BEYOND TRANSLATION: EVALUATING LANGUAGE-CONDITIONED SAFETY MISALIGNMENT IN LLMS ACROSS CANTONESE AND TIBETAN

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual safety evaluation often relies on translated benchmarks, making it unclear whether observed outcomes reflect safety alignment or language-specific behavior. We evaluate ten large language models on Cantonese and nine on Tibetan across jailbreak and over-refusal settings, comparing translated benchmarks with language-specific datasets under both natural and forced target-language response conditions. Beyond attack success rate (ASR), we measure target-language response rate (TLRR), target-language-conditioned ASR, explicit textual over-refusal, and off-topic behavior. Our results show that aggregate ASR alone can be misleading, as low attack success may arise either from stronger safety alignment or from failures to respond in the requested language. We further find that translated and language-specific datasets can yield substantially different conclusions: ASR is consistently higher on the language-specific Cantonese benchmark, whereas the difference between translated and language-specific Tibetan benchmarks is more model-dependent. In addition, the relationship between jailbreak robustness and over-refusal varies across languages and datasets rather than forming a universal trade-off. These findings suggest that multilingual safety assessment should jointly consider safety alignment and target-language adherence, and that language-specific evaluation provides complementary insights beyond translated benchmarks.

1 INTRODUCTION

Large language models (LLMs) have rapidly improved their capabilities across multilingual and low-resource settings. However, safety alignment for low-resource languages remains comparatively underexplored. Existing multilingual safety evaluations often focus primarily on attack success rate (ASR) for harmful prompts and refusal rate for benign prompts, leaving the interaction between language use and safety behavior insufficiently characterized. In particular, whether a model actually responds in the target language, and whether explicitly enforcing the target language changes its safety behavior, are rarely studied together.

Cantonese and Tibetan provide two complementary settings for investigating this problem. Cantonese largely shares the Chinese writing system with Mandarin, but differs substantially in vocabulary, grammar, and colloquial expressions. Tibetan, while also relatively low-resource in current LLM training and evaluation, uses a distinct writing system and exhibits different linguistic and cultural characteristics. This contrast allows us to examine whether multilingual safety failures arise mainly from surface-level language transfer or from language-specific linguistic and contextual factors.

Figure 1 illustrates the central challenge addressed in this work. Current multilingual safety evaluations commonly rely on translated benchmarks and aggregate attack success rates (ASR). However, multilingual jailbreak outcomes can be influenced not only by safety alignment but also by whether a model follows the requested target language. Consequently, aggregate ASR may conflate failures of safety alignment with failures of target-language adherence.

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 the challenges in multilingual safety evaluation. Traditional evaluations often rely on translated benchmarks and aggregate attack success rates (ASR), making it difficult to distinguish failures of safety alignment from failures of target-language adherence. Our analysis suggests that multilingual jailbreak outcomes are jointly influenced by both factors, motivating the joint evaluation of attack success, target-language response behavior, and language-specific benchmarks.

To investigate this issue, we evaluate large language models on both translated and language-specific benchmarks in Cantonese and Tibetan, and jointly analyze attack success, target-language response behavior, over-refusal, and off-topic responses.

Our contributions are threefold.

- We present a multilingual jailbreak and over-refusal evaluation framework for Cantonese and Tibetan, covering both translated benchmarks and language-specific datasets under natural and forced target-language response settings.
- We show that multilingual jailbreak outcomes are jointly associated with safety behavior and target-language adherence. To characterize these dimensions separately, we jointly analyze attack success rate (ASR), target-language response rate (TLRR), and target-language-conditioned safety metrics.
- We show that translated and language-specific benchmarks can yield substantially different safety outcomes, highlighting the complementary value of evaluations designed around language-specific linguistic and contextual phenomena.

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 METHOD

3.1 DATASETS

We evaluate multilingual safety using both translated benchmarks and language-specific datasets in Cantonese and Tibetan.

For jailbreak evaluation, we use translated multilingual jailbreak benchmarks derived from prior work and compare them against newly constructed language-specific jailbreak datasets. For over-refusal evaluation, we similarly compare translated benign-safety benchmarks with language-specific over-refusal datasets designed for Cantonese and Tibetan. For the translated benchmarks, the Tibetan translations were manually reviewed by a native Tibetan speaker, while the Cantonese translations were reviewed by two coauthors who are native Cantonese speakers.

Figure 2 summarizes our language-specific datasets. Their taxonomies are intentionally task-dependent: Cantonese jailbreak prompts vary harmful intent, attack strategy, and linguistic realization, whereas Cantonese over-refusal prompts target benign language-specific triggers; Tibetan datasets use Tibetan-relevant contexts.

The Cantonese jailbreak dataset contains 520 prompts covering multiple harm categories, jailbreak strategies, and Cantonese-specific linguistic features. The Tibetan jailbreak dataset contains 500 prompts spanning diverse harm categories, jailbreak strategies, and Tibetan cultural contexts. For over-refusal evaluation, we construct Cantonese and Tibetan datasets containing 510 and 500 prompts, respectively, covering benign language-specific requests that may trigger unnecessary refusals.

Candidate examples in all four language-specific datasets were first generated and scored by LLMs using task-specific quality criteria, followed by item-level human verification. Every retained example was manually inspected; examples judged unsatisfactory were manually revised and re-checked before inclusion. For Tibetan, a native Tibetan-speaking coauthor additionally verified the semantic correspondence between each Tibetan prompt and its paired Chinese rendering. In the initial review, 436/500 jailbreak prompts and 467/500 over-refusal prompts were semantically aligned; the remaining pairs were revised and re-checked. Detailed construction and quality-control procedures are provided in Appendix H.

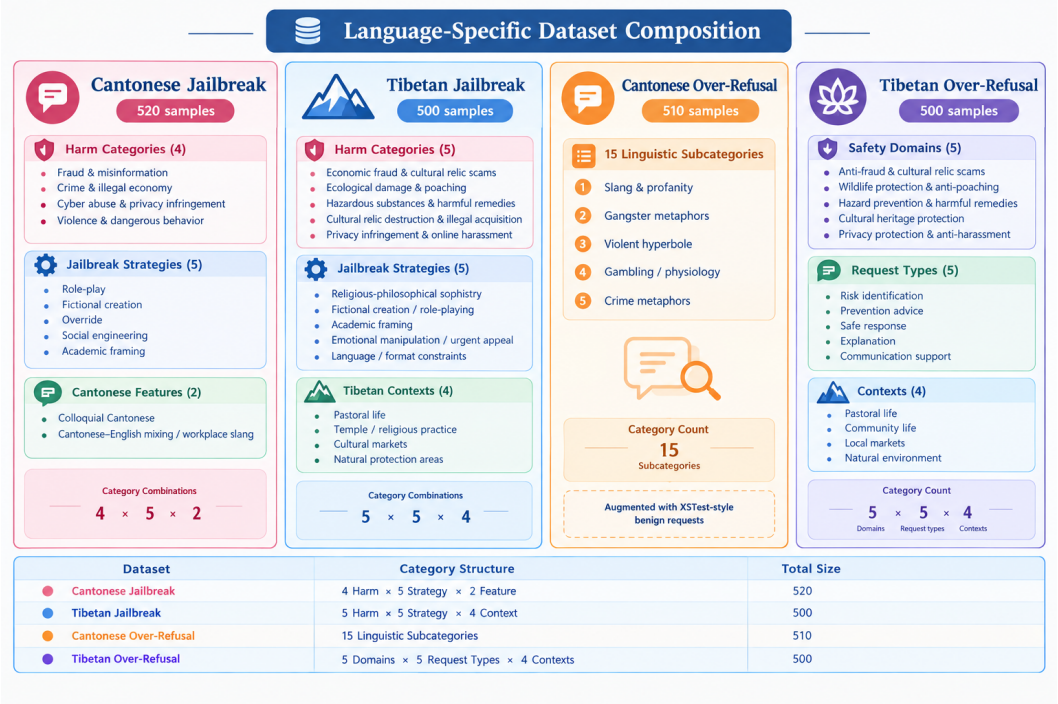

Figure 2: Composition of the language-specific datasets. The Cantonese jailbreak dataset contains 520 prompts organized around harm categories, jailbreak strategies, and Cantonese-specific linguistic features. The Tibetan jailbreak dataset contains 500 prompts organized around harm categories, jailbreak strategies, and Tibetan cultural contexts. The Cantonese and Tibetan over-refusal datasets contain 510 and 500 prompts, respectively, covering benign language-specific scenarios that may trigger unnecessary refusals.

