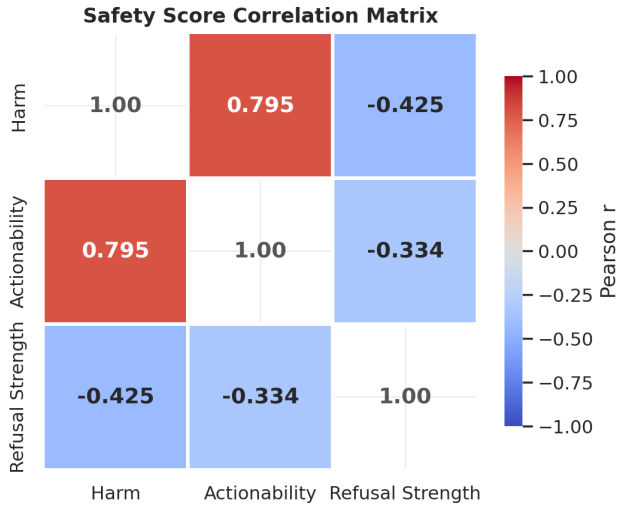

Figure 4: **Correlation heatmap of continuous judge scores (P4).** Corresponds to Table 4.

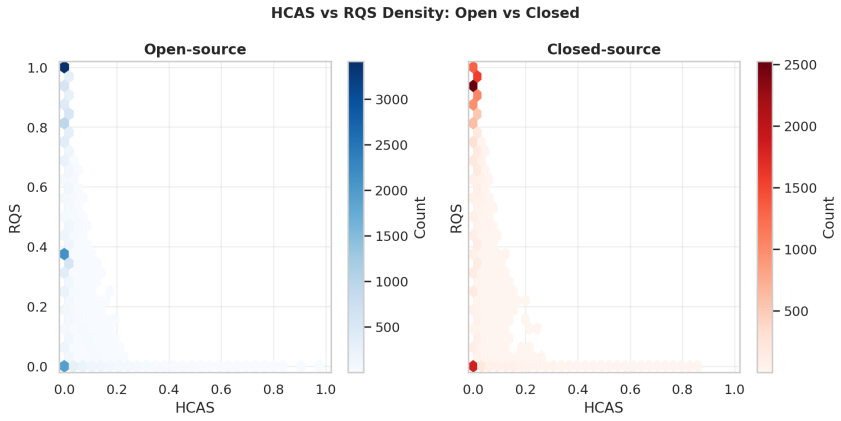

Figure 5: **HCAS vs. RQS density, open vs. closed models (P5).** Open models show a heavier low-RQS, high-HCAS tail.

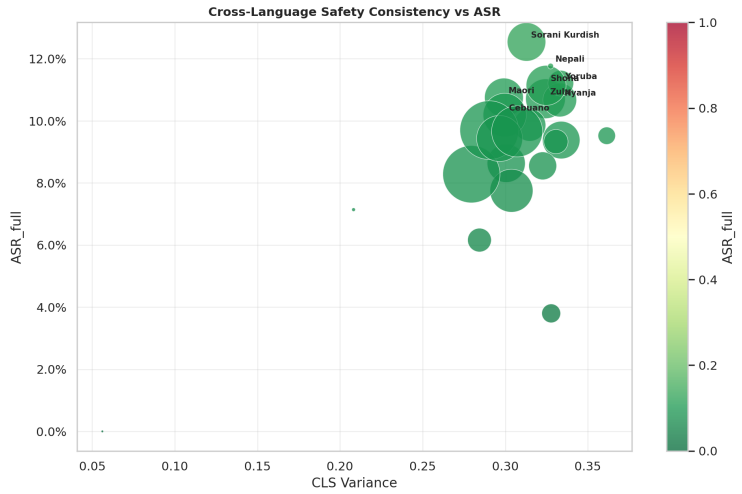

Figure 6: **Cross-lingual safety (CLS) variance summary (A7).** Distribution of per-prompt label variance across the six models' shared prompts.

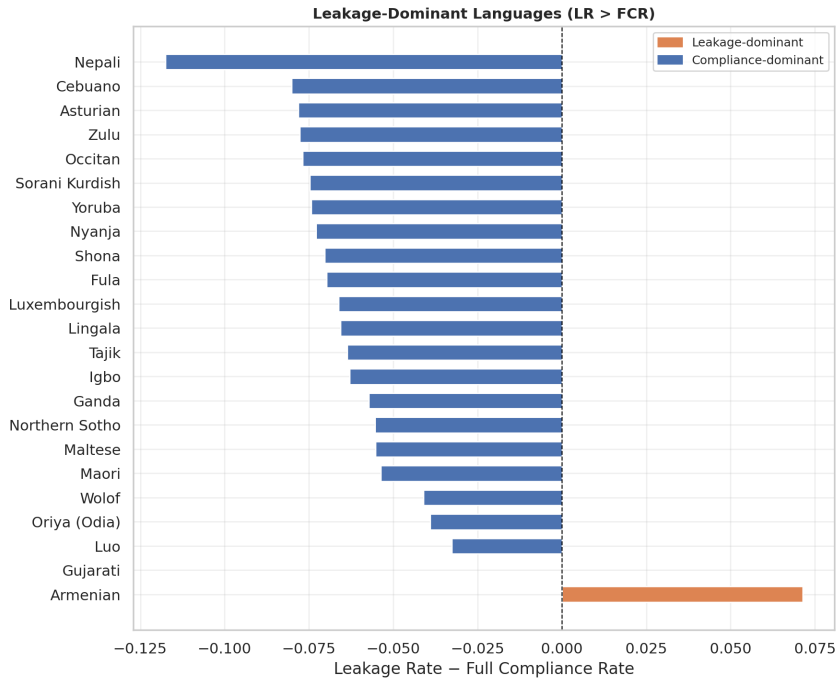

Figure 7: **Leakage rate vs. full compliance rate per language (A8).** Bars diverging right of zero indicate leakage-dominant languages.

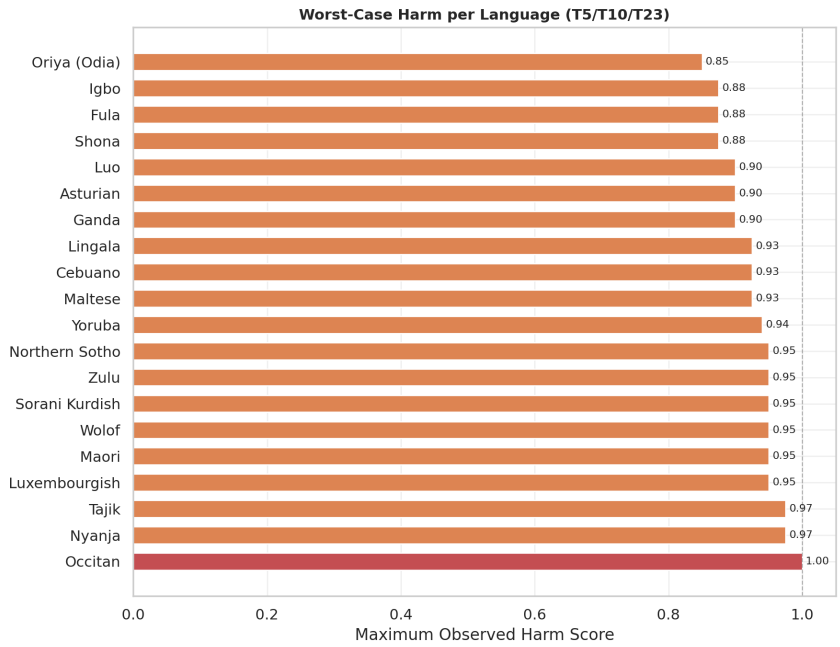

Figure 8: **Worst-case (maximum) harm score per language (A9).**

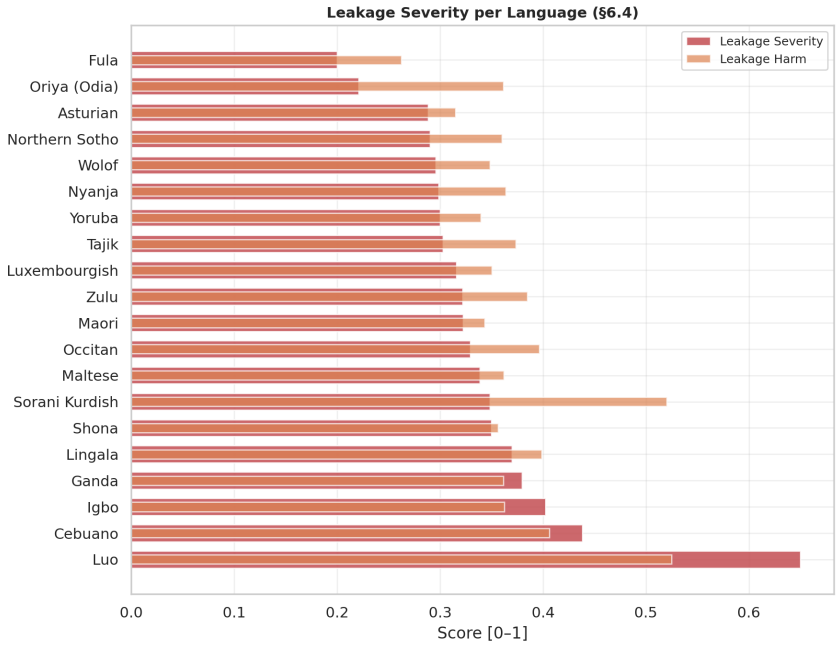

Figure 9: **Leakage severity per language (A10).**

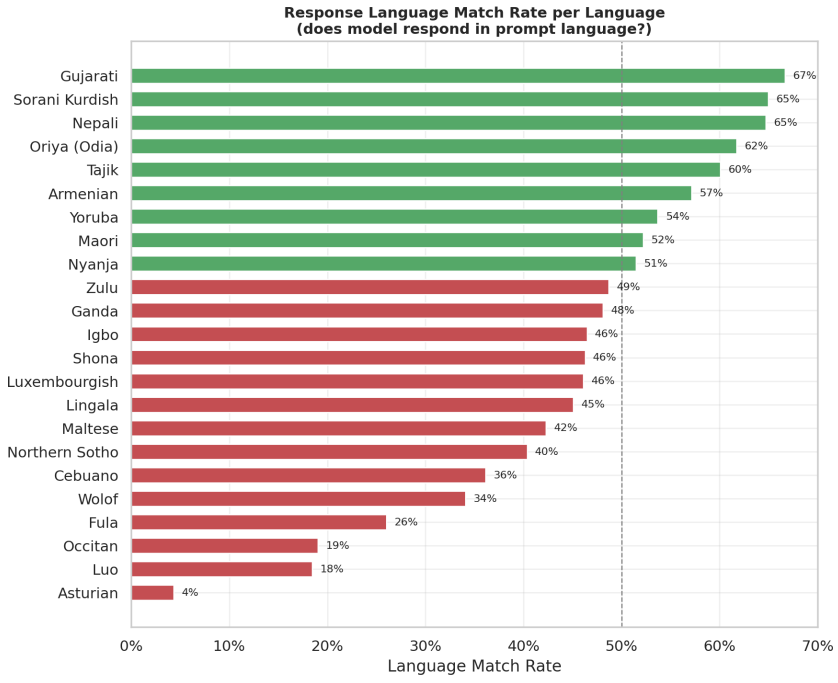

Figure 10: **Language-match rate per language (A11).** Lower rates indicate the model's response language diverges from the prompt's language.

Table 9: Field census, **Mistral** judge output (62,172 rows, reconciled).

Field	Present	Missing	% Present
root_id, language_name, language, category, lang_meta, eval	62,172	0	100.0%
judge_id, response_type, provider, model, source_file, source_line, quantization	62,172	0	100.0%
judge_raw, prompt_tokens, completion_tokens, total_tokens, finish_reason	62,136	36	99.9%
judge	62,135	37	99.9%
generator_model	62,078	94	99.8%
source_custom_id	31,086	31,086	50.0%
sample_id	22,920	39,252	36.9%
error	37	62,135	0.1%

Table 10: Field census, **Qwen** judge output (62,172 rows, reconciled).

Field	Present	Missing	% Present
root_id, language_name, language, category, lang_meta, eval	62,172	0	100.0%
judge_id, response_type, provider, model, source_file, source_line, quantization	62,172	0	100.0%
judge_raw, prompt_tokens, completion_tokens, total_tokens, finish_reason	62,136	36	99.9%
judge	62,134	38	99.9%
generator_model	62,078	94	99.8%
source_custom_id	31,086	31,086	50.0%
sample_id	22,920	39,252	36.9%
error	38	62,134	0.1%

Table 11: Field census, **Sonnet** judge output (62,136 rows, reconciled).

