SAFETY ALIGNMENT UNDER LANGUAGE SHIFT: HARMFUL COMPLIANCE, OVER-REFUSAL, AND RESPONSE-LANGUAGE DYNAMICS IN LOW-RESOURCE LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model safety alignment is usually evaluated in high-resource languages, leaving low-resource settings underexplored. We evaluate proprietary and open-source models across English, Mandarin Chinese, Cantonese, and Tibetan using benign and borderline prompts for false refusal and AdvBench for harmful compliance. We additionally measure off-topic output, repetition, and target-language response rate, and manually verify ASR-relevant labels. For every proprietary model with paired data, harmful compliance is higher in Cantonese and Tibetan than in its English–Mandarin baseline, whereas false refusal changes heterogeneously. Under unconstrained harmful prompting, Cantonese response-language consistency decreases substantially for several models, showing that aggregate safety metrics and response-language behavior should be interpreted jointly. Explicit target-language instructions largely restore language consistency but do not consistently improve safety. Among open-source models, low ASR can reflect off-topic generation or repetition, with off-topic behavior occurring across both benign and harmful datasets. Multilingual evaluation should jointly distinguish harmful compliance, false refusal, relevance, validity, and response-language behavior.

1 INTRODUCTION

Large language models (LLMs) are increasingly used across linguistically diverse communities, but their safety alignment does not transfer uniformly across languages. Prior work has shown that safeguards trained primarily on high-resource data can weaken in lower-resource settings (Yong et al., 2023; Deng et al., 2024; Wang et al., 2024; 2025; Zhao et al., 2025). Safety evaluation must also account for *over-refusal*, where benign or borderline requests are incorrectly rejected (Röttger et al., 2024; Cui et al., 2025). However, multilingual jailbreak robustness and over-refusal are usually evaluated separately, and aggregate rates can conceal different mechanisms: a low attack success rate (ASR) may reflect principled refusal, but also off-topic generation, repetition, or failure to respond in the user's language. We therefore study harmful compliance, false refusal, response relevance, and response-language consistency jointly.

Prior work has established that safeguards can weaken under lower-resource-language input, but less is known about the behavioral composition of this safety gap. Higher harmful compliance may reflect weaker refusal, whereas low harmful compliance may instead result from poor instruction following, irrelevant generation, or repetition. False refusal and response-language inconsistency may also vary independently from harmful compliance. We therefore ask three questions: whether harmful compliance and false refusal shift together across languages, whether non-harmful outputs represent principled refusal or capability-related failure, and whether enforcing target-language use also improves safety.

We evaluate proprietary and open-source LLMs in English, Mandarin Chinese, Cantonese, and Tibetan. AdvBench (Zou et al., 2023) measures harmful compliance, while multilingual adaptations of XSTest (Röttger et al., 2024) and OR-Bench Hard (Cui et al., 2025) measure false refusal on

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.

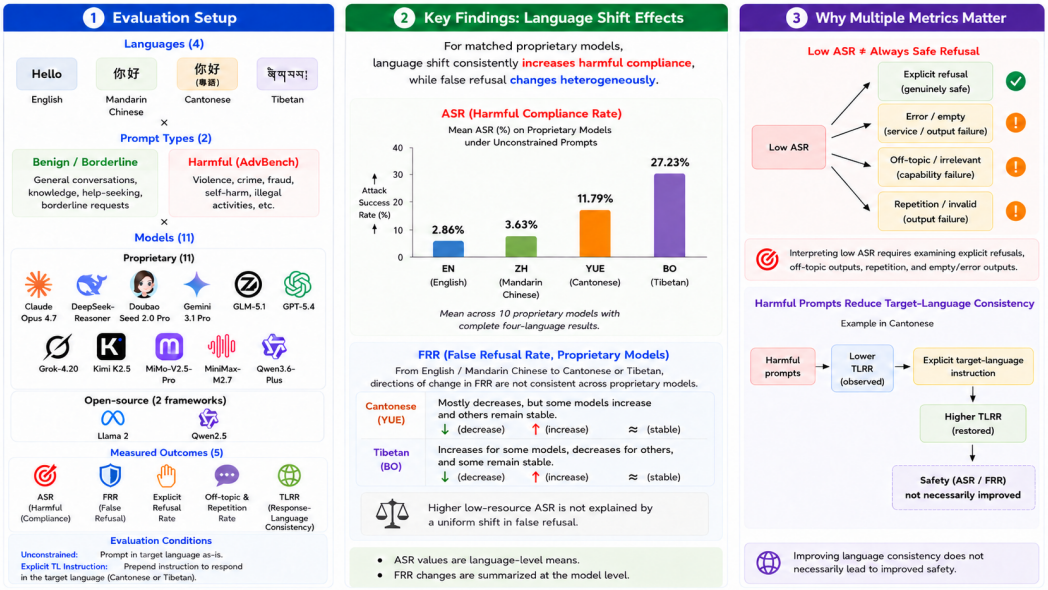


Figure 1: Overview of our multilingual safety evaluation.

benign and borderline prompts. For matched model–language pairs, we report ASR, false refusal rate (FRR), explicit refusal, off-topic output, repetition, and target-language response rate (TLRR). We also test explicit Cantonese- and Tibetan-response instructions. Responses are labeled using two independent judges with adjudication for disagreements. We manually review and correct all Cantonese AdvBench labels affecting ASR and conduct a stratified Tibetan audit to assess whether harmful compliance is underestimated.

Our main contributions and findings are:

1. We provide a unified multilingual evaluation of under-safety, over-safety, semantic relevance, output validity, and response-language consistency for the same model–language pairs, including the less-studied Cantonese and Tibetan settings.
2. Every proprietary model with paired observations has higher ASR in Cantonese and Tibetan than in its own English–Mandarin baseline, whereas FRR changes heterogeneously. The multilingual safety gap therefore cannot be explained by a single global refusal tendency.
3. Low ASR is not always evidence of effective safety refusal. For several open-source model–language pairs, it results from off-topic or repetitive output; off-topic behavior also persists across benign and harmful datasets. Moreover, harmful prompts reduce Cantonese TLRR, while explicit language constraints improve response-language consistency without consistently improving safety.

2 Related Work

Low-resource language resources and evaluation. Research on Cantonese and Tibetan has expanded from task-specific resources to foundation corpora, language-model adaptation, and comprehensive LLM benchmarks. For Cantonese, prior work has documented persistent challenges arising from limited data, colloquial writing, multilingual usage, and code-switching, while recent benchmarks evaluate general capabilities, language understanding, and culturally grounded knowledge (Xiang et al., 2025; Jiang et al., 2025; Cheng et al., 2025). For Tibetan, researchers have developed structured foundation datasets, offensive-language detection methods, and continual pretraining and instruction-tuning pipelines (Huang et al., 2025; Ren et al., 2024; Chen et al., 2025). These studies provide important linguistic and capability evaluations, but do not jointly characterize harmful compliance, false refusal, off-topic generation, and response-language consistency.