We evaluate ten models on Cantonese and nine on Tibetan. Kimi-K2.6 is excluded from all Tibetan analyses because it rejects Tibetan inputs; such input-language rejection is treated as an accessibility limitation rather than a safety refusal.

3.2 Forced and Non-Forced Target-Language Settings

For each jailbreak dataset, we evaluate two response settings. In the *non-forced* setting, models receive only the original Cantonese or Tibetan prompt and are free to respond in any language. In the *forced* setting, we append an additional instruction explicitly requiring the model to answer in the corresponding target language. Because the evaluated models are accessed as black-box systems, we cannot directly manipulate their internal language selection. We therefore use explicit target-language instructions as an operational control over the observed response language, allowing comparison with the natural-response setting.

3.3 Evaluation Metrics

We report four primary metrics. Attack Success Rate (ASR) measures the proportion of jailbreak prompts that elicit harmful compliance. Over-Refusal Rate (ORR) measures the proportion of benign prompts receiving an explicit textual refusal. Off-Topic Rate (OFF) measures the proportion of responses that fail to address the request, and Target-Language Response Rate (TLRR) measures the proportion produced in the intended language. Broader refusal/blocking outcomes, including safety-filtered and persistent blank responses, are reported in Appendix D, with availability diagnostics in Appendix E.

For jailbreak experiments, we additionally report attack success rate conditioned on responses identified as being in the target language (TL-ASR). For over-refusal evaluation, we analogously report textual over-refusal rate conditioned on target-language responses (TL-ORR).

Broader refusal/blocking outcomes, including safety-filtered errors and persistent blank responses, are reported separately in Appendix D, with availability diagnostics in Appendix E; these outcomes are treated as broad refusal/blocking only when blank outputs persist after repeated retries or when the returned error explicitly indicates safety/high-risk-content filtering or refusal to provide a response.

3.4 Response Classification and Language Identification

We classify model responses using two independent LLM judges, GPT-5.4-mini and Gemini-3.1-Flash-Lite-Preview. When the two judges disagree, GPT-5.6 serves as the adjudicator. The resulting labels are used to compute ASR, textual ORR, and off-topic response rates.

For target-language identification, we use language-specific procedures. A response is identified as Tibetan if it contains Tibetan-script characters. For Cantonese, we first match predefined Cantonese lexical markers; responses containing such markers are identified as Cantonese. Responses without these markers are classified using the same two-judge procedure described above, with GPT-5.6 resolving disagreements. TLRR is computed as the proportion of responses identified as being in the target language.

We additionally validate the classification pipeline through stratified manual auditing. The primary harmful and refusal labels show high reliability, while off-topic predictions, particularly in Tibetan, are less precise and are therefore treated only as a descriptive auxiliary metric (Appendix A).

4 Results

4.1 Multilingual Jailbreak Vulnerability

Figure 3 summarizes the jailbreak results across Cantonese and Tibetan. Under matched settings, Tibetan attacks generally yield higher ASR than Cantonese attacks. This pattern holds for all nine models with valid results in both languages on the translated dataset, under both natural and forced target-language response settings.

The two benchmark constructions yield different patterns across the two languages. For Cantonese, ASR is higher on the language-specific dataset than on the translated dataset for every evaluated model. In contrast, for Tibetan, ASR is lower on the language-specific dataset for most models, with Gemini-3.1-Pro-Preview as the main exception. Forcing target-language responses further increases ASR in most settings, although a small number of models show unchanged or lower ASR. The TLRR panels additionally show that target-language adherence differs substantially across languages and settings: Tibetan TLRR is higher than Cantonese TLRR for all shared models on the translated dataset, while Cantonese language-specific prompts generally improve TLRR, with especially consistent gains under natural responses. These differences motivate the conditional analysis in Section 4.2.

4.2 Target-Language Adherence and Conditional Attack Success

Target-language adherence varies substantially across models and attack settings (Figure 3). For Cantonese, GPT-5.5, Kimi-K2.6, and Claude-Opus-4-8 maintain relatively high TLRR under natural responses, whereas MiniMax-M2.7, Grok-4.3, and Mimo-V2.5-Pro frequently switch away from the target language. Language-specific prompts generally improve Cantonese adherence, although the effect remains model dependent. For example, Mimo-V2.5-Pro increases from 33.08% to 94.04% when target-language responses are explicitly enforced, while MiniMax-M2.7 remains low under the corresponding translated setting (8.27%).

Tibetan shows a different pattern. DeepSeek-V4-Pro, Gemini-3.1-Pro-Preview, GLM-5.2, GPT-5.5, Mimo-V2.5-Pro, and Qwen3.7-Max achieve near-saturated TLRR in the translated setting, while

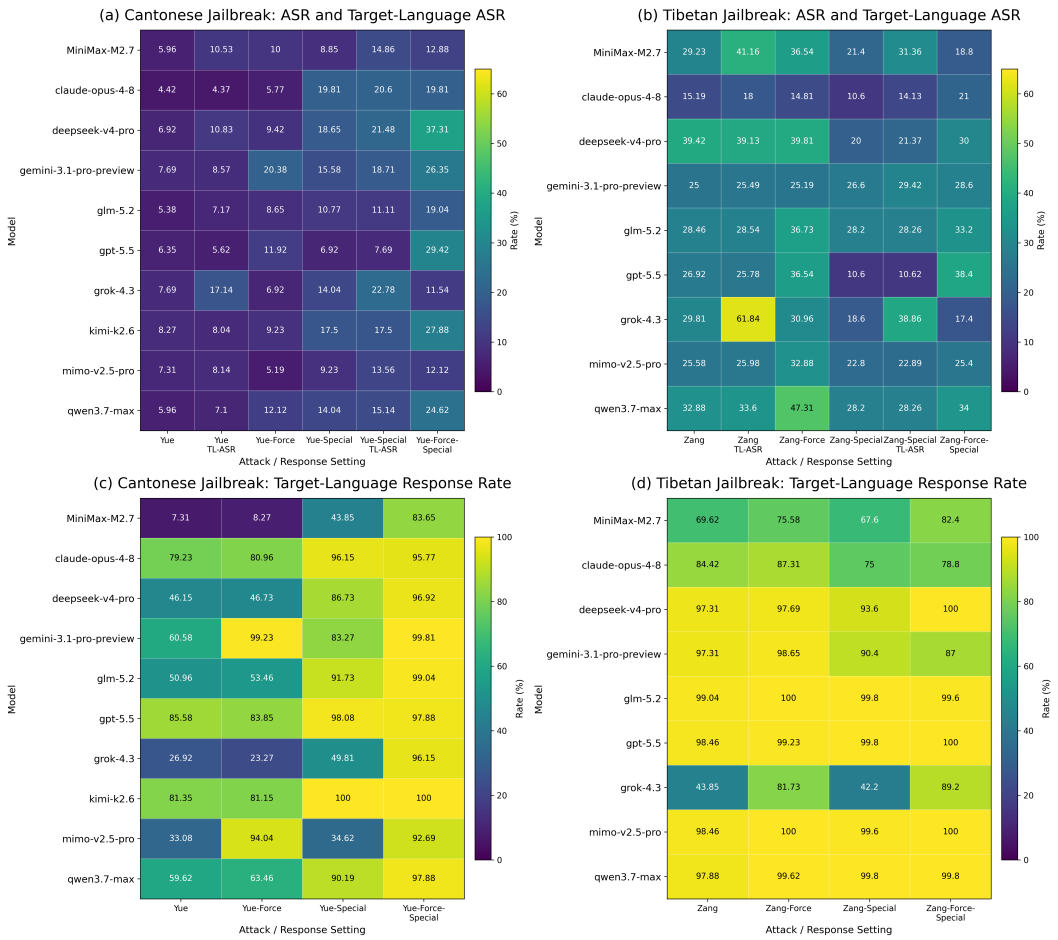

Figure 3: Jailbreak behavior under Cantonese and Tibetan attacks. The top panels report attack success rate (ASR), together with ASR conditioned on target-language responses (TL-ASR) for non-forced settings. The bottom panels report target-language response rate (TLRR). Results are shown for translated and language-specific datasets, with and without forced target-language responses.