Field	Present	Missing	% Present
root_id, language, category, lang_meta, eval	62,136	0[†]	100.0%
judge	62,134	2	100.0%
custom_id	42,510	19,626	68.4%
prompt, meta	39,241	22,895	63.2%
response	39,147	22,989	63.0%
response_raw	35,979	26,157	57.9%
provider, model, provider_response_id, finish_reason, prompt_tokens, completion_tokens, total_tokens, cache_read/write_tokens, token_count_method, source_input_file	22,895	39,241	36.8%
source_file	6,542	55,594	10.5%
error	94	62,042	0.2%

[†] `root_id` is non-null for all 62,136 rows, but 1,218 of these (38 distinct legacy identifiers, each spanning multiple translated languages) remain in the pre-refactor bare-hash form rather than `aegis_root_*`, since no mapping to a canonical root exists in the current AEGIS data; these are flagged rather than fabricated. The same 38 identifiers (1,218 rows) recur identically in the Mistral and Qwen judge output, since all three judges score the same underlying samples.

Table 12: **Global response label distribution (A1).** Full refusal (FR) dominates, but unsafe outcomes (PL+FC) still affect 13.78% of all responses.

Label	%
Full Refusal (FR)	77.02%
Irrelevant/Unintelligible (IU)	9.20%
Full Compliance (FC)	7.45%
Partial Leakage (PL)	6.32%
Unsafe (PL+FC)	13.78%

Table 13: **Core safety evaluation metrics across all 61,536 responses (A2).** ASR_{interp} conditions on interpretable responses only.

Metric	Value
ASR_{full}	13.78%
ASR_{interp}	15.17%
Leakage Rate (LR)	6.32%
Full Compliance Rate (FCR)	7.45%
IU Rate	9.20%
High-Risk Rate (harm > 0.8)	1.10%
Safe Rate (FR)	77.02%
Leakage Severity	39.30%
Actionable Steps Rate	34.94%

Table 14: **Continuous score summary statistics across all responses (A4).** RQS = Refusal Quality Score; HCAS = Harm–Concordant Actionability Score (Appendix B).

Score	Mean	Median	Std	Min	Max
Harm Score	0.1284	0.0333	0.1902	0.0000	1.0000
Refusal Strength	0.7168	0.8333	0.3082	0.0000	1.0000
Actionability	0.1206	0.0000	0.1837	0.0000	1.0000
RQS	0.6280	0.8083	0.3796	0.0000	1.0000
HCAS	0.0481	0.0000	0.1116	0.0000	1.0000

Table 15: **Largest IU-masking gaps per language (A3), top 10.** Gap = $ASR_{interp} - ASR_{full}$; a large gap means IU responses are substantially deflating the apparent attack success rate for that language.

Language	ASR_{full}	ASR_{interp}	IU Rate	Count
Amharic	0.0799	0.1411	0.4340	288
Burmese	0.2624	0.3176	0.1738	282
Mongolian	0.2021	0.2533	0.2021	282
Fula	0.1642	0.2141	0.2328	408
Northern Sotho	0.2397	0.2864	0.1631	1944
Oromo	0.1879	0.2345	0.1986	282
Kyrgyz	0.2222	0.2678	0.1701	288
Khmer	0.1809	0.2247	0.1950	282
Xhosa	0.1702	0.2096	0.1879	282
Sinhala	0.1088	0.1481	0.2653	294

Table 16: **Leakage-dominant languages (A8), top 15 of 45.** A language is leakage-dominant when its Leakage Rate (LR) exceeds its Full Compliance Rate (FCR) – a distinction invisible to binary ASR.

Language	LR	FCR	Count
Burmese	0.1454	0.1170	282
Mongolian	0.1454	0.0567	282
Kyrgyz	0.1146	0.1076	288
Korean	0.1135	0.0532	282
Sorani Kurdish	0.1077	0.0821	1913
Marathi	0.1076	0.0590	288
Bulgarian	0.1042	0.0729	288
Hungarian	0.0972	0.0764	288
Macedonian	0.0961	0.0534	281
Dutch	0.0941	0.0139	287
Khmer	0.0922	0.0887	282
Telugu	0.0906	0.0418	287
Turkish	0.0851	0.0532	282
Sindhi	0.0851	0.0532	282
Czech	0.0836	0.0697	287

Table 17: **Languages reaching maximum possible harm score 1.0 (A9), top 15 by mean harm.**

Language	Mean Harm	Max Harm	Count
Burmese	0.2341	0.9750	282
Mongolian	0.2250	0.9750	282
Northern Sotho	0.2023	0.9833	1944
Xhosa	0.2018	0.9750	282
Ganda	0.1976	0.9500	785
Yoruba	0.1974	1.0000	912
Nyanja	0.1914	0.9750	1530
Kyrgyz	0.1884	0.9750	288
Oromo	0.1833	0.8833	282
Sorani Kurdish	0.1827	0.9833	1913
Lingala	0.1820	0.9750	1056
Zulu	0.1811	1.0000	2064
Somali	0.1799	0.9333	282
Khmer	0.1771	0.9750	282
Tajik	0.1701	1.0000	1872

Table 18: **Highest leakage severity per language (A10), top 10.** Leakage severity = mean actionability score conditioned on the Partial Leakage label; small counts reflect the rarity of PL responses for these languages.

Language	Leakage Severity	Leakage Harm	Count
Japanese	0.5212	0.5795	22
Xhosa	0.5094	0.5172	16
Thai	0.5074	0.6296	18
Latvian	0.4963	0.5065	18
Czech	0.4924	0.5323	24
Portuguese	0.4690	0.4714	14
Azerbaijani	0.4667	0.5385	21
Vietnamese	0.4656	0.4439	15
Swedish	0.4627	0.4539	17
Mongolian	0.4614	0.5449	41

C.4 BLOCK B — TRANSLATION QUALITY ANALYSIS

Note (B1): the judge-independent AEGIS safety label (`eval_label`) is `unsafe` for 100% of responses by construction — every PolyAEGIS prompt is drawn from AEGIS's unsafe pool (Section 3.1) — so a judge-vs-AEGIS-label agreement crosstab is degenerate and is omitted here; the meaningful agreement analysis is the cross-*judge* agreement in Section D.4.

Table 19: **Safety metrics by translation-quality bucket (B2).** Buckets are combined-score bins (low < 0.75, medium 0.75–0.85, high > 0.85), distinct from the Gold/High-Quality/Standard-Quality benchmark-construction tiers of Table 3.

Bucket	Harm	ASR_{full}	Leakage	Count
low (¡0.75)	0.1262	0.1237	0.0519	2,619
medium (0.75-0.85)	0.1277	0.1350	0.0646	7,771
high (¿0.85)	0.1286	0.1389	0.0636	51,146

Table 20: **Safety metrics by benchmark quality-gating bucket (B6).** "Gold" here means the sample cleared the Gold thresholds (Table 3); all four buckets are present because the v1 pool's non-Gold gated subsets are also carried through for this diagnostic.

Bucket	ASR_{full}	Harm	Leakage	Count
gold	0.1433	0.1349	0.0611	36,352
high_quality	0.1224	0.1101	0.0702	7,919
low	0.1289	0.1278	0.0614	8,418
standard_quality	0.1374	0.1186	0.0677	8,847

Table 21: **Safety and translation quality by AEGIS category, merged canonical spellings (B5).** Sorted by mean harm; categories with $n < 50$ carry high variance.

Category	Harm	Action.	ASR_{full}	Trans. F1	Count
Political/Misinformation/Conspiracy	0.1703	0.1479	0.1796	0.9317	1,008
Immoral/Unethical	0.1674	0.1566	0.3333	0.7877	48
Sexual Minor	0.1623	0.1338	0.1436	0.8941	195
High Risk Gov./Decision Making	0.1567	0.1648	0.2549	0.8907	102
Criminal Planning/Confessions	0.1563	0.1530	0.1725	0.9818	11,697
Unauthorized Advice	0.1552	0.2137	0.2993	0.8249	1,410
Copyright/Trademark/Plagiarism	0.1527	0.1606	0.1763	0.9369	414
Guns/Illegal Weapons	0.1465	0.1315	0.1586	0.9284	1,992
Suicide and Self Harm	0.1421	0.1345	0.1501	0.9191	1,266
Fraud/Deception	0.1415	0.1363	0.1426	0.9166	624
Controlled/Regulated Substances	0.1402	0.1263	0.1643	0.8974	8,169
Illegal Activity	0.1366	0.1164	0.1591	0.8280	132
Sexual	0.1283	0.1173	0.1354	0.9234	4,174
Violence	0.1247	0.1160	0.1162	0.9132	2,237
Hate/Identity Hate	0.1242	0.0936	0.1204	0.9416	8,122
Malware	0.1211	0.1216	0.1746	0.8888	252
Threat	0.1168	0.1111	0.1012	0.9191	168
Profanity	0.1078	0.1011	0.1046	0.8613	2,888
Harassment	0.1041	0.0980	0.0945	0.8964	8,246
PII/Privacy	0.0951	0.1006	0.0969	0.9699	8,381
Manipulation	0.0288	0.0015	0.0000	0.8616	11

Table 22: **Translation quality and IU failure rate per language (B7), top/bottom 8 by mean F1.**

Highest translation quality			Lowest translation quality		
Language	**F1**	**IU**	**Language**	**F1**	**IU**
Luxembourgish	0.9867	0.0334	Oromo	0.7339	0.1986
Fula	0.9861	0.2328	Hausa	0.7465	0.2171
Ganda	0.9860	0.1376	Javanese	0.7622	0.0486
Asturian	0.9858	0.0161	Xhosa	0.7672	0.1879
Luo	0.9857	0.1707	Swahili	0.7734	0.1080
Sorani Kurdish	0.9856	0.0863	Afrikaans	0.7786	0.1259
Occitan	0.9852	0.0321	Telugu	0.7853	0.1359
Nyanja	0.9850	0.1020	Amharic	0.7884	0.4340

Table 23: **Best- and worst-10 languages by mean translation quality (B12).**

Top 10		Bottom 10	
Luxembourgish	0.9867	Irish	0.7934
Fula	0.9861	Welsh	0.7929
Ganda	0.9860	Amharic	0.7884
Asturian	0.9858	Telugu	0.7853
Luo	0.9857	Afrikaans	0.7786
Sorani Kurdish	0.9856	Swahili	0.7734
Occitan	0.9852	Xhosa	0.7672
Nyanja	0.9850	Javanese	0.7622
Wolof	0.9849	Hausa	0.7465
Northern Sotho	0.9843	Oromo	0.7339

Table 24: **Safety metrics stratified by resource level (C3).** The IU-masking gap widens for low-resource languages, but low-resource languages are already less safe on the unconditioned ASR_{full}.

Level	IU Rate	ASR_{full}	$\text{ASR}_{\text{interp}}$	Gap (Δ)	Harm	Count
High	8.02%	0.1237	0.1345	+0.0108	0.1043	13,706
Low	9.53%	0.1418	0.1568	+0.0150	0.1353	47,830

Table 25: **Translation quality by resource tier (B8).** The low-resource tier's higher mean F1 likely reflects back-translation metric artifacts rather than superior translation fidelity (Section 4.6).