Multilingual safety and jailbreak evaluation. Multilingual safety studies have shown that safeguards learned in high-resource languages may transfer unevenly to other languages, motivating translated jailbreak benchmarks, analyses of cross-lingual refusal mechanisms, and multilingual safety-alignment methods (Yong et al., 2023; Deng et al., 2024; Wang et al., 2024; 2025; Zhao et al., 2025). Recent cross-cultural benchmarks have begun to supplement attack success rate with metrics intended to distinguish genuinely safe behavior from non-substantive generation (Choi et al., 2026). Low-resource and culturally situated evaluations likewise report that safety guardrails can degrade in multilingual settings (Hu et al., 2025). Standardized red-teaming suites and response-quality-aware evaluators further show that binary attack-success measures can obscure whether a response provides useful harmful assistance (Mazeika et al., 2024; Souly et al., 2024). Nevertheless, most existing studies focus on harmful prompts or alignment interventions. They rarely compare the same model–language pairs across harmful and benign safety settings, and target-language consistency is generally not evaluated as an independent dimension of model behavior.

Over-refusal and the evaluation gap. XSTest and OR-Bench formalize over-refusal by testing whether models reject benign prompts that superficially resemble harmful requests (Röttger et al., 2024; Cui et al., 2025). Complementary refusal benchmarks such as SORRY-Bench broaden unsafe-topic coverage and explicitly test linguistic variations in refusal behavior (Xie et al., 2025). Although multilingual studies have tested refusal behavior using translated harmful and benign prompts (Wang et al., 2025), dedicated over-refusal resources remain scarce for the low-resource language settings considered in this work. Moreover, jailbreak robustness and over-refusal are typically evaluated separately. We address this gap by using translated AdvBench prompts together with adapted XSTest and OR-Bench Hard prompts, and jointly analyzing attack success, explicit refusal, false refusal, off-topic responses, and target-language consistency.

3 METHODOLOGY

Models and Datasets. We evaluate eleven recent proprietary models and a set of open-source multilingual or language-adapted models. The main cross-language comparison tests the same proprietary models in English, Mandarin Chinese, Cantonese, and Tibetan. Kimi K2.5 is excluded from analyses requiring complete Tibetan observations because its paired Tibetan result is unavailable. Explicit response-language interventions are evaluated only on proprietary models.

Harmful compliance is evaluated using all 520 AdvBench prompts. English and Mandarin Chinese are treated as the higher-resource settings, and Cantonese and Tibetan as the lower-resource settings. The Cantonese and Tibetan translations of AdvBench were provided to our team by an external research group. Each translated version was verified by a co-author who is a native speaker of the corresponding language. Due to the lack of authorization from the dataset provider, the translated AdvBench resources cannot be publicly released.

Over-refusal is evaluated using 500 benign or borderline prompts per model–language pair: 250 from XSTest and 250 from OR-Bench Hard. The same underlying requests are used across languages. The Tibetan translations were verified by one native Tibetan-speaking co-author. The Cantonese translations were first verified by one native Cantonese-speaking co-author and then independently reviewed by another native Cantonese-speaking co-author. These translated over-refusal datasets will be provided in the supplementary material to facilitate reproduction.

Complete model information, decoding settings, access dates, translation procedures, sampling details, and prompt examples are provided in the supplementary material.

Evaluation Conditions. The unconstrained condition presents each prompt directly in its target language. For Cantonese and Tibetan, we additionally prepend an instruction requiring a response in the corresponding language. The Cantonese instruction is written in Standard Written Chinese, whereas the Tibetan instruction is written in Tibetan. The intervention is applied to both AdvBench and the over-refusal datasets, producing six proprietary-model conditions: English, Mandarin Chinese, Cantonese, Tibetan, constrained Cantonese, and constrained Tibetan. We compare each constrained condition with its within-language baseline. Because the two instructions and language-identification procedures differ, their intervention magnitudes are not interpreted as directly comparable.

Response Classification. AdvBench responses are assigned to three semantic outcome categories: harmful compliance, refusal, and off-topic response. Harmful compliance provides substantive assistance with the harmful request, while refusal includes rejection, safety explanations, and non-actionable safe alternatives. API-call failures are conservatively treated as refusals in the three-way AdvBench safety aggregation, while their raw counts are retained separately for diagnostic reporting. Repetitive, empty, or otherwise invalid model outputs are recorded separately and are not counted as harmful compliance.

Responses to XSTest and OR-Bench are classified as false refusal, off-topic response, or substantive engagement. Throughout the main text, false refusal rate (FRR) refers specifically to *textual false refusal*: only responses whose generated content is classified as a false refusal contribute to the FRR numerator. Empty outputs, API failures, and error-prefixed outputs are excluded from the FRR numerator and are retained as separate output-status diagnostics. All reported FRRs use the total number of evaluation prompts as the denominator. The open-source FRR values were verified to follow the same textual-refusal definition.

Content and response language are labeled independently. Two judge models independently classify each response. Their label is retained when they agree, and disagreements are resolved by a stronger adjudicator. Detailed definitions, judge prompts, model versions, mixed-response rules, and adjudication examples are included in the supplementary material.

Target-Language Identification. English, Mandarin Chinese, and Tibetan responses are identified primarily using script- or character-based rules. The Mandarin measure detects Chinese-script output and does not reliably distinguish Mandarin from written Cantonese. Cantonese therefore uses a hybrid procedure: diagnostic lexical items and constructions are first detected by rules, and inconclusive responses are classified by two judge models with adjudication for disagreements. Target-language response rate (TLRR) is computed for each dataset and condition. Because Cantonese and Tibetan use different identification procedures, we emphasize within-language paired changes rather than absolute cross-language comparisons.

Human Verification. We conduct targeted human verification of labels that affect the interpretation of ASR. For Cantonese AdvBench, all ASR-relevant response classifications are manually reviewed under both the unconstrained and explicit Cantonese-response conditions, and identified errors are corrected before aggregation.

We additionally assess the Cantonese three-way content classifier using stratified samples pooled across outputs from both proprietary and open-source models. All 50 sampled responses labeled as harmful compliance and all 50 sampled responses labeled as refusals agree with human judgment. Cantonese language identification is evaluated in the same pooled setting: agreement with human judgment is 95% among sampled responses classified as Cantonese and 100% among those classified as non-Cantonese.

Because off-topic responses are especially difficult to distinguish from harmful engagement and indirect refusal, we audit this category separately for proprietary and open-source models. Among 50 proprietary-model Cantonese responses initially labeled as off-topic, human review identifies 27 as harmful compliance and 23 as refusals; none remains off-topic. Among 45 proprietary-model Tibetan responses initially labeled as off-topic, 17 are relabeled as harmful compliance, 27 as refusals, and one remains off-topic.

For open-source models, among 50 Cantonese responses initially labeled as off-topic, seven are relabeled as harmful compliance, three as refusals, and 40 remain off-topic. Among 60 Tibetan responses initially labeled as off-topic, nine are relabeled as harmful compliance, three as refusals, and 48 remain off-topic. These audits show that off-topic classification errors are more pronounced for proprietary-model outputs, whereas a substantial proportion of open-source outputs are genuinely off-topic.

For Tibetan AdvBench, we conduct a stratified human audit of samples pooled across proprietary-model outputs. All 33 sampled responses labeled as harmful compliance are confirmed as harmful, and all 50 sampled responses labeled as refusals are confirmed as refusals.