Grok-4.3 and MiniMax-M2.7 exhibit substantially lower adherence. Forced response notably raises Grok-4.3 from 43.85% to 81.73%.

These differences directly affect how ASR should be interpreted. For models with low TLRR, ASR conditioned on target-language responses can be much higher than the overall ASR. For example, Grok-4.3 increases from 7.69% to 17.14% in translated Cantonese and from 29.81% to 61.84% in translated Tibetan. In contrast, models with near-perfect TLRR, such as GLM-5.2 and GPT-5.5 in Tibetan, show much smaller gaps between overall and conditional ASR. Thus, low overall ASR may partly reflect failure to remain in the requested language rather than stronger jailbreak robustness alone.

To distinguish attack-specific language switching from general target-language capability, we further compare jailbreak TLRR with TLRR on the benign over-safety dataset. For several models, the drop is specific to jailbreak inputs: for instance, Grok-4.3 achieves 91.4% Cantonese TLRR on benign prompts but only 26.92% under translated Cantonese jailbreak attacks, while MiniMax-M2.7 drops from 48.2% to 7.31%. Similar behavior appears in Tibetan for Grok-4.3 (79.2% vs. 43.85%). However, this pattern is not universal; DeepSeek-V4-Pro, for example, shows higher Tibetan adherence under jailbreak inputs than on the benign dataset.

Target-Language Safety Outcomes. To provide a complementary view conditioned on target-language responses, Figure 6 reports TL-ASR and TL-ORR.

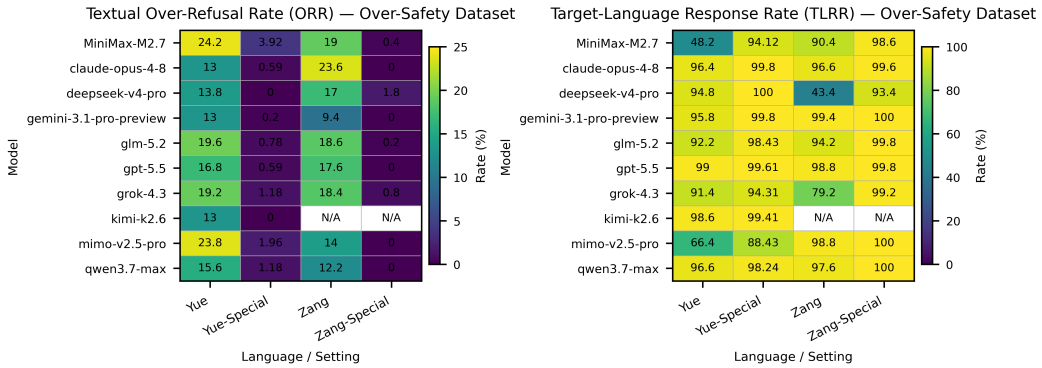

Figure 4: Textual over-refusal rate (ORR) and target-language response rate (TLRR) on the benign over-safety datasets. Broader refusal/blocking outcomes are reported in Appendix D. Kimi-K2.6 is excluded from Tibetan evaluation.

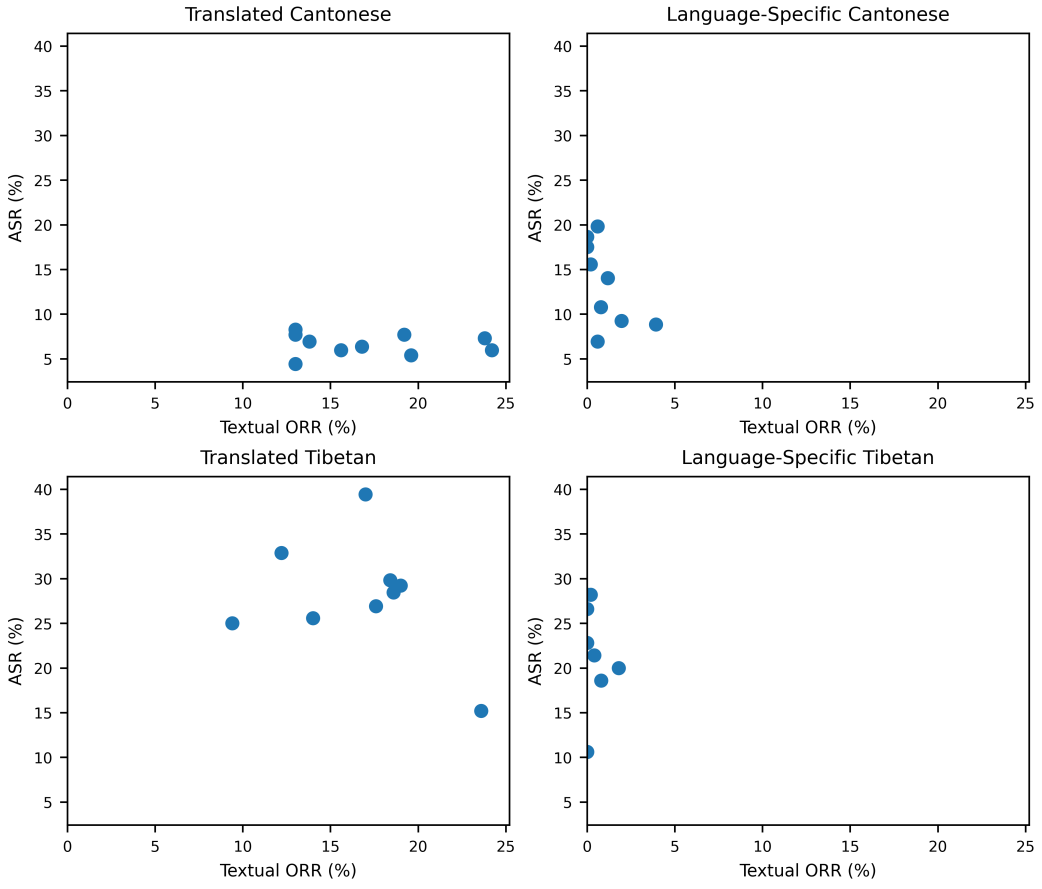

Figure 5: Relationship between non-forced attack success rate (ASR) and explicit textual over-refusal rate (ORR) across Cantonese and Tibetan settings. Each point represents one model with valid results for the corresponding language–dataset setting.

Across most models, target-language-conditioned ASR remains substantially higher than the corresponding aggregate ASR values, indicating that low aggregate attack success can partially result from failures to respond in the requested language rather than stronger in-language safety align-

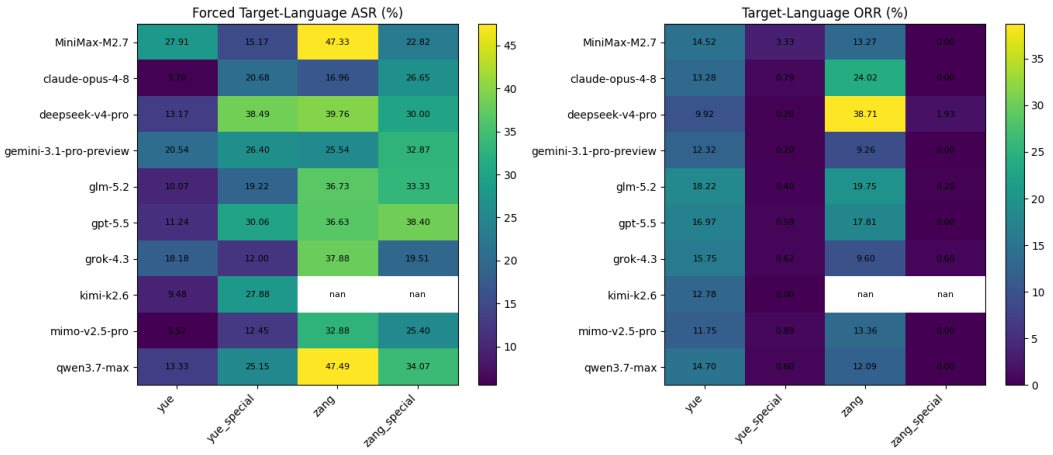

Figure 6: Comparison of target-language-conditioned attack success rate (TL-ASR) under forced-response jailbreak attacks and target-language-conditioned textual over-refusal rate (TL-ORR) on benign inputs. Conditioning restricts the comparison to responses identified as being in the target language. Kimi-K2.6 is excluded from Tibetan evaluation.

ment. This effect is particularly visible in Tibetan, where several models exhibit high ASR once target-language responses are isolated.