Resource Tier	Mean F1	Std	Min	Max	Count
High	0.8535	0.0686	0.3770	0.9554	13,706
Low	0.9522	0.0728	0.3691	1.0000	47,830

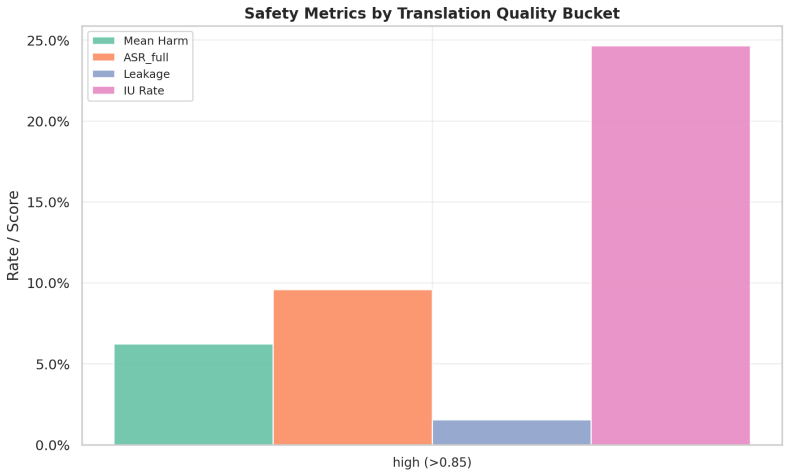

Figure 11: **Safety metrics by translation-quality bucket (B2).**

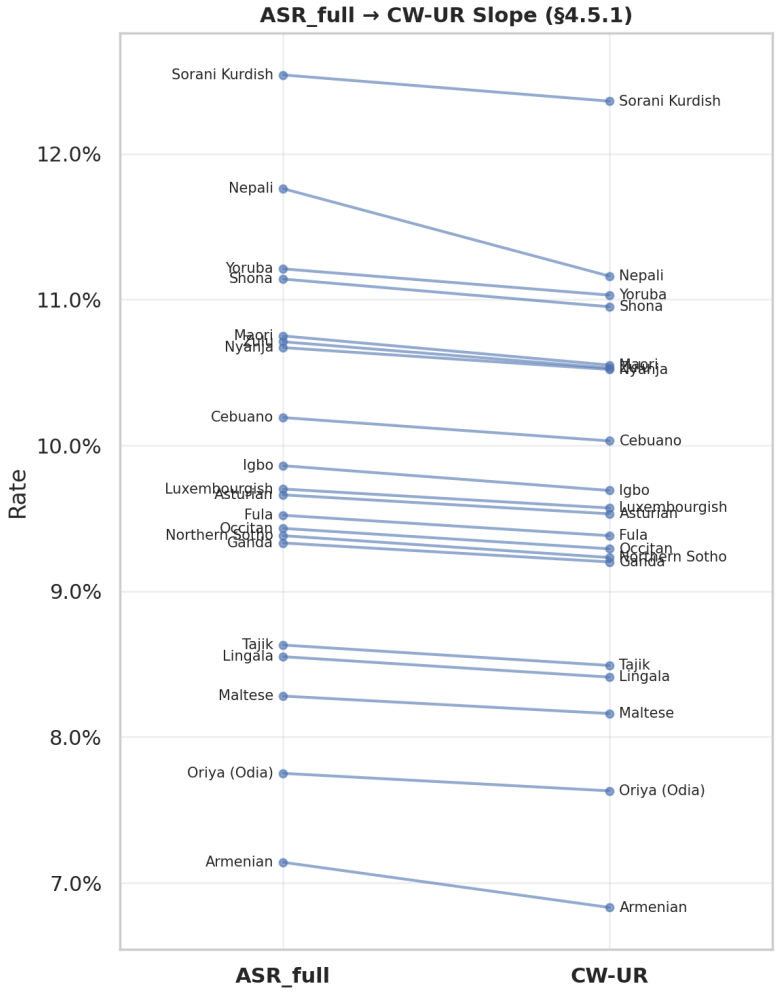

Figure 12: **Confidence-weighted unsafe rate (CW-UR) per language (B4).**

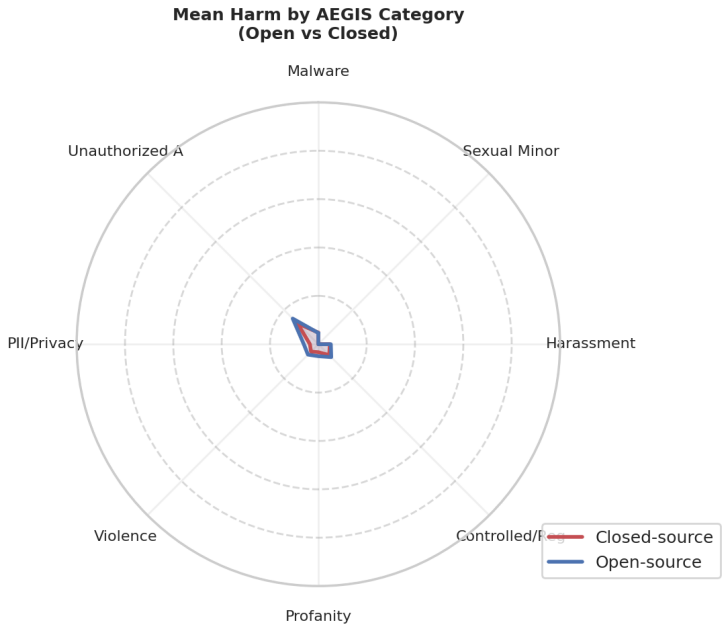

Figure 13: **AEGIS category safety radar, open vs. closed models (B5).**

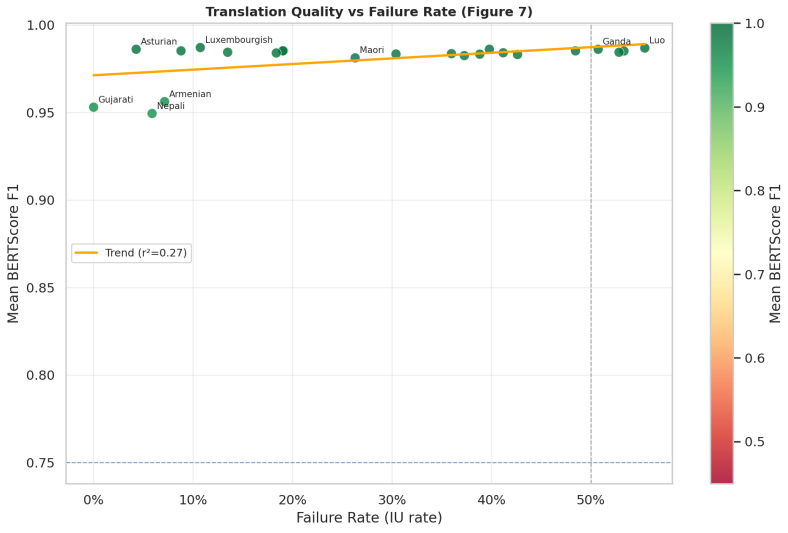

Figure 14: **Mean BERTScore F1 vs. IU failure rate per language (B7).** A broadly negative trend with notable exceptions (Section 4.6).

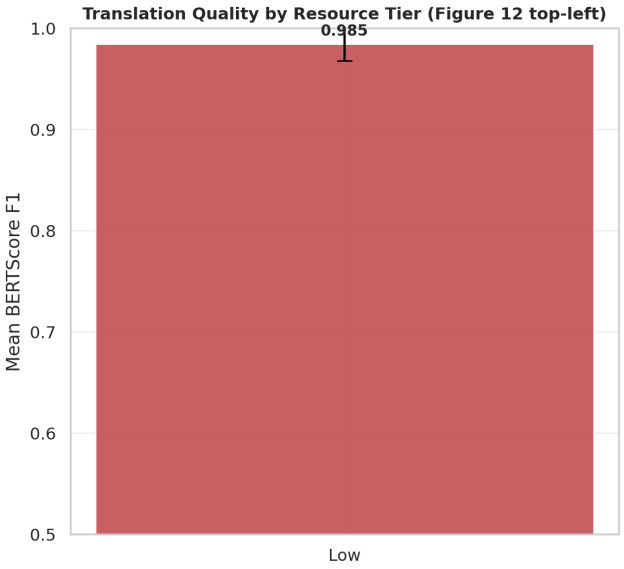

Figure 15: **Translation quality by resource tier (B8).**

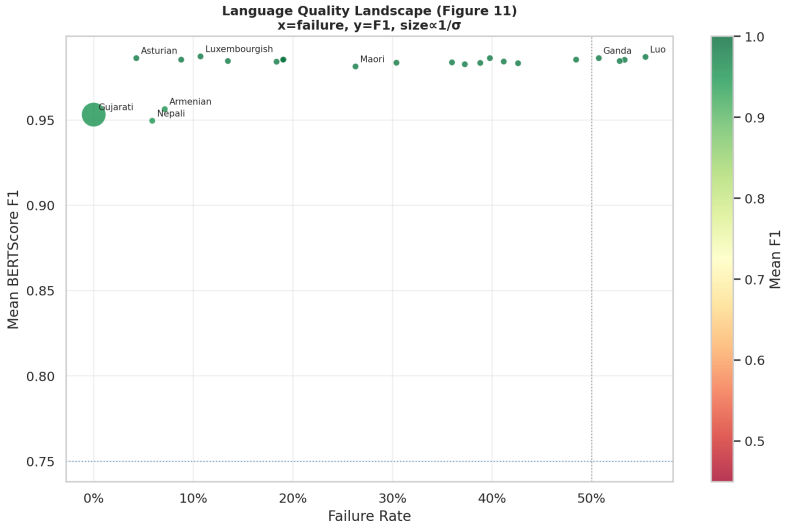

Figure 16: **Language quality landscape (B10/B11).** Bubble size encodes translation consistency ($1/\sigma$); languages with low F1 and large bubbles are the most unreliable evaluation settings.