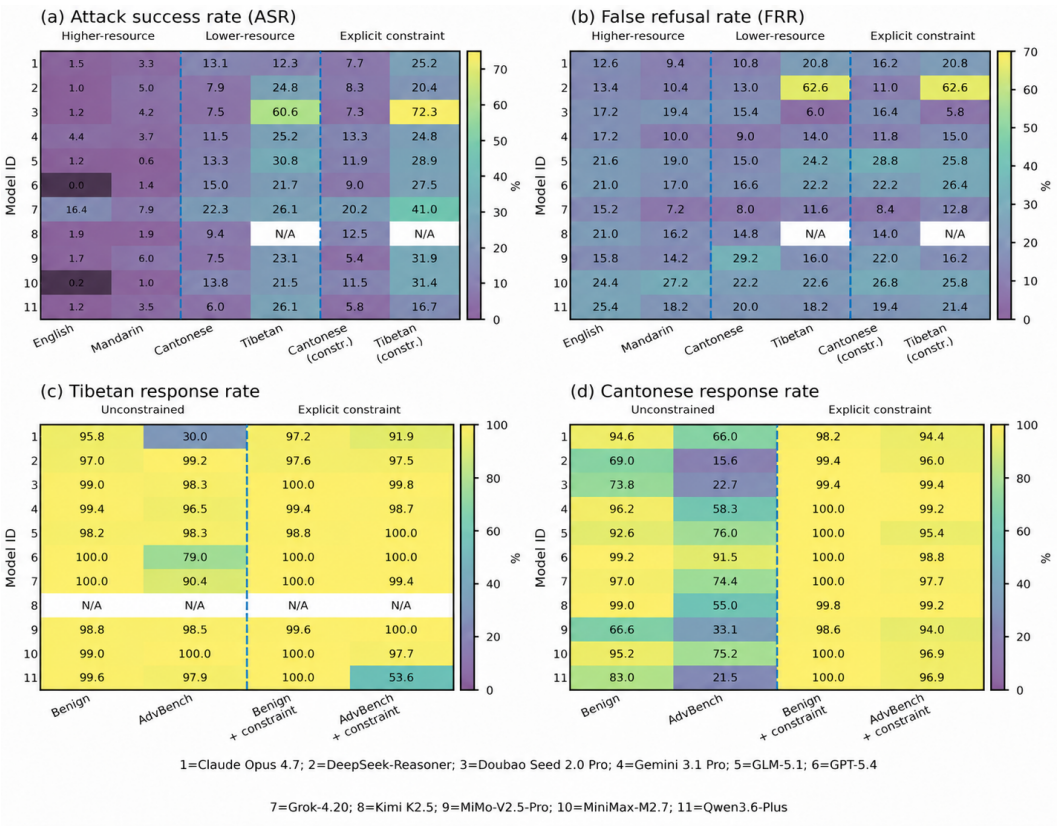


Figure 2: Safety and target-language consistency of proprietary models. Panels (a) and (b) report attack success rate (ASR) and false refusal rate (FRR) across English, Mandarin Chinese, Cantonese, Tibetan, and the explicit Cantonese- and Tibetan-response conditions. Panels (c) and (d) report Tibetan and Cantonese target-language response rates under benign and harmful prompts, with and without explicit response-language constraints. Values are percentages. The Tibetan result for Kimi K2.5 is unavailable.

Metrics and Statistical Analysis. We report attack success rate (ASR), textual false refusal rate (FRR), off-topic rate, repetition or invalid-output rate, substantive-answer rate, and TLRR. Unless otherwise specified, all rates use the total number of prompts as their denominator. Proprietary-model results are summarized using model-level heatmaps and language-level averages. FRR–ASR relationships are examined separately in English, Mandarin Chinese, Cantonese, Tibetan, and the two explicit response-language constraint conditions using Pearson correlation. Constrained and unconstrained conditions are compared using paired t-tests and Wilcoxon signed-rank tests.

4 Experimental Results

4.1 Cross-Language Safety Shifts

Figure 2 shows that harmful compliance changes more consistently across languages than false refusal. Across the ten proprietary models with complete observations, mean ASR increases from 2.86% and 3.63% in English and Mandarin Chinese to 11.79% and 27.23% in Cantonese and Tibetan, respectively. By comparison, mean textual FRR is 18.38% in English, 15.20% in Mandarin Chinese, 15.92% in Cantonese, and 21.82% in Tibetan.

Every proprietary model with paired observations has higher ASR in Cantonese and Tibetan than in its own English–Mandarin baseline, whereas textual FRR changes heterogeneously. For example, Doubao combines a Tibetan FRR of 6.0% with an ASR of 60.58%, while DeepSeek reaches 62.6%