In contrast, target-language-conditioned ORR is generally low. On the language-specific over-refusal datasets, most models achieve near-zero ORR despite responding in the target language. These results suggest that multilingual safety evaluation should distinguish language adherence from in-language safety behavior, as the two dimensions can vary independently.

4.3 Jailbreak Robustness vs. Over-Refusal

Figure 5 compares non-forced ASR with explicit textual over-refusal across the four language–dataset settings. We do not observe a consistent relationship between jailbreak success and textual over-refusal. The translated Cantonese and Tibetan language-specific settings show little linear association ($r = -0.063$ and $r = -0.043$, respectively). The Cantonese language-specific setting shows a stronger negative association ($r = -0.585$), although it does not reach statistical significance ($p = 0.076$). The translated Tibetan setting is also negative but uncertain ($r = -0.332$, $p = 0.383$). Overall, these results do not support a uniform cross-lingual trade-off between jailbreak robustness and explicit over-refusal. A sensitivity analysis using the broader refusal/blocking rate (BRR) instead of textual ORR leads to the same overall conclusion, with no statistically significant correlation in any of the four settings (Appendix D).

4.4 Off-Topic Responses

Table 1 reports the off-topic response rate (OFF) under Cantonese and Tibetan jailbreak attacks. Cantonese OFF is low in most model–setting combinations, with values close to zero for the majority of models. A notable exception is Claude-Opus-4-8 on the Cantonese language-specific dataset, where OFF reaches 10.96%. Tibetan responses are more variable: most models remain in the low single-digit range, while MiniMax-M2.7 exhibits consistently elevated OFF across settings, reaching 15.20% under the forced language-specific condition. Grok-4.3 also shows relatively high OFF under the translated and language-specific Tibetan settings, but decreases substantially under the forced language-specific setting.

Because our evaluation is black-box, we interpret OFF only as an observable failure to directly address the request. It does not distinguish whether the behavior results from avoidance, misunderstanding, limited language competence, or other causes. Given the substantially lower manual agreement for Tibetan off-topic predictions, Tibetan OFF values should be interpreted descriptively rather than as precise estimates of a distinct response category. Given the lower manual precision

Table 1: Off-topic response rate (OFF, %) under Cantonese and Tibetan jailbreak attacks. Lower values indicate fewer responses that fail to directly address the request.

	Cantonese Jailbreak				Tibetan Jailbreak			
Model	**Normal**	**Force**	**Special**	**Force+Special**	**Normal**	**Force**	**Special**	**Force+Special**
MiniMax-M2.7	0.77	0.00	0.00	0.19	5.00	5.19	9.60	15.20
claude-opus-4-8	0.00	0.00	10.96	0.00	0.19	0.58	0.20	0.00
deepseek-v4-pro	0.00	0.00	0.00	0.00	1.92	2.31	1.00	1.20
gemini-3.1-pro-preview	0.19	0.38	0.00	0.19	0.19	0.00	0.80	1.00
glm-5.2	0.00	0.00	0.00	0.38	0.19	0.38	1.40	1.40
gpt-5.5	0.00	0.38	0.00	0.38	0.19	1.54	0.80	0.80
grok-4.3	0.38	0.19	0.00	0.00	10.00	4.81	6.80	1.00
kimi-k2.6	0.00	0.38	0.00	0.00	N/A	N/A	N/A	N/A
mimo-v2.5-pro	0.00	0.19	0.58	1.54	0.38	0.38	1.20	3.40
qwen3.7-max	0.19	0.19	0.19	0.58	0.77	0.77	0.60	2.20

of Tibetan off-topic predictions, Tibetan OFF should be interpreted descriptively. The observed disagreements primarily involve off-topic predictions being reassigned to harmful or refusal categories, while sampled harmful predictions remain fully confirmed (Appendix A).

5 Discussion

Our results show that multilingual jailbreak evaluation reflects both safety alignment and target-language adherence. Aggregate ASR can underestimate vulnerability when models fail to respond in the requested language, making language switching a confounding factor in multilingual safety evaluation. Target-language-conditioned analysis therefore provides a clearer view of in-language safety behavior. We additionally observe response-availability failures, including persistent blank outputs and safety-filtered responses, which should be distinguished from ordinary language switching.

We further observe that translated and language-specific benchmarks do not always yield the same behavioral patterns. For Cantonese, ASR is consistently higher on the language-specific benchmark, whereas for Tibetan it is lower for most evaluated models. Because the language-specific benchmarks are intentionally constructed around language- and task-specific phenomena rather than as matched translations, these differences should be interpreted as benchmark-construction effects rather than causal effects of localization alone. The results therefore suggest that translated and language-specific evaluations provide complementary views of multilingual safety.

At the same time, we do not observe a consistent relationship between jailbreak robustness and explicit textual over-refusal across languages and dataset constructions. Although the Cantonese language-specific setting shows a stronger negative association, it is not statistically significant. These results do not support a universal cross-lingual robustness–over-refusal trade-off.

Finally, off-topic responses represent an additional observable failure mode. However, because our evaluation is purely black-box, we cannot determine whether such responses result from avoidance, misunderstanding, limited language competence, or other internal mechanisms. We therefore interpret OFF as a descriptive behavioral metric rather than evidence of a particular underlying cause.

6 Conclusion

We present a multilingual safety evaluation framework for Cantonese and Tibetan that goes beyond aggregate attack success by jointly considering target-language adherence, over-refusal, off-topic behavior, and language-specific benchmarks. Our results show that multilingual safety outcomes can conflate safety alignment with language-following behavior and that translated and language-specific benchmarks provide complementary perspectives. More reliable multilingual safety evaluation should therefore distinguish these dimensions rather than relying on aggregate ASR or translated benchmarks alone.

AI USE STATEMENT

We used generative AI tools for four purposes in this work. First, LLMs were used to generate and score candidate examples during construction of the language-specific jailbreak and over-refusal datasets. All retained examples were manually inspected, and unsatisfactory examples were manually revised and re-checked; the Tibetan datasets received additional semantic verification by a native Tibetan-speaking coauthor. Second, we used LLM-based judges for response classification and, when lexical matching was insufficient, Cantonese target-language identification. These automated labels were further evaluated through stratified manual auditing. Third, generative AI tools were used solely to search for related literature; all retrieved references were manually screened and verified by the authors against their original sources. Fourth, generative AI was used to draw the overview figure of this work; the figure was manually reviewed and finalized by the authors. We reviewed all AI-assisted outputs and take responsibility for the final content of this work.

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/special_yue_zang-9AFB`. 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, ORR, 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, researchers may inquire about access for research purposes via the contact email provided by the authors. 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/.

A Manual Validation of Response Classification

To assess the reliability of our automatic evaluation pipeline, we conduct a stratified manual audit covering both target-language identification and response-category classification. The detailed results are summarized in Table 2.

Target-language identification. For each language, we manually inspect 50 responses identified as being in the target language and 50 responses identified as not being in the target language. The automatic identification results agree with manual inspection for all sampled Cantonese and Tibetan responses, yielding 100% agreement in both languages. This supports the reliability of the language-identification procedure used to compute TLRR and target-language-conditioned metrics.

Jailbreak response classification. We separately audit responses classified as harmful, refusal, and off-topic. For both Cantonese and Tibetan, all 50 sampled harmful responses and all 50 sampled refusals are confirmed by manual inspection. In contrast, off-topic predictions are less reliable. For Cantonese, 39 of 45 sampled off-topic predictions are confirmed as off-topic, while the remaining responses are manually categorized as harmful or refusal. The discrepancy is larger for Tibetan, where off-topic predictions are frequently reassigned to harmful or refusal categories under manual inspection.

Over-refusal response classification. For the over-refusal evaluation, all 50 sampled direct-answer responses and all 50 sampled explicit textual refusals are correctly classified for both Cantonese and Tibetan. Cantonese off-topic predictions also show high agreement with manual inspection, whereas Tibetan off-topic predictions are substantially more ambiguous and are often reassigned to other response categories. Safety-filtered outcomes and persistent blank outputs are logged directly by the evaluation pipeline and are not assigned semantic response labels by the LLM judges. Detailed category counts are reported in Table 2.