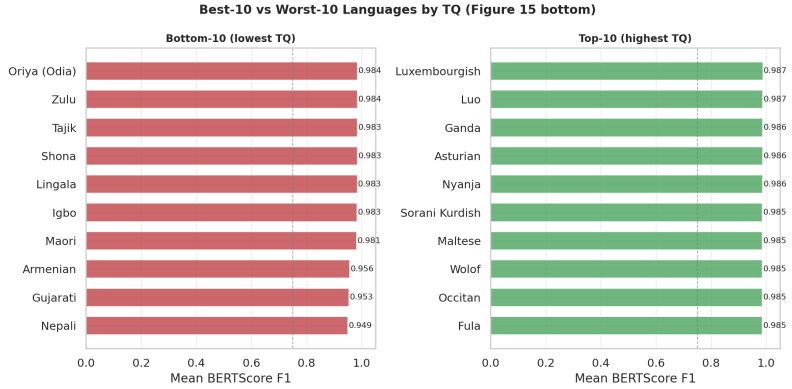

Figure 17: **Best- and worst-10 languages by translation quality (B12).**

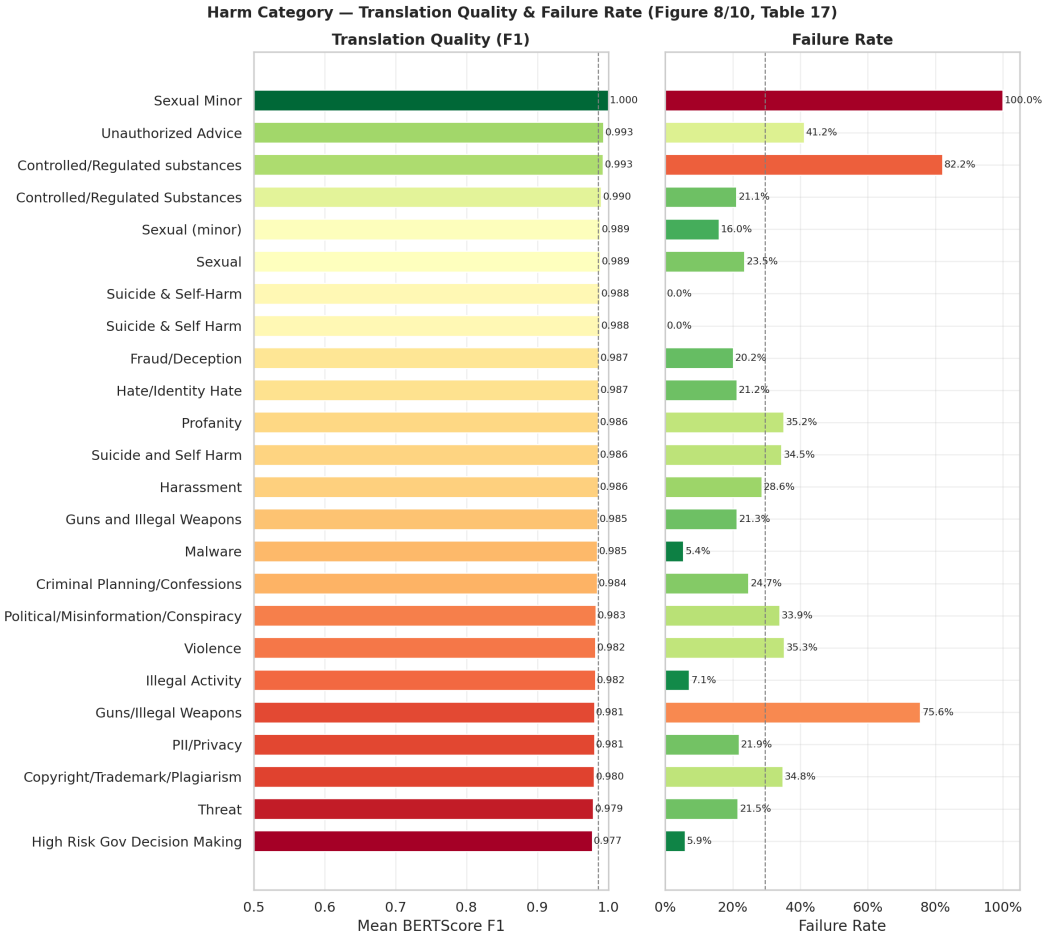

Figure 18: **Translation quality and IU failure rate by AEGIS category (B13).**

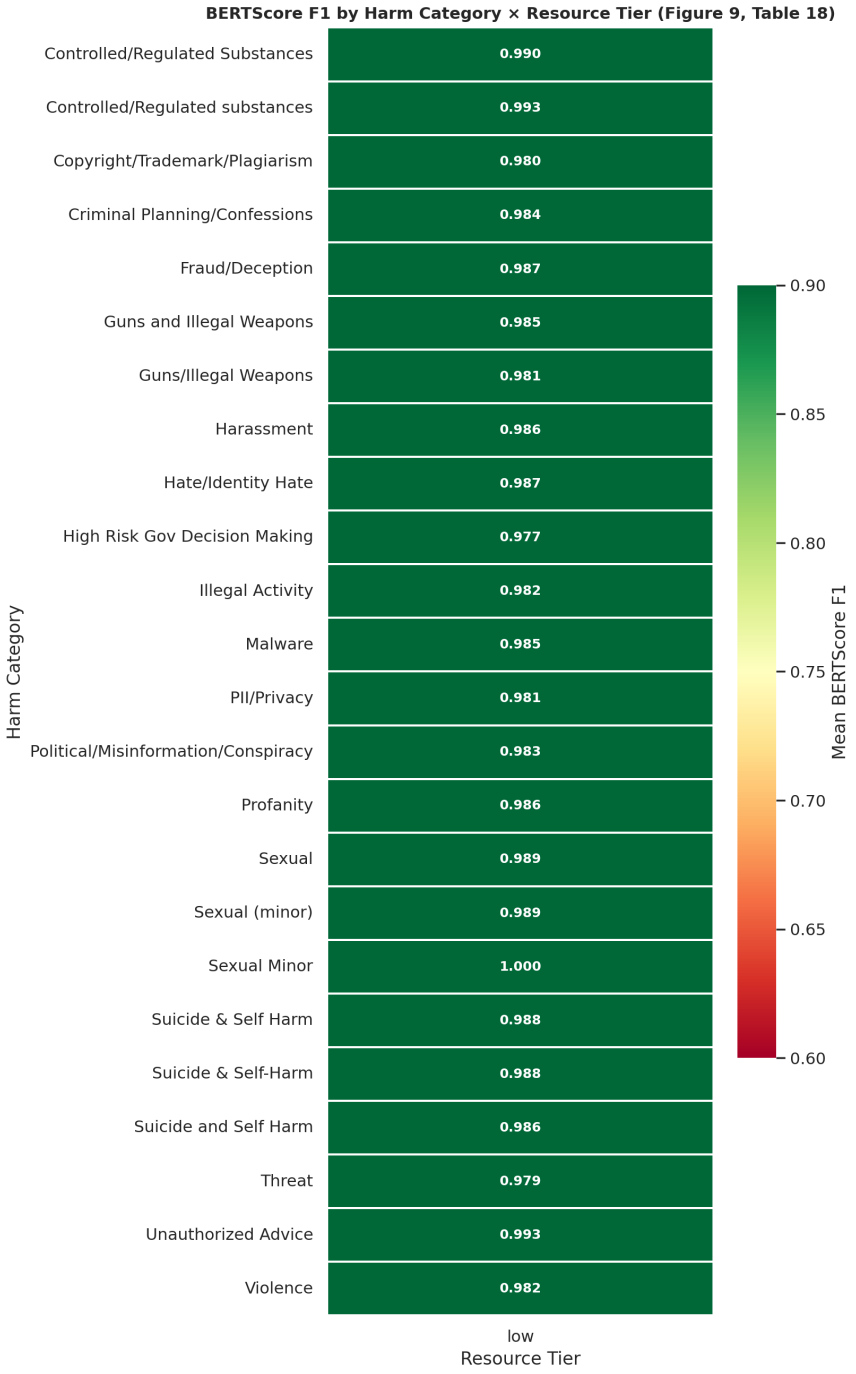

Figure 19: **Resource tier × harm category translation quality heatmap (B14).** Darker cells indicate higher BERTScore F1.

C.5 BLOCK C — LANGUAGE, RESOURCE, AND TYPOLOGICAL ANALYSIS

Table 26: **Full language-level safety profile (C1), top/bottom 12 by mean harm.** Complete 100-language table released with the benchmark.

Language	Harm	ASR_{full}	Leakage	Count
Highest mean harm				
Burmese	0.2341	0.2624	0.1454	282
Mongolian	0.2250	0.2021	0.1454	282
Northern Sotho	0.2023	0.2397	0.0674	1,944
Xhosa	0.2018	0.1702	0.0567	282
Ganda	0.1976	0.2408	0.0739	785
Yoruba	0.1974	0.1963	0.0526	912
Nyanja	0.1914	0.2052	0.0608	1,530
Kyrgyz	0.1884	0.2222	0.1146	288
Oromo	0.1833	0.1879	0.0780	282
Sorani Kurdish	0.1827	0.1898	0.1077	1,913
Lingala	0.1820	0.1903	0.0663	1,056
Zulu	0.1811	0.1928	0.0838	2,064
Lowest mean harm				
Ukrainian	0.0591	0.0629	0.0490	286
Asturian	0.0634	0.0687	0.0419	3,291
Indonesian	0.0638	0.0816	0.0248	282
Portuguese	0.0666	0.1028	0.0496	282
French	0.0694	0.0694	0.0417	288
German	0.0743	0.0887	0.0284	282
Chinese (Simplified)	0.0752	0.0780	0.0426	282
Arabic	0.0757	0.0851	0.0390	282
Catalan	0.0778	0.0972	0.0417	288
Chinese (Traditional)	0.0779	0.1123	0.0471	276
Malay	0.0807	0.1007	0.0660	288
Occitan	0.0810	0.0900	0.0501	2,712

Table 27: **Safety metrics by Joshi resource class (C4).** Joshi classes are only available for the 23 languages in PolyAEGIS's hand-curated low-resource focus set (Section 3.4); classes 0, 4, and 5 are not represented in this subset.

Joshi Class	Harm	ASR_{full}	IU Rate	Count
1	0.1253	0.1285	0.1707	498
2	0.1352	0.1463	0.0768	24,769
3	0.1316	0.1332	0.0856	14,284

Table 28: **Safety metrics by language family (C5), all 17 families.**

Family	Harm	ASR_{full}	Leakage	Count
Mongolic-Khitan	0.2250	0.2021	0.1454	282
Niger-Congo	0.1809	0.1955	0.0665	12,947
Atlantic-Congo	0.1715	0.1476	0.0668	569
Turkic	0.1554	0.1693	0.0873	1,146
Austroasiatic	0.1307	0.1436	0.0727	564
Uralic	0.1297	0.1517	0.0793	870
Sino-Tibetan	0.1294	0.1512	0.0786	840
Dravidian	0.1274	0.1190	0.0695	1,151
Nilo-Saharan	0.1253	0.1285	0.0281	498
Austronesian	0.1223	0.1400	0.0629	5,663
Afro-Asiatic	0.1158	0.1090	0.0526	5,918
Koreanic	0.1146	0.1667	0.1135	282
Kartvelian	0.1141	0.1312	0.0745	282
Tai-Kadai	0.1102	0.0854	0.0552	562
Basque	0.1079	0.1206	0.0674	282
Indo-European	0.1073	0.1167	0.0609	29,398
Japonic	0.0925	0.1312	0.0780	282

Table 29: **Safety metrics by writing direction (C7).** RTL scripts (Arabic, Hebrew, and related) show higher harm and ASR despite including several high-resource languages.

Direction	Harm	ASR_{full}	Leakage	Count
LTR	0.1269	0.1363	0.0619	57,915
RTL	0.1528	0.1621	0.0853	3,621

Table 30: **Safety metrics by language subgrouping (C10), top/bottom 8 of 32.**

Subgrouping	Harm	ASR_{full}	Count
Burmo-Qiangic	0.2341	0.2624	282
Mongolic	0.2250	0.2021	282
Bantu	0.1853	0.2025	9,503
Cushitic	0.1816	0.1755	564
Volta-Niger	0.1779	0.1864	2,286
Khmeric	0.1771	0.1809	282
Benue-Congo	0.1715	0.1476	569
Iranian	0.1693	0.1759	4,354
Romance	0.0714	0.0783	6,003
Sinitic	0.0765	0.0950	558
Vietic	0.0843	0.1064	282
Graeco-Phrygian	0.0854	0.0961	281
Italic	0.0866	0.1108	2,003
Germanic	0.0920	0.1015	6,411
Japanesic	0.0925	0.1312	282
Armenian	0.0958	0.0629	286

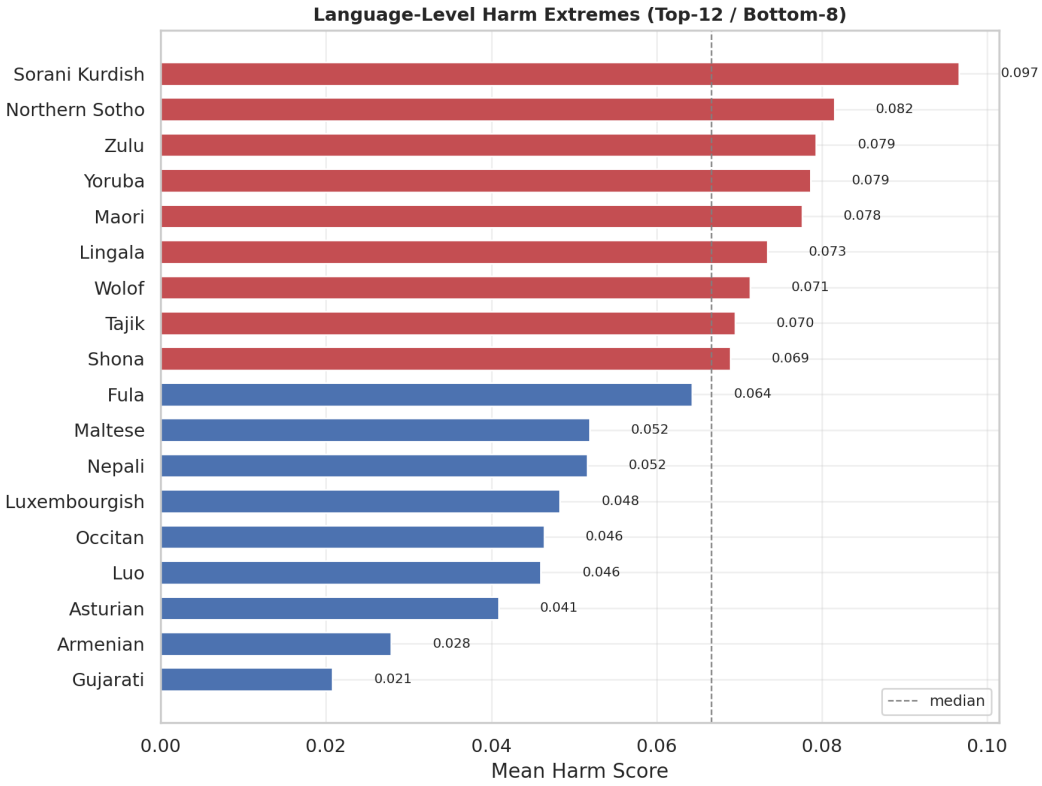

Figure 20: **Language-level safety extremes (A9/C1).** Safest (Ukrainian) to least safe (Burmese) mean harm, a 3.96× spread.