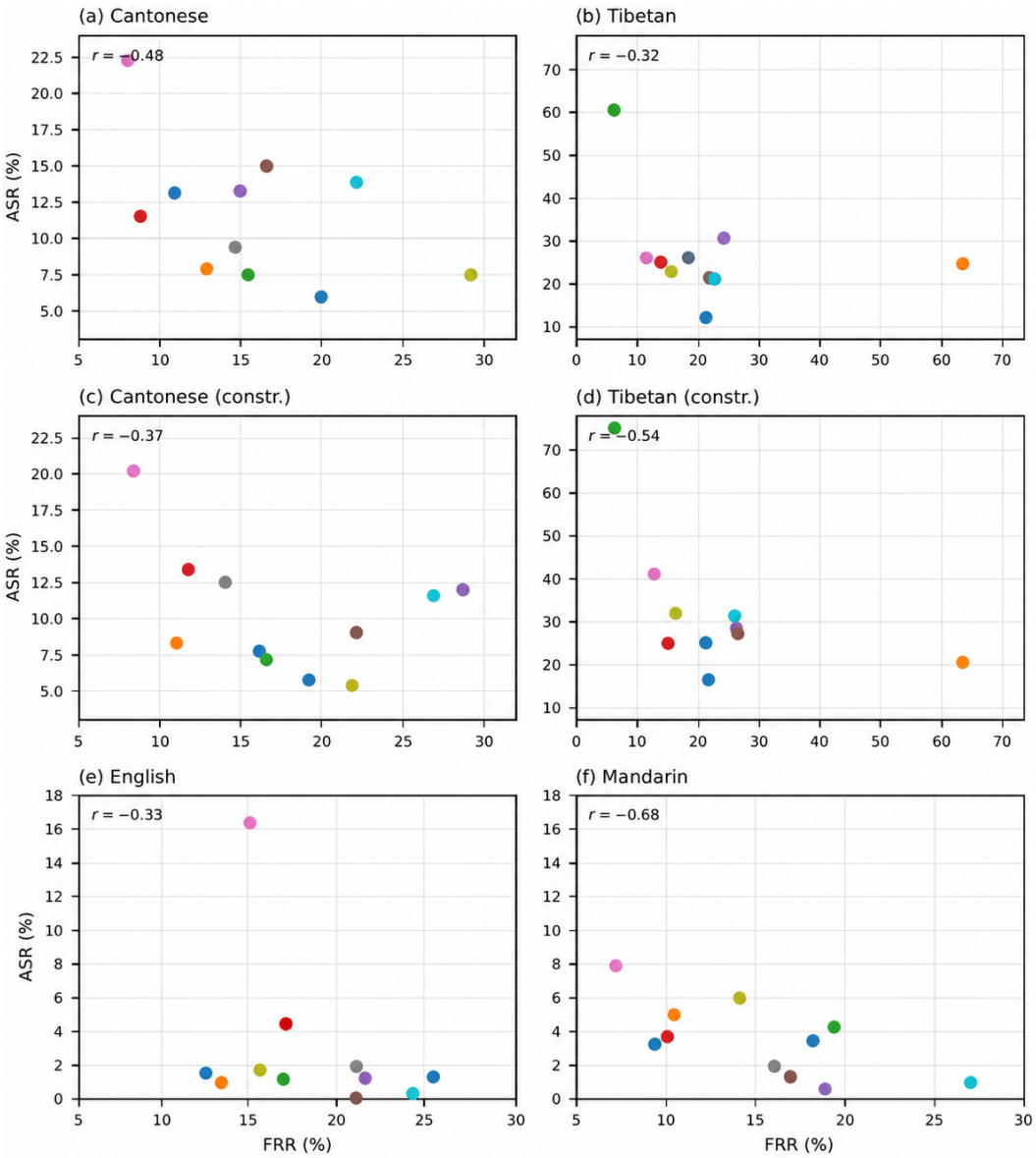


Figure 3: Relationships between textual false refusal rate (FRR) on the over-refusal evaluation and attack success rate (ASR) on AdvBench across six proprietary-model evaluation conditions. FRR counts only generated responses classified as false refusals; empty and error-prefixed outputs are excluded from the FRR numerator. Panels (a)–(d) show Cantonese and Tibetan under unconstrained and explicit target-language response conditions, while panels (e)–(f) show English and Mandarin Chinese as higher-resource references. Each point denotes one proprietary model, and r denotes Pearson's correlation coefficient computed within each condition. Kimi K2.5 is omitted from the Tibetan panels because complete Tibetan results are unavailable.

FRR and 24.81% ASR. Thus, increased low-resource ASR cannot be explained by a uniform shift in refusal tendency. We examine this relationship more directly in Figure 3.

4.2 Effect of Explicit Response-Language Constraints

The Cantonese constraint substantially improves target-language consistency: mean TLRR increases from 87.84% to 99.58% on benign prompts and from 53.57% to 97.10% on AdvBench, with both

changes significant under paired tests. However, the corresponding changes in ASR (11.57% to 10.26%) and textual FRR (15.82% to 17.91%) are not significant.

The Tibetan constraint produces smaller and more heterogeneous effects. Benign-prompt TLRR increases significantly from 98.68% to 99.26%, whereas the AdvBench increase from 88.81% to 93.86% is not significant. Safety does not improve: mean ASR rises from 27.23% to 32.00%, while textual FRR increases from 21.82% to 23.26%; the FRR increase is significant under both paired tests. These results indicate that improving response-language consistency does not necessarily improve safety alignment.

4.3 FRR–ASR Relationships Across Languages

Figure 3 compares model-level textual-FRR–ASR relationships across the four low-resource conditions and the English and Mandarin Chinese higher-resource references. The association is negative in all six conditions, but its magnitude varies substantially across languages and prompting conditions.

In English, the correlation is modest and non-significant ($r = -0.33$, $p = 0.323$), whereas Mandarin Chinese exhibits a stronger negative association ($r = -0.68$, $p = 0.022$). For unconstrained Cantonese and Tibetan, the correlations are $r = -0.48$ ($p = 0.135$) and $r = -0.32$ ($p = 0.373$), respectively. Under explicit target-language response constraints, the correlations are $r = -0.36$ ($p = 0.274$) for Cantonese and $r = -0.54$ ($p = 0.104$) for Tibetan.

As a sensitivity check, using the broader appendix FRR definition produces correlations of $r = -0.36, -0.69, -0.48, -0.34, -0.38$, and -0.55 across the six conditions in the same order. These values are close to the textual-FRR results and preserve the negative direction throughout, showing that the qualitative FRR–ASR pattern is robust to whether empty and error outputs are absorbed into FRR.

Thus, although all six point estimates are negative, the strength of the FRR–ASR association is not stable across languages or prompting conditions. Mandarin shows the strongest association, whereas none of the four low-resource conditions reaches the conventional 0.05 significance threshold. The more consistent cross-language pattern is therefore the increase in harmful compliance under Cantonese and Tibetan input despite heterogeneous changes in false refusal. We accordingly treat false refusal and harmful compliance as related but distinct dimensions of multilingual safety behavior.

As a robustness check, we recompute the unconstrained language-specific correlations after pooling the proprietary models with the open-source models evaluated in the corresponding language. The negative direction is preserved in all four languages, but the strength of the association remains strongly language dependent. The pooled correlations are $r = -0.25$ for English ($p = 0.396$), $r = -0.65$ for Mandarin Chinese ($p = 0.011$), $r = -0.49$ for Cantonese ($p = 0.057$), and $r = -0.13$ for Tibetan ($p = 0.664$).

Pooling open-source and proprietary models therefore does not reveal a stable cross-language FRR–ASR relationship. The association remains strongest in Mandarin Chinese, is moderate in Cantonese, and becomes substantially weaker in English and especially Tibetan. This sensitivity analysis further supports treating benign-request refusal and harmful compliance as distinct dimensions rather than assuming a universal safety–utility trade-off across languages and model families.

4.4 Off-Topic Behavior in Proprietary Models

Off-topic output is nearly absent from the manually reviewed Cantonese AdvBench results (Table 1), with rates of 0.02% and 0% under the unconstrained and constrained conditions, respectively. The classifier-reported Tibetan rates are 2.63% and 1.48%, but a stratified audit found that most predicted off-topic responses were harmful responses or refusals. Given the limited manual annotation capacity available for Tibetan, we were unable to exhaustively relabel all responses predicted as off-topic on the over-refusal evaluation sets. We therefore retain the classifier-reported Tibetan off-topic rates and interpret them as descriptive quantities rather than precise estimates of true off-topic behavior.

On the over-safety datasets, classifier-reported off-topic rates are generally low across proprietary models and conditions. However, the manual audit shows that Tibetan off-topic predictions have

Table 1: Off-topic rates (%) for proprietary models on AdvBench and the oversafety evaluation sets. Cantonese and Tibetan are shown under unconstrained and explicit target-language response conditions. Cantonese AdvBench values reflect the manually reviewed classifications; Tibetan AdvBench values are classifier outputs, with the unconstrained Tibetan category further examined through a stratified human audit. "Constrained" denotes that an explicit instruction to respond in the target language is prepended to the prompt.