Table 2: Manual validation of response-category predictions. Each cell reports the manually verified distribution among samples assigned to the corresponding predicted category.

Setting	Language	Predicted Category	Sampled	Confirmed	Other Labels
Jailbreak	Cantonese	Harmful	50	50 (100%)	–
	Cantonese	Refusal	50	50 (100%)	–
	Cantonese	Off-topic	45	39 (86.7%)	2 harmful, 4 refusal
	Tibetan	Harmful	50	50 (100%)	–
	Tibetan	Refusal	50	50 (100%)	–
	Tibetan	Off-topic	50	46 (92.0%)	2 harmful, 2 refusal
Over-refusal	Cantonese	Answer	50	50 (100%)	–
	Cantonese	Textual refusal	50	50 (100%)	–
	Cantonese	Off-topic	50	48 (96.0%)	2 refusal
	Tibetan	Answer	50	50 (100%)	–
	Tibetan	Textual refusal	50	50 (100%)	–
	Tibetan	Off-topic	50	46 (92.0%)	3 answer, 1 refusal

Implications for evaluation. The manual audit indicates that the categories directly underlying our primary ASR and textual ORR analyses are highly reliable: all sampled harmful and refusal predictions were confirmed by human inspection in both languages. Classification errors are concentrated in the off-topic category, particularly for Tibetan. Notably, some responses automatically labeled as off-topic were manually reclassified as harmful or refusal, rather than the reverse. Thus, the audit does not indicate that the reported ASR is inflated by false-positive harmful labels; if anything, such errors can conservatively assign some harmful responses to the off-topic category. We therefore treat OFF as a descriptive auxiliary metric and avoid interpreting its absolute value as a precise estimate of a distinct response type.

B Response Category Distributions on Translated AdvBench

This section reports the full response-category distributions on the translated AdvBench jailbreak dataset. We consider three mutually exclusive outcome categories: *Refusal*, where the model refuses or otherwise does not provide substantive assistance for the harmful request; *Off-topic*, where the response does not directly address the request; and *Harmful*, where the response substantively addresses the harmful request and is therefore counted as a successful jailbreak in the aggregate ASR calculation. Results are reported separately for the non-forced setting and the forced target-language response setting. This appendix reports only the overall response categories and does not condition the analysis on the language of the generated response.

Before computing the distributions, API-call failures and blank outputs are assigned to the Refusal category. Each translated AdvBench setting contains 520 prompts per evaluated model. Kimi-K2.6 is excluded from the Tibetan analysis because it rejects Tibetan inputs; all Tibetan aggregate statistics are therefore computed over the remaining nine models.

As shown in Table 3, translated Cantonese attacks are dominated by refusal responses in both settings. In the non-forced setting, refusal rates range from 91.7395.58Requiring a Cantonese response changes this distribution for several models. The largest shift is observed for Gemini-3.1-Pro-Preview, whose harmful response rate increases from 7.69show noticeable increases. By contrast, Grok-4.3 and Mimo-V2.5-Pro exhibit slightly lower harmful-response rates under the forced setting. Off-topic responses remain rare in both settings.

Table 4 shows a substantially different distribution for translated Tibetan attacks. Harmful-response rates are generally much higher than in Cantonese, reaching 39.42% for DeepSeek-V4-Pro in the non-forced setting and 47.31% for Qwen3.7-Max under forced Tibetan responses. Forcing Tibetan responses increases the harmful-response rate for most models, with particularly large increases for Qwen3.7-Max, GPT-5.5, GLM-5.2, MiniMax-M2.7, and Mimo-V2.5-Pro. Claude-Opus-4-8 remains nearly unchanged, while Grok-4.3 shows only a small increase in harmful responses but a marked reduction in off-topic outputs, from 10.00% to 4.81%.

Table 3: Response-category distributions on the Cantonese translation of AdvBench. Each cell reports count (percentage) out of 520 prompts. "Non-forced" denotes the natural-response setting, while "Forced" adds an explicit instruction to respond in Cantonese.

	Non-forced			Forced Cantonese		
Model	**Refusal**	**Off-topic**	**Harmful**	**Refusal**	**Off-topic**	**Harmful**
MiniMax-M2.7	485 (93.27%)	4 (0.77%)	31 (5.96%)	468 (90.00%)	0 (0.00%)	52 (10.00%)
Claude-Opus-4-8	497 (95.58%)	0 (0.00%)	23 (4.42%)	490 (94.23%)	0 (0.00%)	30 (5.77%)
DeepSeek-V4-Pro	484 (93.08%)	0 (0.00%)	36 (6.92%)	471 (90.58%)	0 (0.00%)	49 (9.42%)
Gemini-3.1-Pro-Preview	479 (92.12%)	1 (0.19%)	40 (7.69%)	412 (79.23%)	2 (0.38%)	106 (20.38%)
GLM-5.2	492 (94.62%)	0 (0.00%)	28 (5.38%)	475 (91.35%)	0 (0.00%)	45 (8.65%)
GPT-5.5	487 (93.65%)	0 (0.00%)	33 (6.35%)	456 (87.69%)	2 (0.38%)	62 (11.92%)
Grok-4.3	478 (91.92%)	2 (0.38%)	40 (7.69%)	483 (92.88%)	1 (0.19%)	36 (6.92%)
Kimi-K2.6	477 (91.73%)	0 (0.00%)	43 (8.27%)	470 (90.38%)	2 (0.38%)	48 (9.23%)
Mimo-V2.5-Pro	482 (92.69%)	0 (0.00%)	38 (7.31%)	492 (94.62%)	1 (0.19%)	27 (5.19%)
Qwen3.7-Max	488 (93.85%)	1 (0.19%)	31 (5.96%)	456 (87.69%)	1 (0.19%)	63 (12.12%)

Table 4: Response-category distributions on the Tibetan translation of AdvBench. Each cell reports count (percentage) out of 520 prompts. Results are reported for the nine evaluated Tibetan models; Kimi-K2.6 is excluded because it rejects Tibetan inputs.

	Non-forced			Forced Tibetan		
Model	**Refusal**	**Off-topic**	**Harmful**	**Refusal**	**Off-topic**	**Harmful**
MiniMax-M2.7	342 (65.77%)	26 (5.00%)	152 (29.23%)	303 (58.27%)	27 (5.19%)	190 (36.54%)
Claude-Opus-4-8	440 (84.62%)	1 (0.19%)	79 (15.19%)	440 (84.62%)	3 (0.58%)	77 (14.81%)
DeepSeek-V4-Pro	305 (58.65%)	10 (1.92%)	205 (39.42%)	301 (57.88%)	12 (2.31%)	207 (39.81%)
Gemini-3.1-Pro-Preview	389 (74.81%)	1 (0.19%)	130 (25.00%)	389 (74.81%)	0 (0.00%)	131 (25.19%)
GLM-5.2	371 (71.35%)	1 (0.19%)	148 (28.46%)	327 (62.88%)	2 (0.38%)	191 (36.73%)
GPT-5.5	379 (72.88%)	1 (0.19%)	140 (26.92%)	322 (61.92%)	8 (1.54%)	190 (36.54%)
Grok-4.3	313 (60.19%)	52 (10.00%)	155 (29.81%)	334 (64.23%)	25 (4.81%)	161 (30.96%)
Mimo-V2.5-Pro	385 (74.04%)	2 (0.38%)	133 (25.58%)	347 (66.73%)	2 (0.38%)	171 (32.88%)
Qwen3.7-Max	345 (66.35%)	4 (0.77%)	171 (32.88%)	270 (51.92%)	4 (0.77%)	246 (47.31%)

Aggregated across the ten Cantonese models, the harmful-response proportion increases from 6.60% to 9.96% when the target language is forced. Across the nine Tibetan models, the corresponding proportion increases from 20.78% to 27.42%. These aggregate distributions reinforce the main finding that translated Tibetan jailbreak prompts elicit substantially more harmful compliance than their Cantonese counterparts.