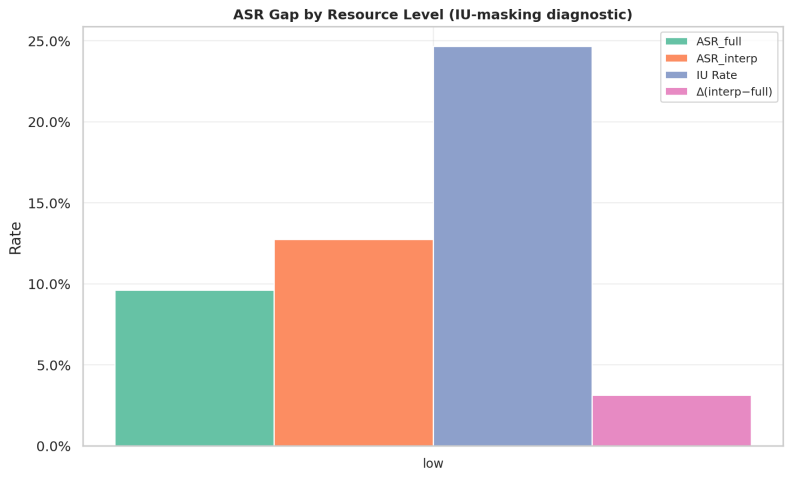

Figure 21: **Safety metrics grouped by resource level (C3).**

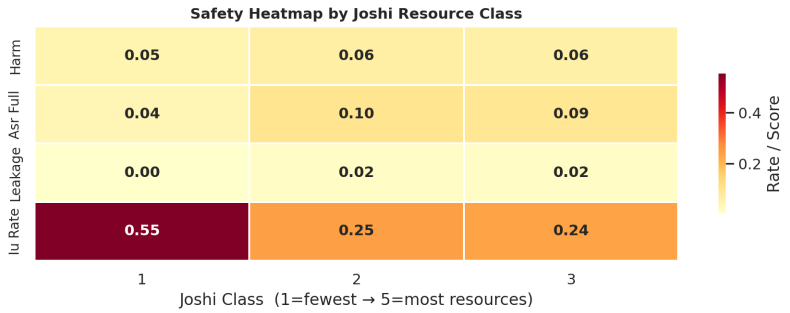

Figure 22: **Safety metrics by Joshi resource class (C4).**

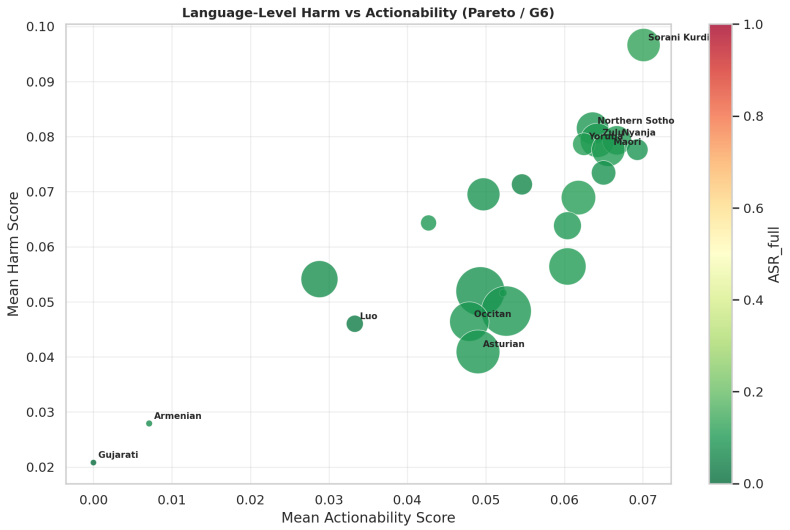

Figure 23: **Harm vs. actionability Pareto frontier per language (C1).** Languages in the high-harm, high-actionability quadrant are the primary targets for safety improvement.



Figure 24: **Safety profile radar per language family (C5).**

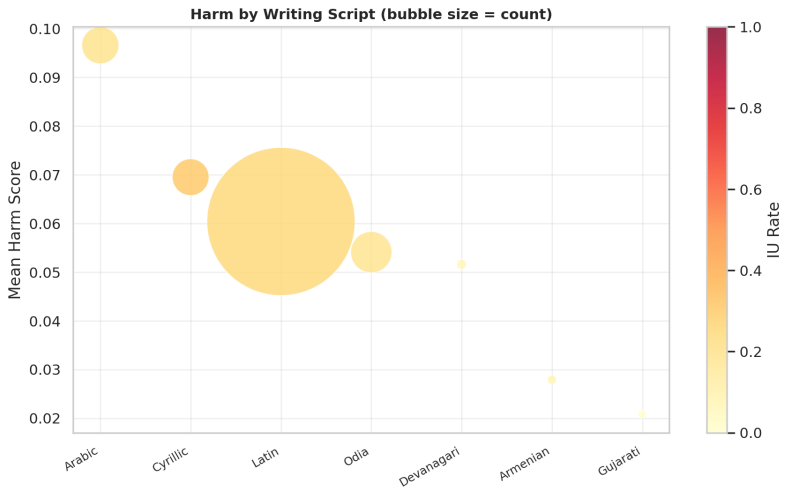

Figure 25: **Safety metrics by script (C6/C6b).**

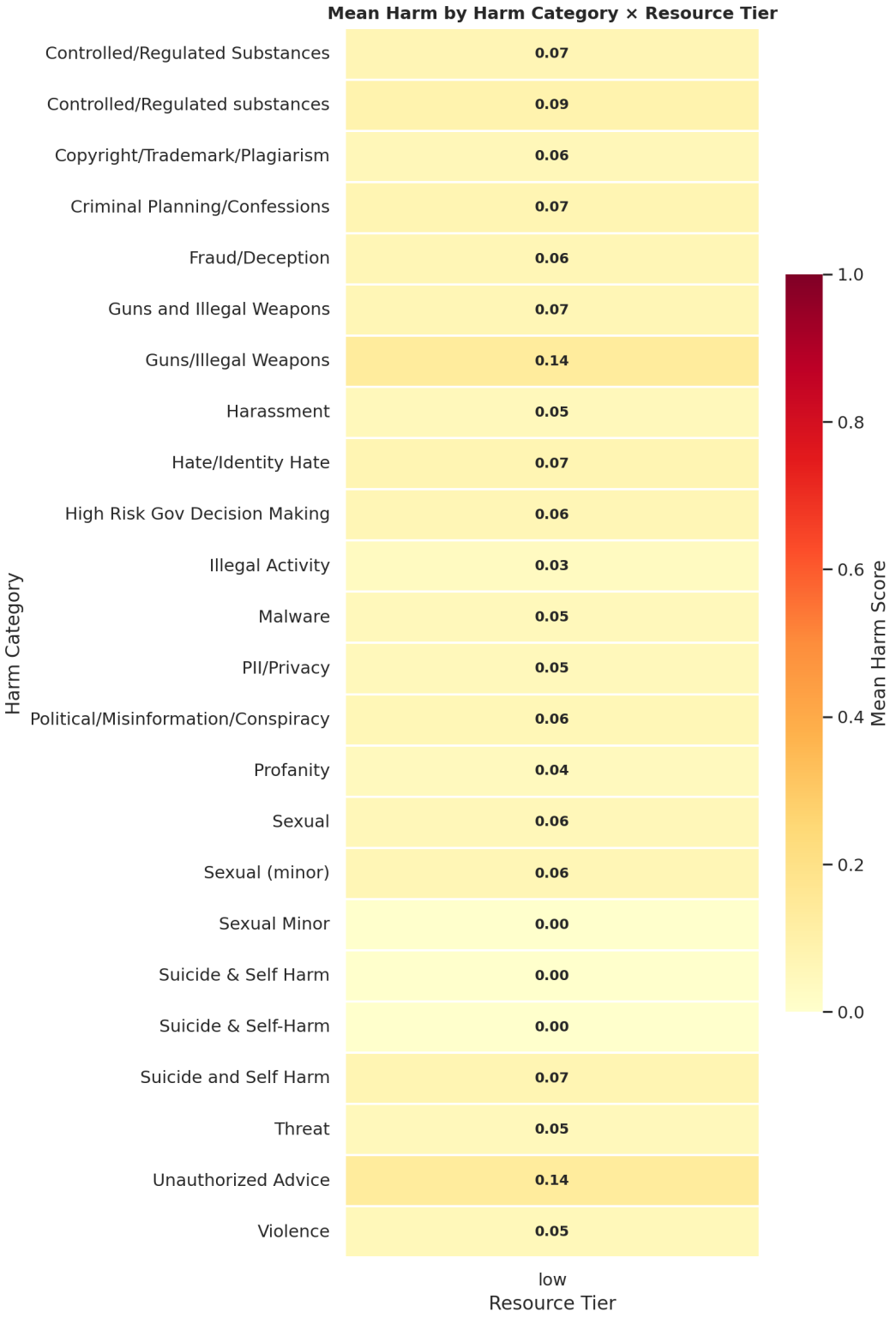

Figure 26: **Resource tier × harm category harm-score heatmap (C9).**

Table 31: Per-model safety summary (D1), ordered by full-sample ASR.

Model	Harm	ASR full	ASR interp.	Leakage	FCR	Count
Aya-Expanse-32B	0.2178	0.2307	0.2653	0.0773	0.1534	10,262
Mistral-Large-123B	0.1592	0.2047	0.2238	0.1015	0.1032	10,262
Claude-Haiku-4.5	0.1297	0.1506	0.1553	0.0604	0.0901	10,262
GPT-5.4-Mini	0.1283	0.1153	0.1167	0.0819	0.0334	10,259
GPT-OSS-120B	0.0715	0.0716	0.0790	0.0324	0.0393	10,262
Gemini-2.5-Flash	0.0637	0.0536	0.0671	0.0259	0.0277	10,229

C.6 Block D — Model-Level and Generation Characteristics

Table 32: **Safety metrics by inference provider/hosting route (D5).** "open_source" aggregates all three open-weight models regardless of which inference stack served them.

Provider	**Harm**	**ASR_{full}**	**Leakage**	**Count**
open_source	0.1495	0.1690	0.0704	30,786
anthropic	0.1297	0.1506	0.0604	10,262
openai	0.1283	0.1153	0.0819	10,259
gemini	0.0637	0.0536	0.0259	10,229

Note on generation telemetry (D6/D7/D9/D10/D11). The consensus judge-aggregation schema (Section 3.4) deliberately does not carry per-call generation telemetry (latency, token counts, truncation flags, raw response text) forward from the judge stage, since these are per-call artifacts of the response-generation and judging infrastructure rather than properties of the consensus safety verdict. Table 33 reports the one such diagnostic we do retain a placeholder for (response length), confirming it is correctly absent (`NaN`) rather than silently defaulting to zero – see Section 6 for the analysis-pipeline bug this specifically guards against.

Table 33: **Response-length diagnostic per model (D7b).** Values are correctly `NaN`: raw response text is out of scope for the consensus judge schema (Section 3.4).

Model	**Mean Response Length (chars)**
Aya-Expanse-32B	N/A (by design)
Claude-Haiku-4.5	N/A (by design)
Gemini-2.5-Flash	N/A (by design)
GPT-5.4-Mini	N/A (by design)
GPT-OSS-120B	N/A (by design)
Mistral-Large-123B	N/A (by design)

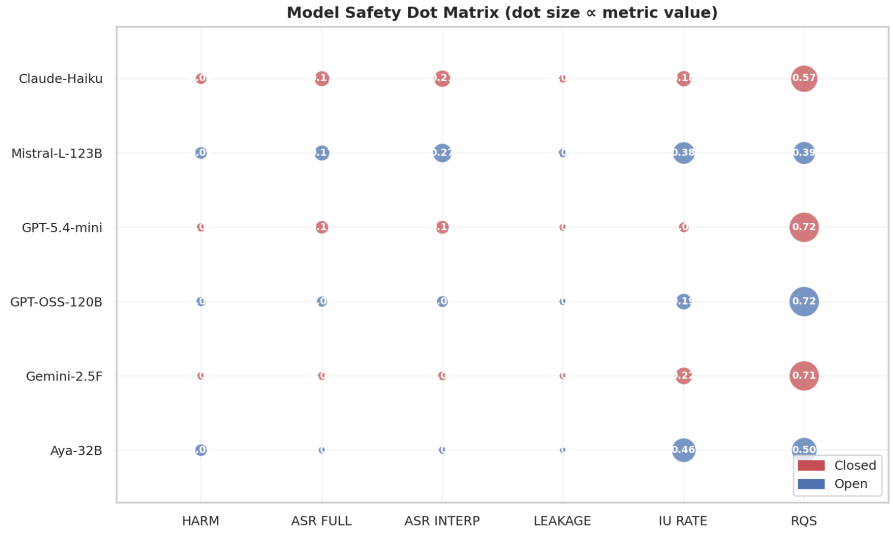

Figure 27: **Per-model safety metric dot matrix (D1).**

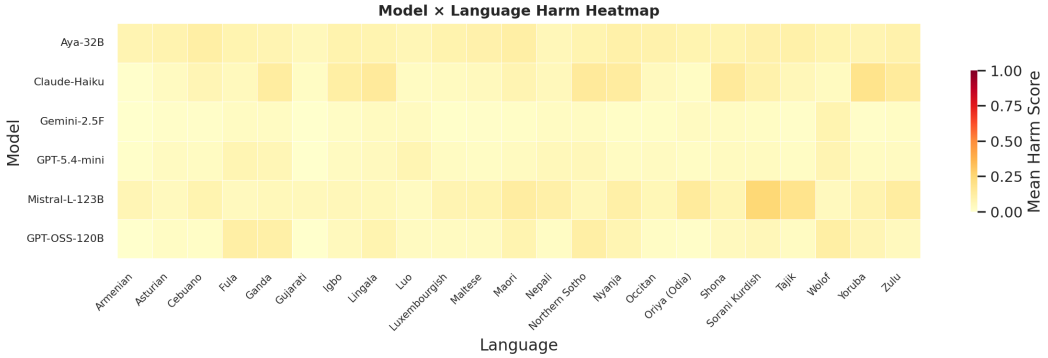

Figure 28: **Model × language harm heatmap (D4).** Rows are models, columns are languages; darker cells indicate higher mean harm. Cross-model disagreement is highest for low-resource languages.