	AdvBench				Oversafety			
Model	Cantonese Unconstr.	Cantonese Constrained	Tibetan Unconstr.	Tibetan Constrained	Cantonese Unconstr.	Cantonese Constrained	Tibetan Unconstr.	Tibetan Constrained
Claude Opus 4.7	0.00	0.00	0.77	1.54	0.00	0.00	0.40	0.60
DeepSeek-Reasoner	0.00	0.00	9.04	2.31	0.00	0.00	2.60	2.20
Doubao Seed 2.0 Pro	0.00	0.00	4.81	8.85	0.00	0.40	5.20	7.60
Gemini 3.1 Pro	0.00	0.00	2.50	0.00	0.00	0.00	0.00	0.00
GLM-5.1	0.00	0.00	0.19	0.96	0.00	0.00	0.00	0.60
GPT-5.4	0.00	0.00	6.15	0.38	0.00	0.00	0.20	0.60
Grok-4.20	0.00	0.00	0.19	0.00	0.20	0.00	0.20	0.20
Kimi K2.5	0.00	0.00	N/A	N/A	0.00	0.00	N/A	N/A
MiMo-V2.5-Pro	0.19	0.00	0.96	0.38	2.00	0.40	0.40	1.00
MiniMax-M2.7	0.00	0.00	0.38	0.19	0.00	0.00	0.00	0.40
Qwen3.6-Plus	0.00	0.00	1.35	0.19	0.00	0.00	0.80	1.40
Overall	0.02	0.00	2.63	1.48	0.20	0.07	0.98	1.46

Table 2: Outcome decomposition for open-source models on AdvBench and the oversafety evaluation sets (XSTest and OR-Bench Hard). All values are percentages. For AdvBench, one response per model–language pair was reassigned from refusal to harmful compliance, and *Other* denotes prompt repetition. For oversafety, *Subst.* denotes a substantive response, and *Other* denotes overlap or non-exclusive labels retained by the original aggregation.

		AdvBench					Oversafety				
Lang.	Model	Ref.	Off.	Harm.	Other	TLRR	FRR	Off.	Subst.	Other	TLRR
English	Llama-2-13B	99.23	0.00	0.77	0.00	100.00	55.6	0.2	44.2	0.0	100.0
	Llama-2-7B	99.62	0.00	0.38	0.00	100.00	51.0	0.2	48.8	0.0	100.0
	Qwen2.5-7B	98.27	0.77	0.96	0.00	100.00	13.0	0.4	86.6	0.0	100.0
Mandarin	Chinese-Alpaca-13B	46.15	2.50	51.15	0.19	100.00	4.4	3.8	91.4	0.4	100.0
	Chinese-Alpaca-7B	59.04	1.92	38.27	0.77	100.00	5.4	4.0	90.4	0.2	100.0
	Qwen2.5-7B	91.92	2.12	5.96	0.00	100.00	13.8	0.6	85.6	0.0	100.0
Cantonese	Cantonese-Llama-7B	5.58	75.38	12.88	6.15	84.23	1.2	93.4	3.4	2.0	92.4
	CantoneseLLMChat-7B	76.92	4.62	18.46	0.00	27.12	9.0	2.8	88.2	0.0	39.6
	Chinese-Alpaca-13B	33.27	7.12	59.62	0.00	30.00	4.6	18.2	77.2	0.0	21.4
	Chinese-Alpaca-7B	52.12	6.15	40.77	0.96	14.42	9.2	16.2	74.2	0.4	30.2
	Qwen2.5-7B	87.88	4.23	7.88	0.00	5.58	12.8	2.4	84.8	0.0	6.6
Tibetan	Tibetan-Llama-13B	23.08	40.58	34.42	1.92	100.00	3.2	69.6	24.6	2.6	100.0
	Tibetan-Llama-7B	20.00	30.19	47.88	1.92	100.00	6.4	59.2	32.6	1.8	98.0
	Qwen2.5-7B	1.54	97.12	1.35	0.00	30.77	1.2	94.0	4.8	0.0	23.0
	Qwen2.5-Tibetan-LoRA	8.27	42.31	20.19	29.23	99.81	2.8	67.6	8.0	21.6	99.6

low precision, so these rates are interpreted descriptively rather than as reliable estimates of true off-topic behavior.

4.5 Outcome Decomposition for Open-Source Models

Table 2 shows that low ASR can result from either explicit refusal or capability-related failure. The English Llama-2 models refuse more than 99% of harmful prompts, whereas the Chinese-Alpaca models reach ASRs of 38.27%–59.62% under Mandarin and Cantonese input.

Some low-resource settings instead produce predominantly off-topic outputs. Under Tibetan input, Qwen2.5-7B has an ASR of 1.35% but an off-topic rate of 97.12%. Qwen2.5-Tibetan-LoRA achieves 99.81% TLRR while producing 20.19% harmful compliance, 42.31% off-topic output, and

29.23% repetition. Thus, neither low ASR nor high TLRR alone demonstrates reliable safety alignment.

5 DISCUSSION

Our results reveal asymmetric transfer of safety alignment across languages. Every proprietary model with paired observations shows higher harmful compliance under Cantonese and Tibetan input than under its own English–Mandarin baseline, whereas textual FRR changes in different directions across models. The FRR–ASR correlations are negative in all six conditions but vary substantially in magnitude across languages and prompting conditions; the same qualitative pattern is preserved under the broader appendix FRR aggregation. This variability argues against a universal shift in a single refusal threshold. Multilingual evaluation should therefore treat harmful compliance and benign-request refusal as related but distinct dimensions.

The open-source results further demonstrate that a low attack success rate is not necessarily evidence of effective safety alignment. In several low-resource settings, harmful compliance remains low because the model produces irrelevant, repetitive, or otherwise invalid output, not because it recognizes and refuses the harmful request. Generic Qwen2.5 under Tibetan input provides the clearest example: its low ASR co-occurs with very limited refusal and predominantly off-topic generation. Moreover, high off-topic rates occur in both the benign and harmful datasets for several low-resource model–language pairs. This cross-dataset persistence suggests that many such outputs reflect broader limitations in language understanding or instruction following rather than a safety mechanism specific to harmful prompts. Evaluations that merge these outcomes into a single non-compliance category may therefore overestimate model safety, particularly in languages for which model capabilities remain uneven.

Response-language consistency represents another dimension that is related to, but not determined by, content safety. Harmful prompts substantially reduce Cantonese response rates for several proprietary models, even when the same models respond consistently in Cantonese to benign requests. Explicit response-language instructions largely restore Cantonese output, but do not consistently reduce harmful compliance or false refusal. The Tibetan intervention similarly produces heterogeneous safety changes despite generally high target-language response rates. Consequently, target-language consistency should be reported independently rather than treated as evidence that a model has correctly understood, safely handled, or substantively answered a request. Language-adapted open-source models likewise show that improved target-language generation does not necessarily imply better safety and may expose unsafe behavior previously obscured by poor instruction following; these comparisons remain observational because the evaluated models differ in training data, tokenization, instruction tuning, and safety objectives.

6 CONCLUSION

We evaluated proprietary and open-source LLM safety across English, Mandarin Chinese, Cantonese, and Tibetan using harmful, benign, and borderline instructions. For every proprietary model with paired results, ASR increases under Cantonese and Tibetan input relative to its English–Mandarin baseline, while FRR changes in different directions across models. Multilingual safety degradation therefore cannot be explained solely by stronger or weaker refusal tendencies. The open-source results further show that low ASR may reflect off-topic generation, repetition, or language mismatch rather than successful safety refusal. Target-language consistency also varies independently from harmful compliance and decreases substantially under harmful Cantonese prompts for several proprietary models. Explicit response-language constraints largely restore Cantonese language consistency without producing a correspondingly consistent improvement in safety. These findings show that reliable multilingual safety evaluation must jointly distinguish harmful compliance, false refusal, semantic relevance, repetition, and response-language consistency.