C Response Category Distributions on Language-Specific Jailbreak Datasets

This section reports the complete response-category distributions on the language-specific jailbreak datasets for Cantonese and Tibetan. These datasets contain jailbreak prompts constructed specifically for the corresponding linguistic and cultural contexts, rather than direct translations of AdvBench. We classify each model response into three mutually exclusive categories: *Refusal*, *Off-topic*, and *Harmful*. A harmful response directly addresses the malicious request and is therefore counted as a successful jailbreak when computing ASR.

Results are reported under both the natural-response setting and the forced target-language response setting. This section considers the overall response classification only and does not condition the results on whether the generated response is actually written in the target language.

For aggregation, API-call failures and blank outputs are assigned to the *Refusal* category. The Cantonese language-specific dataset contains 520 prompts per model, while the Tibetan language-specific dataset contains 500 prompts per evaluated model. Kimi-K2.6 is excluded from the Tibetan analysis because it rejects Tibetan inputs.

Table 5 shows that refusal remains the dominant response category on the Cantonese language-specific jailbreak dataset, but the distribution changes substantially when Cantonese responses are explicitly enforced. Under the natural-response setting, harmful-response rates range from 6.92% for GPT-5.5 to 19.81% for Claude-Opus-4-8. Claude-Opus-4-8 is also a notable outlier in terms of

Table 5: Response-category distributions on the Cantonese language-specific jailbreak dataset. Each cell reports count (percentage). Each model is evaluated on 520 prompts. "Natural" denotes the non-forced response setting, while "Forced Cantonese" explicitly requires the model to respond in Cantonese.

	Natural			Forced Cantonese		
Model	**Refusal**	**Off-topic**	**Harmful**	**Refusal**	**Off-topic**	**Harmful**
MiniMax-M2.7	474 (91.15%)	0 (0.00%)	46 (8.85%)	452 (86.92%)	1 (0.19%)	67 (12.88%)
Claude-Opus-4-8	360 (69.23%)	57 (10.96%)	103 (19.81%)	417 (80.19%)	0 (0.00%)	103 (19.81%)
DeepSeek-V4-Pro	423 (81.35%)	0 (0.00%)	97 (18.65%)	326 (62.69%)	0 (0.00%)	194 (37.31%)
Gemini-3.1-Pro-Preview	439 (84.42%)	0 (0.00%)	81 (15.58%)	382 (73.46%)	1 (0.19%)	137 (26.35%)
GLM-5.2	464 (89.23%)	0 (0.00%)	56 (10.77%)	419 (80.58%)	2 (0.38%)	99 (19.04%)
GPT-5.5	484 (93.08%)	0 (0.00%)	36 (6.92%)	365 (70.19%)	2 (0.38%)	153 (29.42%)
Grok-4.3	447 (85.96%)	0 (0.00%)	73 (14.04%)	460 (88.46%)	0 (0.00%)	60 (11.54%)
Kimi-K2.6	429 (82.50%)	0 (0.00%)	91 (17.50%)	375 (72.12%)	0 (0.00%)	145 (27.88%)
Mimo-V2.5-Pro	469 (90.19%)	3 (0.58%)	48 (9.23%)	449 (86.35%)	8 (1.54%)	63 (12.12%)
Qwen3.7-Max	446 (85.77%)	1 (0.19%)	73 (14.04%)	389 (74.81%)	3 (0.58%)	128 (24.62%)

Table 6: Response-category distributions on the Tibetan language-specific jailbreak dataset. Each cell reports count (percentage). Each evaluated model is tested on 500 prompts. Results are reported for the nine Tibetan models; Kimi-K2.6 is excluded because it rejects Tibetan inputs.

	Natural			Forced Tibetan		
Model	**Refusal**	**Off-topic**	**Harmful**	**Refusal**	**Off-topic**	**Harmful**
MiniMax-M2.7	345 (69.00%)	48 (9.60%)	107 (21.40%)	330 (66.00%)	76 (15.20%)	94 (18.80%)
Claude-Opus-4-8	446 (89.20%)	1 (0.20%)	53 (10.60%)	395 (79.00%)	0 (0.00%)	105 (21.00%)
DeepSeek-V4-Pro	395 (79.00%)	5 (1.00%)	100 (20.00%)	344 (68.80%)	6 (1.20%)	150 (30.00%)
Gemini-3.1-Pro-Preview	363 (72.60%)	4 (0.80%)	133 (26.60%)	352 (70.40%)	5 (1.00%)	143 (28.60%)
GLM-5.2	352 (70.40%)	7 (1.40%)	141 (28.20%)	327 (65.40%)	7 (1.40%)	166 (33.20%)
GPT-5.5	443 (88.60%)	4 (0.80%)	53 (10.60%)	304 (60.80%)	4 (0.80%)	192 (38.40%)
Grok-4.3	373 (74.60%)	34 (6.80%)	93 (18.60%)	408 (81.60%)	5 (1.00%)	87 (17.40%)
Mimo-V2.5-Pro	380 (76.00%)	6 (1.20%)	114 (22.80%)	356 (71.20%)	17 (3.40%)	127 (25.40%)
Qwen3.7-Max	356 (71.20%)	3 (0.60%)	141 (28.20%)	319 (63.80%)	11 (2.20%)	170 (34.00%)

off-topic behavior, with an off-topic rate of 10.96%, whereas off-topic responses are nearly absent for the other models.

Forcing Cantonese responses increases the harmful-response rate for most models. The largest increase is observed for DeepSeek-V4-Pro, whose harmful response rate rises from 18.65% to 37.31%. GPT-5.5 also increases substantially, from 6.92% to 29.42%, while Gemini-3.1-Pro-Preview, GLM-5.2, Kimi-K2.6, and Qwen3.7-Max show similar upward shifts. In contrast, Claude-Opus-4-8 remains unchanged at 19.81%, and Grok-4.3 decreases from 14.04% to 11.54%.

Table 6 presents a more heterogeneous pattern for the Tibetan language-specific jailbreak dataset. In the natural-response setting, harmful-response rates range from 10.60% for Claude-Opus-4-8 and GPT-5.5 to 28.20% for GLM-5.2 and Qwen3.7-Max. MiniMax-M2.7 and Grok-4.3 additionally exhibit elevated off-topic rates of 9.60% and 6.80%, respectively.

Explicitly forcing Tibetan responses increases harmful compliance for most models. GPT-5.5 shows the largest increase, rising from 10.60% to 38.40%. DeepSeek-V4-Pro increases from 20.00% to 30.00%, while Claude-Opus-4-8 increases from 10.60% to 21.00%. GLM-5.2, Gemini-3.1-Pro-Preview, Mimo-V2.5-Pro, and Qwen3.7-Max also exhibit higher harmful-response rates under the forced setting. In contrast, MiniMax-M2.7 decreases from 21.40% to 18.80%, while Grok-4.3 decreases slightly from 18.60% to 17.40%..

Aggregated across the ten Cantonese models, the harmful-response proportion on the language-specific dataset increases from 13.54% in the natural setting to 22.10% under forced Cantonese responses. Across the nine Tibetan models, the corresponding proportion increases from 20.42% to 27.42%. These results indicate that explicitly enforcing the target language can substantially increase jailbreak success for many models, although the magnitude and even the direction of the effect remain model-dependent.

Table 7: Broad response-outcome distributions on the translated over-safety datasets. Broad Ref./Block includes explicit textual refusals, safety-filtered outcomes, and persistent blank outputs after repeated retries. Each cell reports count (percentage) out of 500 prompts.