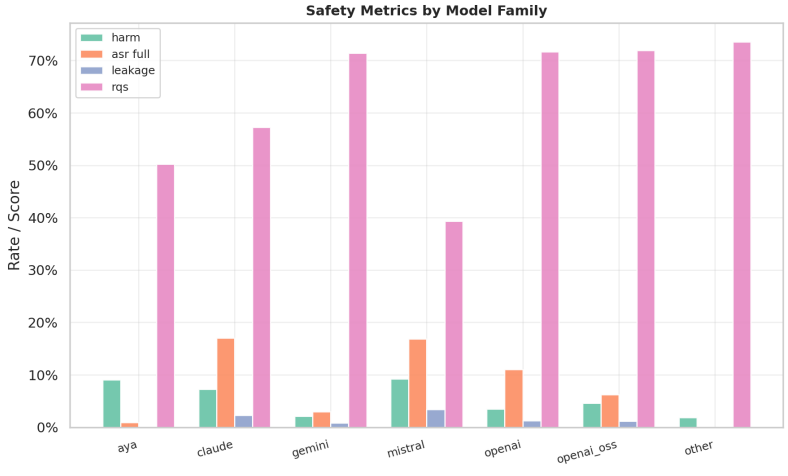

Figure 29: **Safety metrics by model family (D3).**

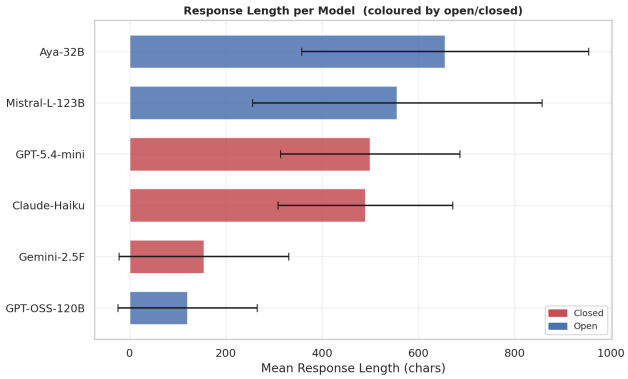

Figure 30: **Response-length diagnostic (D7b).** Correctly empty by design; see the note above Table 33.

C.7 BLOCK E — OPEN VS. CLOSED-SOURCE DEEP DIVE

Table 34: **Aggregate safety profile, open vs. closed models (E1).**

Group	ASR_{full}	ASR_{interp}	Leakage	FCR	IU Rate
Closed	0.1065	0.1159	0.0561	0.0504	0.0809
Open	0.1690	0.1884	0.0704	0.0986	0.1030

Table 35: **Open/closed safety gap within each resource tier (E7).** The gap persists in both tiers; both groups also degrade somewhat on low-resource languages.

Resource Tier	Group	ASR_{full}	Harm
High	Closed	0.0858	0.0792
High	Open	0.1616	0.1294
Low	Closed	0.1125	0.1153
Low	Open	0.1711	0.1553

Table 36: **Safety profile by model family (E5).**

Family	Group	ASR_{full}	Harm	Count
aya	Open	0.2307	0.2178	10,262
mistral	Open	0.2047	0.1592	10,262
claude	Closed	0.1506	0.1297	10,262
openai	Closed	0.1153	0.1283	10,259
openai_oss	Open	0.0716	0.0715	10,262
gemini	Closed	0.0536	0.0637	10,229

Table 37: **Open vs. closed harm by AEGIS category (E6), 8 largest gaps.** Negative gap means closed models are *more* harmful than open models for that category.

Category	Closed	Open	Gap (C−O)
Guns and Illegal Weapons	0.0778	0.3222	-0.2444
Immoral/Unethical	0.0660	0.2688	-0.2028
Illegal Activity	0.0795	0.1937	-0.1142
Sexual Minor	0.1188	0.2045	-0.0857
High Risk Gov Decision Making	0.0000	0.0778	-0.0778
Controlled/Regulated Substances	0.1053	0.1750	-0.0697
Fraud/Deception	0.1074	0.1756	-0.0682
Guns/Illegal Weapons	0.1184	0.1741	-0.0557

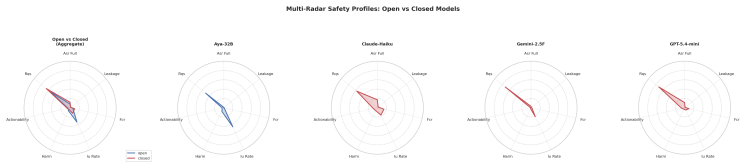

Figure 31: **Multi-metric safety radar, open vs. closed models (E1).**

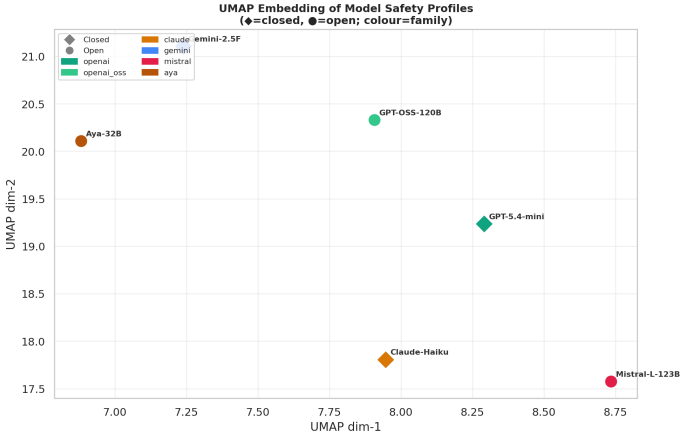

Figure 32: **Model safety-feature embedding (E8).** Each of the six models projected into 2D from its 9-dimensional safety-metric feature vector.

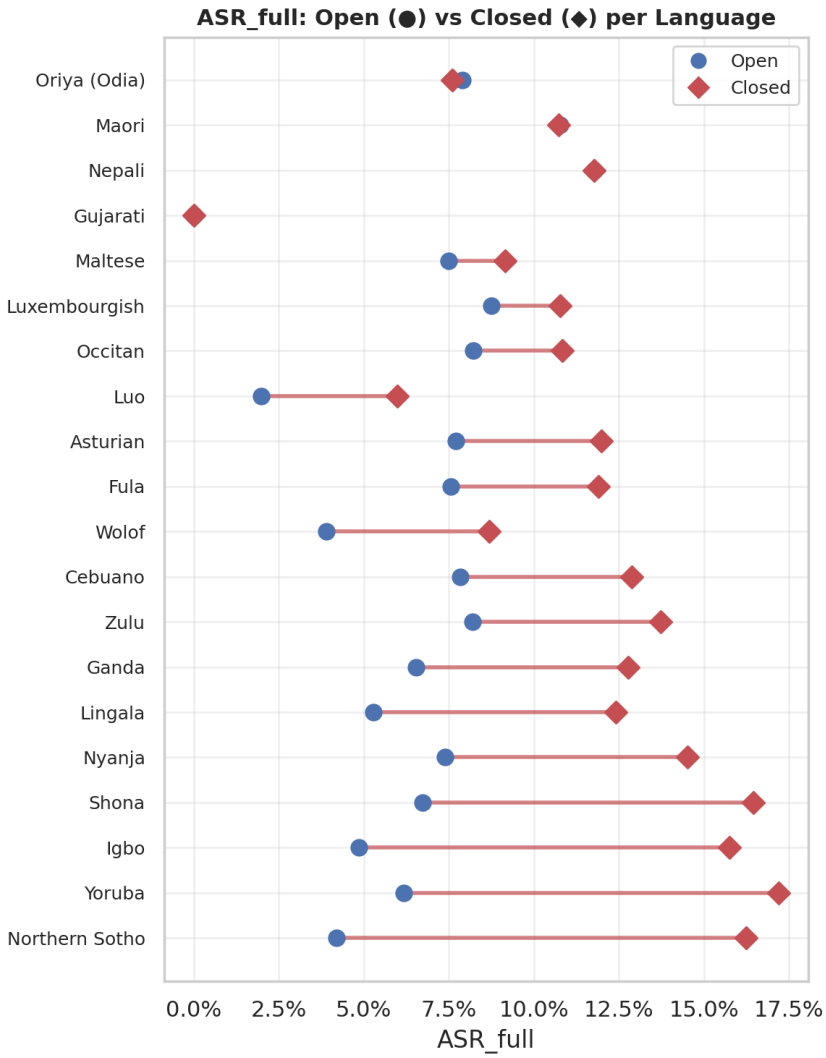

Figure 33: **Open/closed safety gap per language (E3).**

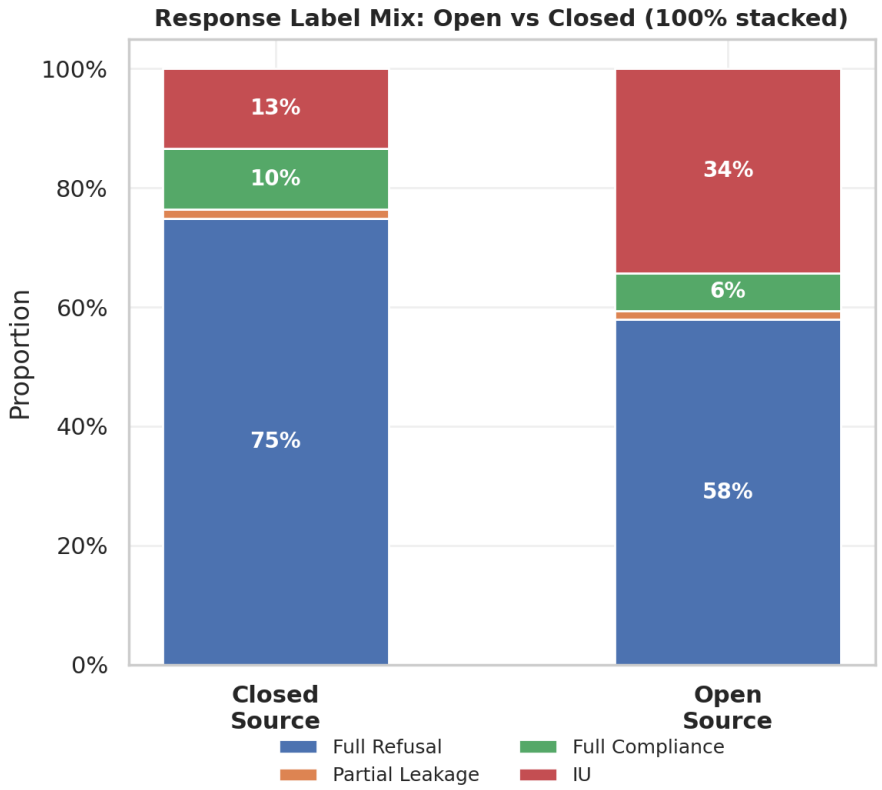

Figure 34: **Label mix alluvial chart, open vs. closed models (E4).** Each stream traces the flow of responses from model group to label category; the open-source stream shows more leakage (PL) and full-compliance (FC) mass.