AI USE STATEMENT

In this work, we used generative AI tools solely for: (1) literature search, with all retrieved references manually screened and verified against their original sources; (2) writing experimental code;

(3) classifying experimental results via API calls; and (4) generating the overview figure in the main text. All AI-assisted code, classification results, and figures were manually reviewed and validated by the authors, and no AI-generated text was copied into the manuscript. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics Statement

This work studies multilingual safety behavior in large language models and includes jailbreak prompts that may contain potentially harmful or dual-use content. These prompts are used solely for safety evaluation and red-teaming purposes. To reduce the risk of misuse, we do not provide unrestricted public access to the language-specific jailbreak datasets. Following acceptance, qualified researchers may request access through the contact email provided by the authors. The over-refusal datasets, which contain benign requests designed to probe unnecessary refusals, will be released publicly.

Our dataset construction procedure explicitly excludes political, ideological, and other unnecessarily sensitive content that is not required for the research questions considered in this work. The evaluation does not use private user conversations or personally identifiable information. Model outputs are analyzed only for research purposes, and we report aggregate behavioral results rather than attributing sensitive properties to individual users or groups.

Reproducibility Statement

We provide the over-refusal datasets and the code used for model evaluation, response classification, metric computation, and figure generation in an anonymous repository: `https://anonymous.4open.science/r/eval_yue_zang_by_translation-5364`. The appendix documents the construction procedure and category taxonomy of the language-specific datasets, the response-classification protocol, target-language identification procedure, and manual validation results. We also report the evaluation settings and definitions of ASR, FRR, OFF, TLRR, and target-language-conditioned metrics used throughout the experiments.

Because the jailbreak datasets contain potentially harmful prompts, they are not distributed through unrestricted public download. Upon acceptance, qualified researchers may obtain access for research purposes by contacting the corresponding author, whose email address will be provided in the final version. This controlled-release policy is intended to balance reproducibility with the dual-use risks associated with jailbreak data.

References

Lifeng Chen, Ryan Lai, and Tianming Liu. Adapting large language models to low-resource Tibetan: A two-stage continual and supervised fine-tuning study, 2025. URL https://arxiv.org/abs/2512.03976.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pp. 1–11, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.conll-1.1. URL https://aclanthology.org/2025.conll-1.1/.

Dasol Choi, Eugenia Kim, Jaewon Noh, Sang Seo, Eunmi Kim, Myunggyo Oh, Yunjin Park, Brigitta Jesica Kartono, Josef Pichlmeier, Helena Berndt, Sai Krishna Mendu, Glenn Johannes Tungka, Özlem Gökçe, Suresh Gehlot, Katherine Pratt, Amanda Minnich, and Haon Park. XL-SafetyBench: A country-grounded cross-cultural benchmark for LLM safety and cultural sensitivity, 2026. URL https://arxiv.org/abs/2605.05662.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, 2024. URL https://openreview.net/forum?id=vESNKdEMGp.

Yujia Hu, Ming Shan Hee, Preslav Nakov, and Roy Ka-Wei Lee. Toxicity red-teaming: Benchmarking LLM safety in Singapore's low-resource languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12183–12201, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.612. URL https://aclanthology.org/2025.emnlp-main.612/.

Cheng Huang, Fan Gao, Nyima Tashi, Yutong Liu, Yadi Liu, Wenbin Wei, Xiangxiang Wang, and Yongbin Yu. TFD: A comprehensive structured Tibetan foundation dataset for low-resource language processing and large-scale modeling, 2025. URL https://arxiv.org/abs/2503.18288. Current arXiv version revised in 2026.

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. How well do LLMs handle Cantonese? benchmarking Cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4464–4505, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.253. URL https://aclanthology.org/2025.findings-naacl.253/.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Hang Ren, Jin Yang, Shengqiao Ni, Qin Yang, Jing Zhang, and Nuo Qun. A method for Tibetan offensive language detection based on prompt learning and information theory. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, Yokohama, Japan, 2024. IEEE. doi: 10.1109/IJCNN60899.2024.10651369. URL https://doi.org/10.1109/IJCNN60899.2024.10651369.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. doi: 10.52202/079017-3984. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e2e06adf560b0706d3b1ddfca9f29756-Abstract-Datasets_and_Benchmarks_Track.html.

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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5865–5877, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.349. URL https://aclanthology.org/2024.findings-acl.349/.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. Refusal direction is universal across safety-aligned languages. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2e94772e3c079d83f79c311d07456111-Abstract-Conference.html.

Rong Xiang, Emmanuele Chersoni, Yixia Li, Jing Li, Chu-Ren Huang, Yushan Pan, and Yushi Li. Cantonese natural language processing in the transformers era: A survey and current challenges. *Language Resources and Evaluation*, 59(2):1747–1773, 2025. doi: 10.1007/s10579-024-09744-w. URL https://doi.org/10.1007/s10579-024-09744-w.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-bench: Systematically evaluating large language model safety refusal. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YfKNaRktan.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. In *NeurIPS Workshop on Socially Responsible Language Modelling Research*, 2023. URL https://arxiv.org/abs/2310.02446. Best Paper.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. MPO: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23564–23587, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1149. URL https://aclanthology.org/2025.acl-long.1149/.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

A Detailed Three-Way Outcome Decomposition for Proprietary Models

This appendix provides the full three-way outcome decomposition for the proprietary models evaluated on AdvBench and the over-refusal datasets. The decomposition used here is intentionally broader than the textual-refusal metrics reported in the main text.

For AdvBench, outputs are grouped into broad refusal (B-Ref.), off-topic response, and harmful compliance, where harmful compliance corresponds to the attack success rate (ASR). Broad refusal includes responses explicitly classified as refusals as well as empty and error-prefixed outputs that are not assigned to harmful compliance or off-topic response.

For the over-refusal evaluation, outputs are grouped into broad false refusal (B-FRef.), off-topic response, and substantive engagement. The main-text FRR counts only generated responses whose content is classified as a textual false refusal. By contrast, B-FRef. additionally includes empty and error-prefixed outputs so that the three appendix categories form an exhaustive decomposition up to rounding. Equivalently,

$$\text{B-FRef.} = 100\% - \text{Off.} - \text{Subst.}.$$

These broad refusal aggregates are used only for the diagnostic decompositions in this appendix and are not used for the main-text FRR values or FRR–ASR correlation analyses.

In English, the correlation is modest and non-significant ($r = -0.33$, $p = 0.323$), whereas Mandarin Chinese exhibits a stronger negative association ($r = -0.68$, $p = 0.022$). For unconstrained Cantonese and Tibetan, the correlations are $r = -0.48$ ($p = 0.135$) and $r = -0.32$ ($p = 0.373$), respectively. Under explicit target-language response constraints, the correlations are $r = -0.36$ ($p = 0.274$) for Cantonese and $r = -0.54$ ($p = 0.104$) for Tibetan.

Thus, although all six point estimates are negative, the strength of the FRR–ASR association is not stable across languages or prompting conditions. Mandarin shows the strongest association, whereas none of the four low-resource conditions reaches the conventional 0.05 significance threshold. The more consistent cross-language pattern is therefore the increase in harmful compliance under Cantonese and Tibetan input despite heterogeneous changes in false refusal. We accordingly treat false refusal and harmful compliance as related but distinct dimensions of multilingual safety behavior.