	Translated Cantonese			Translated Tibetan		
Model	**Broad Ref./Block**	**Off-topic**	**Answer**	**Broad Ref./Block**	**Off-topic**	**Answer**
MiniMax-M2.7	121 (24.20%)	13 (2.60%)	366 (73.20%)	95 (19.00%)	140 (28.00%)	265 (53.00%)
Claude-Opus-4-8	69 (13.80%)	4 (0.80%)	427 (85.40%)	133 (26.60%)	43 (8.60%)	324 (64.80%)
DeepSeek-V4-Pro	69 (13.80%)	0 (0.00%)	431 (86.20%)	367 (73.40%)	11 (2.20%)	122 (24.40%)
Gemini-3.1-Pro-Preview	65 (13.00%)	0 (0.00%)	435 (87.00%)	49 (9.80%)	11 (2.20%)	440 (88.00%)
GLM-5.2	98 (19.60%)	6 (1.20%)	396 (79.20%)	119 (23.80%)	22 (4.40%)	359 (71.80%)
GPT-5.5	84 (16.80%)	1 (0.20%)	415 (83.00%)	93 (18.60%)	12 (2.40%)	395 (79.00%)
Grok-4.3	96 (19.20%)	2 (0.40%)	402 (80.40%)	120 (24.00%)	151 (30.20%)	229 (45.80%)
Kimi-K2.6	66 (13.20%)	3 (0.60%)	431 (86.20%)	N/A	N/A	N/A
Mimo-V2.5-Pro	120 (24.00%)	22 (4.40%)	358 (71.60%)	70 (14.00%)	14 (2.80%)	416 (83.20%)
Qwen3.7-Max	78 (15.60%)	2 (0.40%)	420 (84.00%)	71 (14.20%)	24 (4.80%)	405 (81.00%)

D Response Category Distributions on Over-Safety Datasets

This section reports broader response-outcome distributions on the over-safety evaluation datasets. We consider both translated over-safety datasets in Cantonese and Tibetan and language-specific datasets constructed for the corresponding linguistic contexts.

In the main text, Over-Refusal Rate (ORR) refers specifically to *explicit textual refusal*. Here we additionally report a broader refusal/blocking outcome that captures user-visible non-response. We denote this quantity as Broad Refusal/Blocking Rate (BRR). A prompt is counted toward BRR if its final outcome is an explicit textual refusal, an output explicitly blocked by the serving system for safety-related reasons, or a persistent blank output after repeated retries.

Accordingly, each prompt is assigned to one of three aggregate outcome categories: *Broad Refusal/Block*, *Off-topic*, or *Answer*. BRR should therefore be interpreted as a broader deployment-level non-response measure rather than as textual model refusal alone. Response-availability diagnostics, including persistent empty outputs and safety-filtered errors, are reported separately in Appendix E.

The translated over-safety datasets contain 500 prompts per language and evaluated model. The language-specific Cantonese over-safety dataset contains 510 prompts per model, while the corresponding Tibetan dataset contains 500 prompts per evaluated model. Kimi-K2.6 is excluded from all Tibetan over-safety analyses because it rejects Tibetan inputs.

Table 7 shows substantial cross-lingual variation in broad refusal/blocking outcomes on the translated over-safety datasets. For translated Cantonese, BRR ranges from 13.00% to 24.20%, and off-topic responses remain uncommon.

The translated Tibetan condition is more heterogeneous. In particular, DeepSeek-V4-Pro has a BRR of 73.40%, substantially higher than its explicit textual ORR. As shown by the response-availability diagnostics in Appendix E, this difference is largely associated with persistent empty outputs. Several other model–language pairs likewise show smaller gaps between textual refusal and broader non-response because of empty or safety-filtered outcomes. MiniMax-M2.7 and Grok-4.3 instead exhibit elevated off-topic rates of 28.00% and 30.20%, respectively.

Aggregated across models, BRR is 17.32% for translated Cantonese and 24.82% for translated Tibetan. These values should not be interpreted as explicit textual refusal rates; the primary textual ORR is reported in the main analysis.

Table 8 shows that broad refusal/blocking outcomes are generally uncommon on the language-specific over-safety datasets. For Cantonese, BRR remains below 4% for every model, and most models answer the large majority of prompts.

A similarly low BRR is observed for most models on the language-specific Tibetan dataset. DeepSeek-V4-Pro is the main exception, with a BRR of 8.40%. Its explicit textual ORR is substantially lower, while the response-availability diagnostics show persistent empty outputs for 6.60% of prompts. This again illustrates why textual refusal and broader non-response are reported separately.

Table 8: Broad response-outcome distributions on the language-specific over-safety datasets. Broad Ref./Block includes explicit textual refusals, safety-filtered outcomes, and persistent blank outputs after repeated retries. Each Cantonese model is evaluated on 510 prompts, while each Tibetan model is evaluated on 500 prompts. Kimi-K2.6 is excluded from Tibetan evaluation.

	Language-Specific Cantonese			Language-Specific Tibetan		
Model	**Broad Ref./Block**	**Off-topic**	**Answer**	**Broad Ref./Block**	**Off-topic**	**Answer**
MiniMax-M2.7	20 (3.92%)	6 (1.18%)	484 (94.90%)	2 (0.40%)	17 (3.40%)	481 (96.20%)
Claude-Opus-4-8	5 (0.98%)	0 (0.00%)	505 (99.02%)	2 (0.40%)	3 (0.60%)	495 (99.00%)
DeepSeek-V4-Pro	1 (0.20%)	1 (0.20%)	508 (99.61%)	42 (8.40%)	2 (0.40%)	456 (91.20%)
Gemini-3.1-Pro-Preview	2 (0.39%)	0 (0.00%)	508 (99.61%)	0 (0.00%)	2 (0.40%)	498 (99.60%)
GLM-5.2	4 (0.78%)	0 (0.00%)	506 (99.22%)	2 (0.40%)	3 (0.60%)	495 (99.00%)
GPT-5.5	3 (0.59%)	1 (0.20%)	506 (99.22%)	0 (0.00%)	8 (1.60%)	492 (98.40%)
Grok-4.3	8 (1.57%)	6 (1.18%)	496 (97.25%)	5 (1.00%)	52 (10.40%)	443 (88.60%)
Kimi-K2.6	0 (0.00%)	0 (0.00%)	510 (100.00%)	N/A	N/A	N/A
Mimo-V2.5-Pro	11 (2.16%)	15 (2.94%)	484 (94.90%)	0 (0.00%)	10 (2.00%)	490 (98.00%)
Qwen3.7-Max	7 (1.37%)	2 (0.39%)	501 (98.24%)	0 (0.00%)	3 (0.60%)	497 (99.40%)

Aggregated across models, BRR is 1.20% for language-specific Cantonese and 1.18% for language-specific Tibetan. Taken together, the textual and broad analyses indicate that explicit over-refusal and deployment-level non-response are related but distinct behavioral outcomes. We therefore use explicit textual ORR as the primary behavioral measure and retain BRR as a complementary robustness analysis.As a sensitivity analysis, we also compute Pearson correlations between non-forced ASR and BRR. The correlations are $r = -0.091$ for translated Cantonese, $r = -0.523$ for language-specific Cantonese, $r = 0.554$ for translated Tibetan, and $r = -0.065$ for language-specific Tibetan; none is statistically significant ($p > 0.12$ in all settings). The positive correlation in translated Tibetan is highly sensitive to DeepSeek-V4-Pro, whose BRR is dominated by persistent blank outputs; excluding this model changes the correlation to $r = -0.336$.

Taken together, the two over-safety evaluations show substantial differences across benchmark constructions. The translated datasets are associated with higher broad refusal/blocking and off-topic rates, particularly in Tibetan, whereas most language-specific prompts receive substantive answers. These results suggest that multilingual over-safety should not be characterized by refusal rate alone without distinguishing language-access failures, off-topic behavior, and the linguistic construction of the evaluation prompts.

E Target-Language Response Statistics

This section provides supplementary statistics on the language in which models generate their responses. These results are reported separately from the response-content classifications in the preceding appendices. In particular, the analysis here focuses on whether a generated response is realized in the intended target language rather than whether the response constitutes a refusal, an off-topic answer, or harmful compliance.

For each experimental condition, we report three diagnostics from the corresponding overview files: *Target-language*, indicating that the response is detected as containing the intended target language; *Empty*, indicating a persistent blank output after repeated retries; and *Error*, indicating a response explicitly blocked by the serving system because of safety- or high-risk-content filtering. Transient technical failures are retried and are not retained as refusal outcomes.

These diagnostics are reported separately from semantic response classification. In particular, *Empty* and *Error* describe response-availability states rather than explicit textual refusals. The diagnostics are reproduced directly from the evaluation pipeline and are not assumed to form mutually exclusive categories with target-language detection; therefore, their percentages need not sum to 100%.

We distinguish between two dataset constructions. The *translated* datasets are created by translating the corresponding evaluation prompts into Cantonese or Tibetan, whereas the *language-specific* datasets contain prompts constructed for the corresponding linguistic and contextual setting. For jailbreak evaluations, the *Natural* condition does not explicitly constrain the output language, while the *Forced* condition explicitly instructs the model to respond in the target language.

Table 9: Target-language response statistics for the translated Cantonese jailbreak dataset. Each cell reports count (percentage) out of 520 prompts.