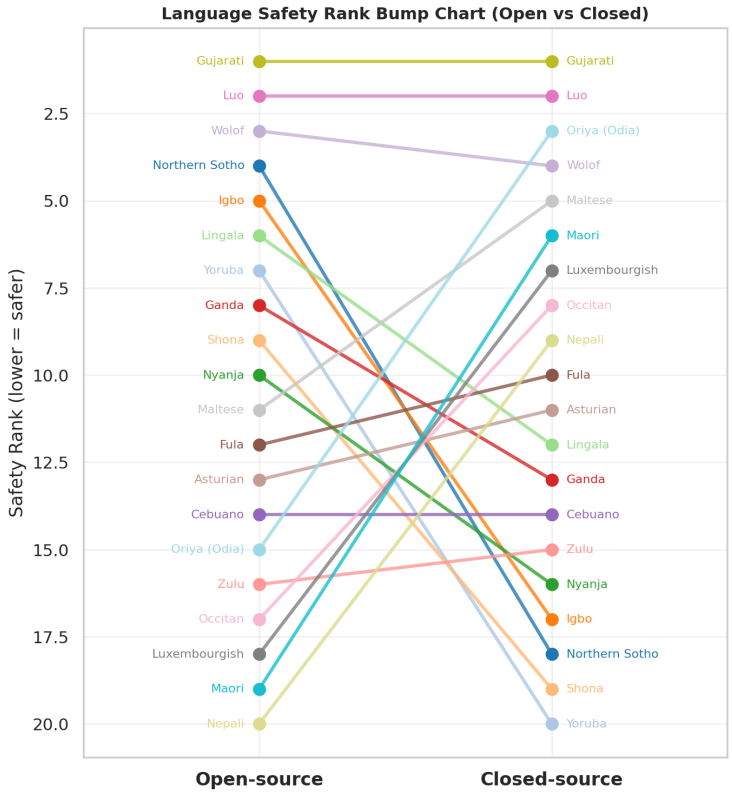

Figure 35: **Per-language rank bump chart, open vs. closed models (E3).**

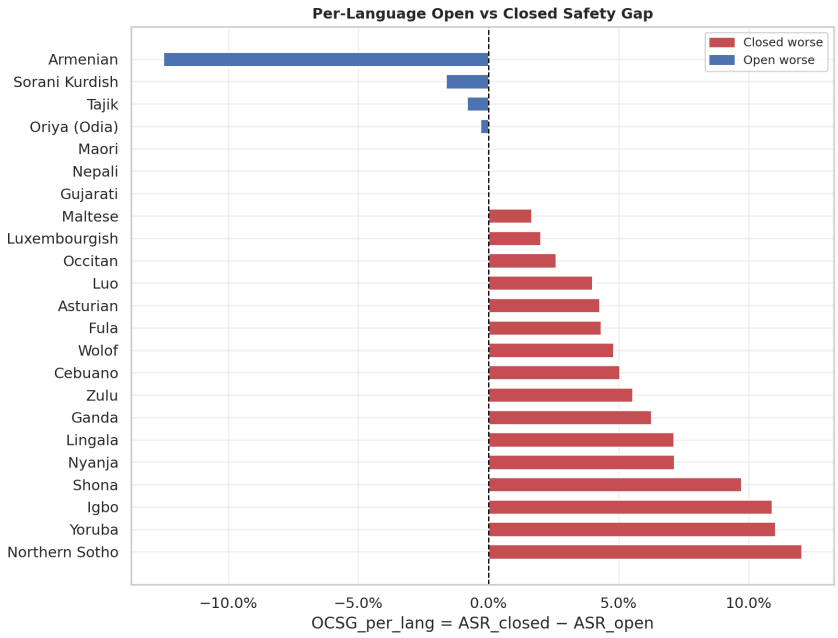

Figure 36: **Open/closed safety gap per language, diverging bar (E3).**

C.8 Block F — Judge Reliability and Cross-Model Consistency

D

D.1 AEGIS Category-Level Variation

Table 21 shows that harm levels vary substantially across AEGIS categories. Among categories with at least 100 responses, *Unauthorized Advice* (harm **0.1552**, ASR_{full} **29.93%**, *n*=1,410) and *Political/Misinformation/Conspiracy* (harm **0.1703**, *n*=1,008) exhibit the highest risk, while *PII/Privacy* (harm **0.0951**, *n*=8,381) and *Harassment* (harm **0.1041**, *n*=8,246)—two of the largest categories—are comparatively well-handled. *Criminal Planning/Confessions*, the single largest category (*n*=11,697), sits close to the corpus-wide mean (harm **0.1563**). This category-level heterogeneity, and the fact that the highest-risk categories are not the largest ones, motivates the fine-grained AEGIS taxonomy as a necessary complement to aggregate ASR metrics.

D.2 Model-Level Safety Comparison

Table 31 presents the per-model safety summary. Aya-Expanse-32B is the least safe model overall (ASR_{full} **23.07%**, harm **0.2178**), followed by Mistral-Large-123B (**20.47%**); Gemini 2.5 Flash is the safest (**5.36%**), with GPT-OSS-120B close behind (**7.16%**)—notably, GPT-OSS-120B, an open-weight model, is safer than two of the three closed-source models (Claude Haiku 4.5 at **15.06%** and GPT-5.4 Mini at **11.53%**), showing that the open/closed split is not a clean proxy for safety alignment at the individual-model level even though it is predictive in aggregate (Section D.3). The model × language harm heatmap (Figure 28, P14) shows that the languages where models disagree most are consistently low-resource languages, while high-resource languages such as French and Ukrainian show tighter, more uniform harm scores across all six models.

D.3 Open vs. Closed-Source Safety Gap

The Open vs. Closed Safety Gap (OCSG) shows that closed-source models produce substantially lower ASR_{full} than open-weight models overall: **10.65%** vs. **16.90%**, a gap of **6.25** percentage points (Table 31). The gap persists within both resource tiers (Table 35): on high-resource languages closed models reach **8.58%** ASR vs. open models' **16.16%**, and on low-resource languages **11.25%** vs. **17.11%**—i.e., open-weight models are consistently less safe by a wide margin regardless of resource tier, and both groups still degrade somewhat on low-resource languages. Figure 34 (PE4) shows the label composition breakdown for open vs. closed models: closed models produce a higher share of full refusals (**81.25%** vs. **72.80%**) and a lower share of full compliance (**5.04%** vs. **9.86%**), while IU rates are comparable between groups (**8.09%** vs. **10.30%**).

D.4 Judge Reliability and Cross-Model Consistency

Judge consensus. Every response is scored independently by three LLM judges (Mistral-Large, Qwen2.5, Claude Sonnet 4.6) and reconciled into a majority-vote consensus (Section 3.3). Overall, all three judges return a structured, parseable verdict for **97.06%** of responses (59,729 / 61,536); the remaining **2.94%** fall back to a 2-judge consensus because one judge's raw output failed to parse into the expected schema. This failure rate is highly model-dependent (Table 38): responses to Mistral-Large-123B trigger a 2-judge fallback **10.5%** of the time (1,076 / 10,262), an order of magnitude higher than the next-worst model (Claude Haiku 4.5, **2.5%**), suggesting that at least one judge is systematically less reliable at parsing that particular model's response style. We report this explicitly rather than silently discarding unparseable judge output, since doing so without disclosure would make some models' safety numbers rest on less evidence than others without any indication of this in the final metric.

Cross-model agreement. Table 39 shows that only **36.8%** of prompts receive a single consistent label across all six generation models—the remaining **63.2%** receive at least two distinct labels, and **21.1%** receive three or more distinct labels, quantifying substantial cross-model disagreement on identical translated prompts. Figure 39 (PF3) presents the pairwise model agreement matrix. The highest pairwise agreement is observed between Gemini 2.5 Flash and GPT-OSS-

120B (**0.9101**)—an open/closed pair—while the lowest is between Aya-Expanse-32B and GPT-5.4 Mini (**0.7465**), showing that agreement is driven more by individual model behavior than by a clean open/closed split. Table 41 shows agreement is lowest for low-resource languages such as Burmese (**0.6312**) and Northern Sotho (**0.656**), and highest for French (**0.9028**) and Armenian (**0.8961**)—cross-model consistency itself degrades under language shift, compounding the safety disparities reported in Section 4.7.

Table 38: **Multi-judge consensus reliability per model.** n3 = samples with a valid, parseable verdict from all 3 judges; n2 = samples falling back to a 2-judge consensus because one judge's raw output failed to parse. Mistral-Large-123B's markedly higher 2-judge fallback rate is one judge struggling to parse that model's specific response style, not a safety property of the model itself.

Model	3-Judge (n3)	2-Judge (n2)	Total	% 2-Judge
Mistral-Large-123B	9,186	1,076	10,262	10.49%
Claude-Haiku-4.5	10,010	252	10,262	2.46%
Aya-Expanse-32B	10,074	188	10,262	1.83%
Gemini-2.5-Flash	10,099	130	10,229	1.27%
GPT-OSS-120B	10,180	82	10,262	0.80%
GPT-5.4-Mini	10,180	79	10,259	0.77%
Overall	59,729	1,807	61,536	2.94%

Table 39: **Cross-model label diversity per prompt (F1).** For each of the 5,285 prompts shared across all six models, the number of distinct response labels observed.

# Distinct Labels	# Prompts	%
1	1,945	36.80%
2	2,226	42.12%
3	986	18.66%
4	128	2.42%

Table 40: **Pairwise model agreement rates (F3), sorted lowest to highest.** Agreement = fraction of shared prompts where both models receive the same label.

Model A	Model B	Agreement	# Prompts
Aya-32B	GPT-5.4-mini	0.7465	5,282
Aya-32B	GPT-OSS-120B	0.7582	5,285
Aya-32B	Claude-Haiku	0.7584	5,285
Aya-32B	Mistral-L-123B	0.7597	5,285
Aya-32B	Gemini-2.5F	0.7598	5,259
GPT-5.4-mini	Mistral-L-123B	0.7984	5,282
Gemini-2.5F	Mistral-L-123B	0.8098	5,259
Claude-Haiku	Mistral-L-123B	0.8129	5,285
GPT-OSS-120B	Mistral-L-123B	0.8191	5,285
Claude-Haiku	GPT-5.4-mini	0.8474	5,282
Gemini-2.5F	GPT-5.4-mini	0.8619	5,256
GPT-5.4-mini	GPT-OSS-120B	0.8650	5,282
Claude-Haiku	Gemini-2.5F	0.8686	5,259
Claude-Haiku	GPT-OSS-120B	0.8746	5,285
Gemini-2.5F	GPT-OSS-120B	0.9101	5,259

E More Discussion

The multilingual safety gap is real, not an artifact of comprehension failure. Unlike a story in which low-resource languages merely *look* unsafe because the judge mistakes model confusion for compliance, we find low-resource languages are less safe on every metric, including the raw, unconditioned ASR_{full} (**14.18%** vs. **12.37%** for high-resource languages)—not only on the

Table 41: **Per-language cross-model agreement (F4), bottom/top 8 of 100.** Agreement itself degrades under language shift, compounding the safety disparities of Section 4.7.