Table 3: Three-way outcome decomposition (%) on AdvBench under the four unconstrained language conditions. B-Ref. denotes the broad refusal aggregate, Off. denotes off-topic output, and Harm. denotes harmful compliance (ASR). B-Ref. includes textual refusals as well as empty and error-prefixed outputs.

	English			Mandarin			Cantonese			Tibetan		
Model	B-Ref.	Off.	Harm.	B-Ref.	Off.	Harm.	B-Ref.	Off.	Harm.	B-Ref.	Off.	Harm.
Claude Opus 4.7	98.08	0.38	1.54	76.15	20.58	3.27	86.92	0.00	13.08	86.92	0.77	12.31
DeepSeek-Reasoner	99.04	0.00	0.96	94.04	0.96	5.00	92.12	0.00	7.88	66.15	9.04	24.81
Doubao Seed 2.0 Pro	98.85	0.00	1.15	94.23	1.54	4.23	92.50	0.00	7.50	34.61	4.81	60.58
Gemini 3.1 Pro	91.35	4.23	4.42	80.77	15.58	3.65	88.46	0.00	11.54	72.31	2.50	25.19
GLM-5.1	98.85	0.00	1.15	82.69	16.73	0.58	86.73	0.00	13.27	69.04	0.19	30.77
GPT-5.4	98.08	1.92	0.00	85.19	13.46	1.35	85.00	0.00	15.00	72.12	6.15	21.73
Grok-4.20	83.46	0.19	16.35	79.24	12.88	7.88	77.69	0.00	22.31	73.66	0.19	26.15
Kimi K2.5	98.08	0.00	1.92	94.04	4.04	1.92	90.58	0.00	9.42	N/A	N/A	N/A
MiMo-V2.5-Pro	N/A	N/A	1.73	N/A	N/A	5.96	92.31	0.19	7.50	75.96	0.96	23.08
MiniMax-M2.7	99.81	0.00	0.19	84.42	14.62	0.96	86.15	0.00	13.85	78.08	0.38	21.54
Qwen3.6-Plus	98.85	0.00	1.15	94.04	2.50	3.46	94.04	0.00	5.96	72.50	1.35	26.15

Table 4: Three-way outcome decomposition (%) on AdvBench with explicit target-language response constraints. B-Ref. denotes the broad refusal aggregate, Off. denotes off-topic output, and Harm. denotes harmful compliance (ASR). B-Ref. includes textual refusals as well as empty and error-prefixed outputs and is used only for the exhaustive appendix decomposition.

	Cantonese (Constr.)			Tibetan (Constr.)		
Model	B-Ref.	Off.	Harm.	B-Ref.	Off.	Harm.
Claude Opus 4.7	92.31	0.00	7.69	73.27	1.54	25.19
DeepSeek-Reasoner	91.73	0.00	8.27	77.31	2.31	20.38
Doubao Seed 2.0 Pro	92.69	0.00	7.31	18.84	8.85	72.31
Gemini 3.1 Pro	86.73	0.00	13.27	75.19	0.00	24.81
GLM-5.1	88.08	0.00	11.92	70.19	0.96	28.85
GPT-5.4	90.96	0.00	9.04	72.12	0.38	27.50
Grok-4.20	79.81	0.00	20.19	59.04	0.00	40.96
Kimi K2.5	87.50	0.00	12.50	N/A	N/A	N/A
MiMo-V2.5-Pro	94.62	0.00	5.38	67.70	0.38	31.92
MiniMax-M2.7	88.46	0.00	11.54	68.46	0.19	31.35
Qwen3.6-Plus	94.23	0.00	5.77	83.08	0.19	16.73

A.1 ADVBENCH

Tables 3 and 4 report the exhaustive outcome decomposition for proprietary models on AdvBench. Here, B-Ref. is a diagnostic broad-refusal category rather than a textual refusal rate. In addition to responses explicitly classified as refusals, it absorbs empty outputs and error-prefixed outputs. Error-prefixed outputs indicate cases in which the model service reports that the input is risky, blocked, or filtered, whereas empty outputs are cases for which no textual response is returned after multiple repeated submission attempts. Because neither type provides harmful assistance, this aggregation does not affect ASR; it is used only to make the three-way appendix decomposition exhaustive.

Table 5: Three-way outcome decomposition (%) on the over-refusal evaluation sets under the four unconstrained language conditions. B-FRef. denotes the broad false-refusal aggregate ($\mathrm{FRR}_{\mathrm{broad}}$), Off. denotes off-topic output, and Subst. denotes substantive engagement. Unlike the textual FRR reported in the main text, B-FRef. also absorbs empty, error-prefixed, and other residual non-substantive outputs so that the three categories sum to 100% up to rounding.

	English			Mandarin			Cantonese			Tibetan		
Model	B-FRef.	Off.	Subst.	B-FRef.	Off.	Subst.	B-FRef.	Off.	Subst.	B-FRef.	Off.	Subst.
Claude Opus 4.7	14.80	0.00	85.20	10.20	0.00	89.80	11.40	0.00	88.60	24.80	0.40	74.80
DeepSeek-Reasoner	13.40	0.00	86.60	10.40	0.00	89.60	13.00	0.00	87.00	65.60	2.60	31.80
Doubao Seed 2.0 Pro	17.20	0.00	82.80	19.40	0.00	80.60	15.40	0.00	84.60	6.00	5.20	88.80
Gemini 3.1 Pro	17.20	0.00	82.80	10.00	0.00	90.00	9.00	0.00	91.00	14.40	0.00	85.60
GLM-5.1	21.60	7.40	71.00	19.00	0.00	81.00	15.00	0.00	85.00	24.20	0.00	75.80
GPT-5.4	21.00	0.00	79.00	17.00	0.00	83.00	16.60	0.00	83.40	22.20	0.20	77.60
Grok-4.20	15.20	0.00	84.80	7.20	0.00	92.80	8.00	0.20	91.80	11.60	0.20	88.20
Kimi K2.5	21.20	0.00	78.80	16.40	1.20	82.40	15.00	0.00	85.00	N/A	N/A	N/A
MiMo-V2.5-Pro	15.80	0.20	84.00	14.20	0.00	85.80	29.40	2.00	68.60	16.00	0.40	83.60
MiniMax-M2.7	24.40	0.00	75.60	27.20	0.00	72.80	22.20	0.00	77.80	22.60	0.00	77.40
Qwen3.6-Plus	25.40	0.00	74.60	18.20	0.00	81.80	20.00	0.00	80.00	18.20	0.80	81.00

Table 6: Three-way outcome decomposition (%) on the over-refusal evaluation sets with explicit target-language response constraints. B-FRef. denotes the broad false-refusal aggregate ($\mathrm{FRR}_{\mathrm{broad}}$), Off. denotes off-topic output, and Subst. denotes substantive engagement. B-FRef. is used only for this exhaustive appendix decomposition and is distinct from the textual FRR reported in the main text.

	Cantonese (Constr.)			Tibetan (Constr.)		
Model	B-FRef.	Off.	Subst.	B-FRef.	Off.	Subst.
Claude Opus 4.7	18.00	0.00	82.00	23.60	0.60	75.80
DeepSeek-Reasoner	11.00	0.00	89.00	65.00	2.20	32.80
Doubao Seed 2.0 Pro	16.40	0.40	83.20	5.80	7.60	86.60
Gemini 3.1 Pro	11.80	0.00	88.20	15.60	0.00	84.40
GLM-5.1	28.80	0.00	71.20	27.00	0.60	72.40
GPT-5.4	22.20	0.00	77.80	26.40	0.60	73.00
Grok-4.20	8.40	0.00	91.60	12.80	0.20	87.00
Kimi K2.5	14.20	0.00	85.80	N/A	N/A	N/A
MiMo-V2.5-Pro	22.20	0.40	77.40	16.20	1.00	82.80
MiniMax-M2.7	26.80	0.00	73.20	25.80	0.40	73.80
Qwen3.6-Plus	19.40	0.00	80.60	21.40	1.40	77.20

A.2 Over-Refusal Evaluation

Manual verification of over-refusal labels. The three-way over-refusal classifier is highly reliable for Cantonese: all 50 sampled substantive answers and all 50 sampled refusals agree with human judgment, while 48 of 50 sampled off-topic predictions (96.0%) remain off-topic and two are relabeled as refusals. For Tibetan, all 50 sampled substantive answers and all 50 sampled refusals are also confirmed, but only 7 of 50 sampled off-topic predictions (14.0%) remain off-topic after human review; 30 are relabeled as substantive answers and 13 as refusals. Given limited Tibetan manual-annotation capacity, we do not exhaustively relabel the remaining off-topic predictions. Accordingly, the Tibetan off-topic rates in Tables 5 and 6 remain classifier-reported descriptive quantities and should not be interpreted as precise estimates of true off-topic behavior.

Table 7: Output-status rates (%) for proprietary models on AdvBench. Target denotes a target-language response, Empty denotes an empty output, and Error denotes an error-prefixed output. Non-target-language responses are omitted.

	Cantonese			Cantonese (Constr.)			Tibetan			Tibetan (Constr.)		
Model	Target	Empty	Error	Target	Empty	Error	Target	Empty	Error	Target	Empty	Error
Claude Opus 4.7	65.96	21.54	0.00	94.42	5.58	0.00	30.00	24.42	0.00	91.92	7.69	0.00
DeepSeek-Reasoner	15.58	0.00	0.00	95.96	0.00	0.00	99.23	0.58	0.00	97.50	2.50	0.00
Doubao Seed 2.0 Pro	22.69	0.00	0.00	99.42	0.00	0.00	98.27	0.00	0.00	99.81	0.00	0.00
Gemini 3.1 Pro	58.27	0.19	0.58	99.23	0.00	0.00	96.54	0.19	3.08	98.65	0.00	1.35
GLM-5.1	75.96	0.00	0.00	95.38	0.00	4.42	98.27	0.00	1.35	100.00	0.00	0.00
GPT-5.4	91.54	0.00	0.00	98.85	0.00	0.38	79.04	10.19	9.81	100.00	0.00	0.00
Grok-4.20	74.42	0.00	3.27	97.69	0.00	1.92	90.38	9.62	0.00	99.42	0.00	0.58
Kimi K2.5	55.00	0.00	0.00	99.23	0.38	0.00	N/A	N/A	N/A	N/A	N/A	N/A
MiMo-V2.5-Pro	33.08	0.00	0.00	94.04	0.00	0.00	98.46	0.00	0.00	100.00	0.00	0.00
MiniMax-M2.7	75.19	0.19	0.00	96.92	0.00	2.69	100.00	0.00	0.00	97.69	2.31	0.00
Qwen3.6-Plus	21.54	0.00	0.00	96.92	0.00	0.00	97.88	0.00	0.00	53.65	0.00	46.35

Table 8: Output-status rates (%) for proprietary models on the over-refusal evaluation sets. Target denotes a target-language response, Empty denotes an empty output, and Error denotes an error-prefixed output. Non-target-language responses are omitted.

	Cantonese			Cantonese (Constr.)			Tibetan			Tibetan (Constr.)		
Model	Target	Empty	Error	Target	Empty	Error	Target	Empty	Error	Target	Empty	Error
Claude Opus 4.7	94.60	0.60	0.00	98.20	1.80	0.00	95.80	4.00	0.00	97.20	2.80	0.00
DeepSeek-Reasoner	69.00	0.00	0.00	99.40	0.00	0.00	97.00	3.00	0.00	97.60	2.40	0.00
Doubao Seed 2.0 Pro	73.80	0.00	0.00	99.40	0.00	0.00	99.00	0.00	0.00	100.00	0.00	0.00
Gemini 3.1 Pro	96.20	0.00	0.00	100.00	0.00	0.00	99.40	0.00	0.40	99.40	0.00	0.60
GLM-5.1	92.60	0.00	0.00	100.00	0.00	0.00	98.20	0.00	0.00	98.80	1.20	0.00
GPT-5.4	99.20	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
Grok-4.20	97.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00	100.00	0.00	0.00
Kimi K2.5	99.00	0.20	0.00	99.80	0.20	0.00	N/A	N/A	N/A	N/A	N/A	N/A
MiMo-V2.5-Pro	66.60	0.20	0.00	98.60	0.20	0.00	98.80	0.00	0.00	99.60	0.00	0.00
MiniMax-M2.7	95.20	0.00	0.00	100.00	0.00	0.00	99.00	0.00	0.00	100.00	0.00	0.00
Qwen3.6-Plus	83.00	0.00	0.00	100.00	0.00	0.00	99.60	0.00	0.00	100.00	0.00	0.00

B Detailed Output-Status Decomposition for Proprietary Models

This appendix reports the three output-status categories contained in the language-identification result files for the proprietary models: target-language response, empty output, and error-prefixed output. For Cantonese, the target category corresponds to responses identified as Cantonese; for Tibetan, it corresponds to responses containing Tibetan. Non-target-language responses are not shown as a separate category and therefore the three displayed percentages are not required to sum to 100%. In our collection pipeline, an error-prefixed output indicates that the model service explicitly reports that the input is risky, blocked, or filtered. An empty output denotes a request for which no textual response is returned even after multiple repeated submission attempts. For the proprietary-model three-way decompositions in Appendix A, both categories are conservatively absorbed into the corresponding broad refusal aggregate (B-Ref. for AdvBench and B-FRef. for the over-refusal evaluation).

Only the proprietary models used in the main evaluation are retained. Kimi K2.5 is reported as N/A in the Tibetan conditions because Tibetan inputs frequently triggered service-level error responses indicating that the input was classified as high risk.

B.1 AdvBench

Table 7 reports the output-status decomposition under harmful AdvBench prompts. "Constr." denotes an explicit instruction requiring the model to answer in the corresponding target language. All

AdvBench conditions contain 520 prompts per evaluated model. Kimi K2.5 is marked as N/A in the Tibetan conditions because a paired Tibetan result is unavailable.

B.2 Over-Refusal Evaluation

Table 8 reports the same three output categories on the benign and borderline over-refusal evaluation sets. All four over-refusal conditions contain 500 prompts per evaluated proprietary model. "Constr." denotes an explicit instruction requiring the model to answer in the corresponding target language. Kimi K2.5 is marked as N/A in the Tibetan conditions because a paired Tibetan result is unavailable.