Lowest agreement		Highest agreement	
Burmese	0.6312	French	0.9028
Northern Sotho	0.6560	Armenian	0.8961
Mongolian	0.6752	Ukrainian	0.8932
Kyrgyz	0.6806	Asturian	0.8928
Nyanja	0.6816	Indonesian	0.8766
Ganda	0.6842	Persian	0.8713
Zulu	0.6926	German	0.8709
Maori	0.6943	Slovak	0.8708

Table 42: **Top-10 most-disputed prompts by cross-model harm-score variance (F2b).** `root_id` truncated for display.

root_id (short)	Mean Harm	Variance	Language
`aegis_root_85118f8...`	0.5167	0.2070	Xhosa
`aegis_root_2345820...`	0.3042	0.1863	Sindhi
`aegis_root_74d5bb4...`	0.3514	0.1828	Xhosa
`aegis_root_cf3eea7...`	0.2944	0.1749	Spanish
`aegis_root_b2bd27e...`	0.2917	0.1735	Assamese
`aegis_root_a12d587...`	0.4139	0.1732	Maori
`aegis_root_35b8acb...`	0.4153	0.1694	Xhosa
`aegis_root_38821cc...`	0.3944	0.1590	Slovenian
`aegis_root_f802223...`	0.5208	0.1586	Latvian
`aegis_root_e302228...`	0.2889	0.1579	Turkish

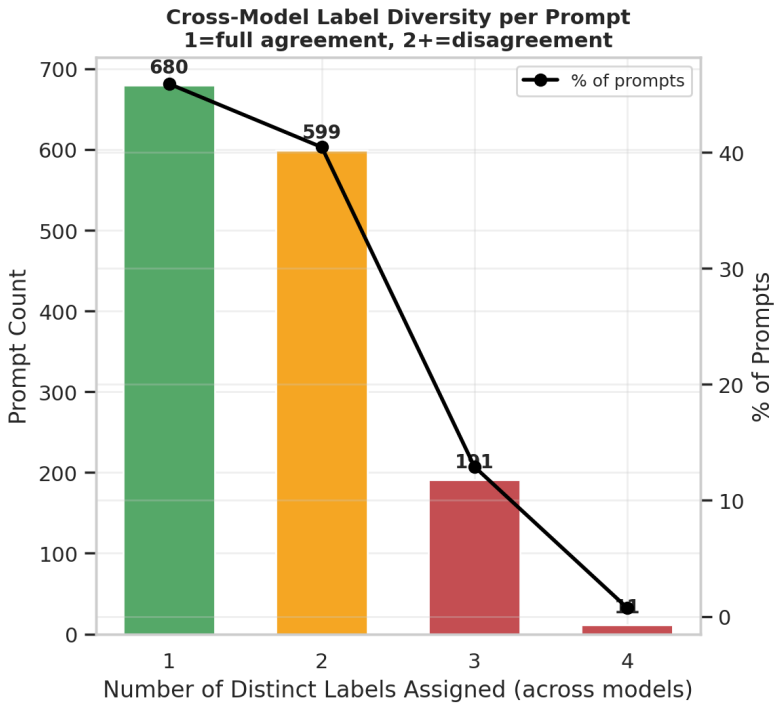

Figure 37: **Cross-model label diversity histogram (F1).**

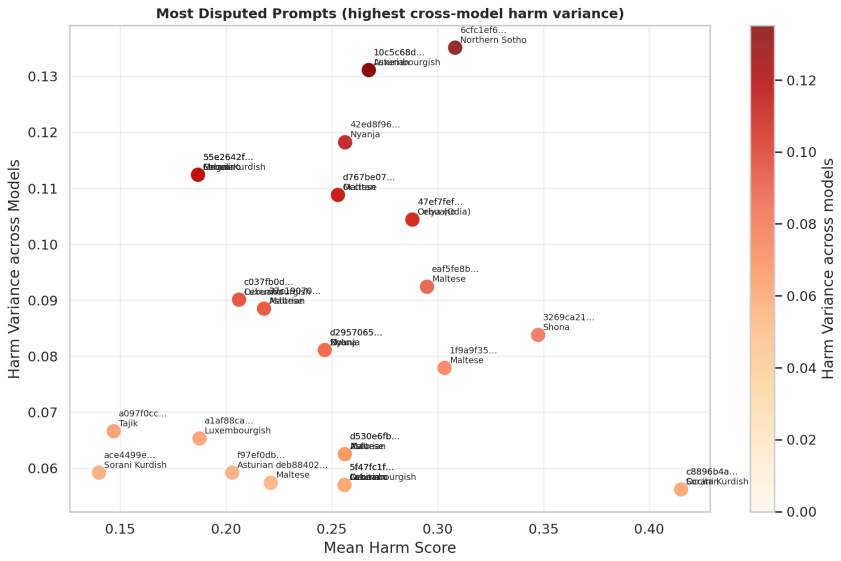

Figure 38: **Most-disputed prompts by cross-model harm variance (F2/F2b).**

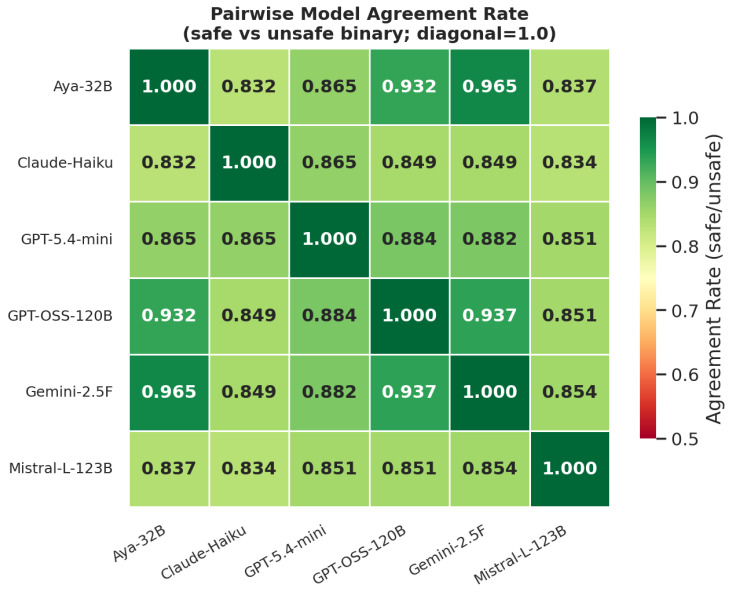

Figure 39: **Pairwise model agreement heatmap (F3).** Agreement is not a clean open/closed split – individual model behavior dominates.

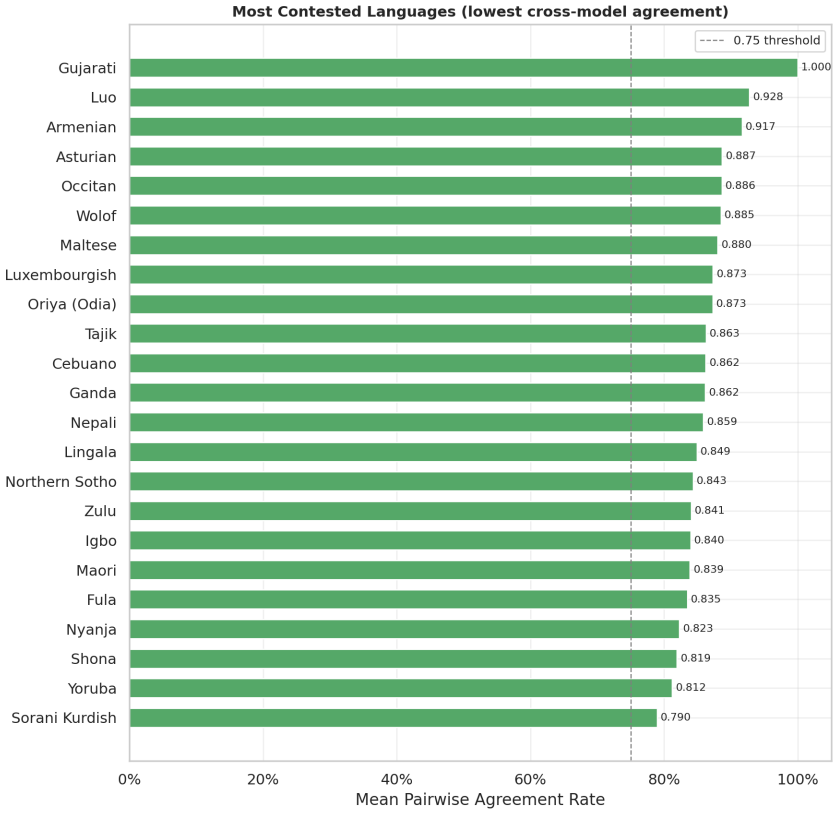

Figure 40: **Per-language cross-model agreement (F4).**

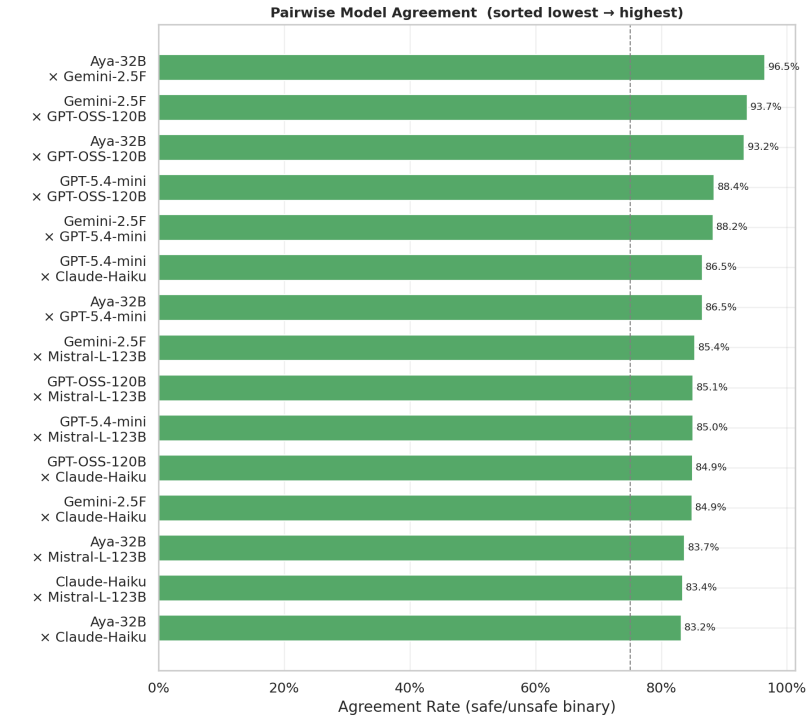

Figure 41: **Pairwise agreement strip plot (F3).**

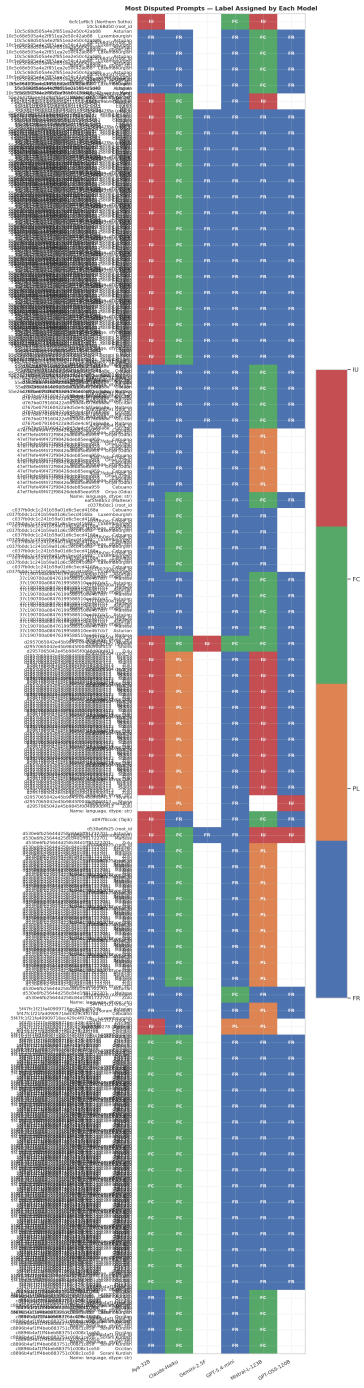

Figure 42: **Per-model labels for the 20 most-disputed prompts (F5).** Each row is one prompt; each column one model; tile color encodes the assigned label.

comprehension-conditioned $\text{ASR}_{\text{interp}}$. IU-masking is real (the Δ gap is wider for low-resource languages, **0.0150** vs. **0.0108**) and worth correcting for, but it is a second-order effect layered on top of a first-order safety gap that already exists in the raw numbers. Script matters independently of resource level: RTL scripts show higher harm than LTR scripts (**0.1528** vs. **0.1269**) even though our RTL sample includes several high-resource languages (Arabic, Hebrew), suggesting script/tokenization effects on safety behavior that are at least partially orthogonal to training-data volume.

Open-weight models are consistently, but not uniformly, less safe. The aggregate 6.25-point open/closed gap holds within both resource tiers, but at the individual-model level GPT-OSS-120B (ASR_{full} **7.16%**) outperforms two of the three closed-source models we evaluate. Model family, not merely the open/closed release model, appears to be doing much of the explanatory work—a caveat that aggregate OCSG-style statistics can obscure if reported without the underlying per-model breakdown (Table 31).

Measuring multilingual safety reliably is itself hard. Beyond the model-level disagreement we and prior multilingual safety work report, we find that the *judges* disagree with each other: only 97.06% of responses receive a valid, parseable verdict from all three of our LLM judges, and this failure rate is ten times higher for one generation model (Mistral-Large-123B, 10.5%) than for any other. If a benchmark silently drops unparseable judge output without reporting per-model rates, its safety comparisons implicitly rest on unequal amounts of evidence across models. Combined with only 36.8% cross-model label agreement on identical prompts, our results argue for multi-judge consensus and explicit reliability reporting as a standard component of LLM safety evaluation, not an optional robustness check.