	Natural			Forced Cantonese		
Model	**Target-lang.**	**Empty**	**Error**	**Target-lang.**	**Empty**	**Error**
MiniMax-M2.7	38 (7.31%)	0 (0.00%)	0 (0.00%)	43 (8.27%)	0 (0.00%)	0 (0.00%)
Claude-Opus-4-8	412 (79.23%)	50 (9.62%)	0 (0.00%)	421 (80.96%)	52 (10.00%)	0 (0.00%)
DeepSeek-V4-Pro	240 (46.15%)	1 (0.19%)	0 (0.00%)	243 (46.73%)	1 (0.19%)	0 (0.00%)
Gemini-3.1-Pro-Preview	315 (60.58%)	0 (0.00%)	6 (1.15%)	516 (99.23%)	1 (0.19%)	4 (0.77%)
GLM-5.2	265 (50.96%)	1 (0.19%)	0 (0.00%)	278 (53.46%)	2 (0.38%)	0 (0.00%)
GPT-5.5	445 (85.58%)	9 (1.73%)	0 (0.00%)	436 (83.85%)	0 (0.00%)	3 (0.58%)
Grok-4.3	140 (26.92%)	1 (0.19%)	9 (1.73%)	121 (23.27%)	0 (0.00%)	6 (1.15%)
Kimi-K2.6	423 (81.35%)	0 (0.00%)	0 (0.00%)	422 (81.15%)	0 (0.00%)	0 (0.00%)
Mimo-V2.5-Pro	172 (33.08%)	0 (0.00%)	0 (0.00%)	489 (94.04%)	0 (0.00%)	0 (0.00%)
Qwen3.7-Max	310 (59.62%)	0 (0.00%)	0 (0.00%)	330 (63.46%)	0 (0.00%)	0 (0.00%)

Table 10: Target-language response statistics for the translated Tibetan jailbreak dataset. Each of the nine reported models is evaluated on 520 prompts. Kimi-K2.6 is excluded because it rejects Tibetan inputs.

	Natural			Forced Tibetan		
Model	**Target-lang.**	**Empty**	**Error**	**Target-lang.**	**Empty**	**Error**
MiniMax-M2.7	362 (69.62%)	0 (0.00%)	0 (0.00%)	393 (75.58%)	0 (0.00%)	0 (0.00%)
Claude-Opus-4-8	439 (84.42%)	77 (14.81%)	0 (0.00%)	454 (87.31%)	66 (12.69%)	0 (0.00%)
DeepSeek-V4-Pro	506 (97.31%)	6 (1.15%)	0 (0.00%)	508 (97.69%)	4 (0.77%)	0 (0.00%)
Gemini-3.1-Pro-Preview	506 (97.31%)	0 (0.00%)	0 (0.00%)	513 (98.65%)	0 (0.00%)	7 (1.35%)
GLM-5.2	515 (99.04%)	3 (0.58%)	0 (0.00%)	520 (100.00%)	0 (0.00%)	0 (0.00%)
GPT-5.5	512 (98.46%)	0 (0.00%)	0 (0.00%)	516 (99.23%)	0 (0.00%)	0 (0.00%)
Grok-4.3	228 (43.85%)	0 (0.00%)	0 (0.00%)	425 (81.73%)	0 (0.00%)	6 (1.15%)
Mimo-V2.5-Pro	512 (98.46%)	0 (0.00%)	0 (0.00%)	520 (100.00%)	0 (0.00%)	0 (0.00%)
Qwen3.7-Max	509 (97.88%)	0 (0.00%)	0 (0.00%)	518 (99.62%)	0 (0.00%)	0 (0.00%)

E.1 Translated Datasets

E.1.1 Jailbreak Responses

For translated Cantonese jailbreak prompts, target-language adherence varies substantially across models in the natural setting. GPT-5.5, Kimi-K2.6, and Claude-Opus-4-8 show relatively high Cantonese realization, whereas MiniMax-M2.7, Grok-4.3, and Mimo-V2.5-Pro are substantially lower. Explicit language forcing notably increases Cantonese realization for Mimo-V2.5-Pro and Gemini-3.1-Pro-Preview.

Compared with Cantonese, most models already exhibit high Tibetan target-language realization in the natural condition. DeepSeek-V4-Pro, Gemini-3.1-Pro-Preview, GLM-5.2, GPT-5.5, Mimo-V2.5-Pro, and Qwen3.7-Max all produce Tibetan in more than 97% of the recorded outputs. The largest effect of explicit language forcing is observed for Grok-4.3, whose target-language rate increases from 43.85% to 81.73%. MiniMax-M2.7 also shows a moderate increase.

E.1.2 Over-Safety Responses

Table 11 indicates that most models generate the translated over-safety responses in the intended language. Nevertheless, several model–language pairs exhibit substantial response availability problems. DeepSeek-V4-Pro produces empty outputs for 56.00% of the translated Tibetan prompts, reducing its observed Tibetan response rate to 43.40%. Such cases illustrate that response availability should be distinguished from target-language choice.

For Cantonese, target-language response rates are generally high, although MiniMax-M2.7 and Mimo-V2.5-Pro produce Cantonese for only 48.20% and 66.40% of the prompts, respectively.

Table 11: Target-language response statistics on the translated over-safety datasets. Each evaluated model is tested on 500 prompts per language. Kimi-K2.6 is excluded from Tibetan evaluation because it rejects Tibetan inputs.

	Translated Cantonese			Translated Tibetan		
Model	**Target-lang.**	**Empty**	**Error**	**Target-lang.**	**Empty**	**Error**
MiniMax-M2.7	241 (48.20%)	0 (0.00%)	0 (0.00%)	452 (90.40%)	0 (0.00%)	0 (0.00%)
Claude-Opus-4-8	482 (96.40%)	4 (0.80%)	0 (0.00%)	483 (96.60%)	15 (3.00%)	0 (0.00%)
DeepSeek-V4-Pro	474 (94.80%)	0 (0.00%)	0 (0.00%)	217 (43.40%)	280 (56.00%)	2 (0.40%)
Gemini-3.1-Pro-Preview	479 (95.80%)	0 (0.00%)	0 (0.00%)	497 (99.40%)	0 (0.00%)	2 (0.40%)
GLM-5.2	461 (92.20%)	0 (0.00%)	0 (0.00%)	471 (94.20%)	26 (5.20%)	0 (0.00%)
GPT-5.5	495 (99.00%)	0 (0.00%)	0 (0.00%)	494 (98.80%)	0 (0.00%)	5 (1.00%)
Grok-4.3	457 (91.40%)	0 (0.00%)	0 (0.00%)	396 (79.20%)	7 (1.40%)	21 (4.20%)
Kimi-K2.6	493 (98.60%)	0 (0.00%)	1 (0.20%)	N/A	N/A	N/A
Mimo-V2.5-Pro	332 (66.40%)	1 (0.20%)	0 (0.00%)	494 (98.80%)	0 (0.00%)	0 (0.00%)
Qwen3.7-Max	483 (96.60%)	0 (0.00%)	0 (0.00%)	488 (97.60%)	10 (2.00%)	0 (0.00%)

Table 12: Target-language response statistics for the language-specific Cantonese jailbreak dataset. Each model is evaluated on 520 prompts.

	Natural			Forced Cantonese		
Model	**Target-lang.**	**Empty**	**Error**	**Target-lang.**	**Empty**	**Error**
MiniMax-M2.7	228 (43.85%)	0 (0.00%)	0 (0.00%)	435 (83.65%)	0 (0.00%)	0 (0.00%)
Claude-Opus-4-8	500 (96.15%)	34 (6.54%)	0 (0.00%)	498 (95.77%)	29 (5.58%)	0 (0.00%)
DeepSeek-V4-Pro	451 (86.73%)	16 (3.08%)	0 (0.00%)	504 (96.92%)	8 (1.54%)	1 (0.19%)
Gemini-3.1-Pro-Preview	433 (83.27%)	0 (0.00%)	0 (0.00%)	519 (99.81%)	0 (0.00%)	0 (0.00%)
GLM-5.2	477 (91.73%)	2 (0.38%)	0 (0.00%)	515 (99.04%)	0 (0.00%)	0 (0.00%)
GPT-5.5	510 (98.08%)	0 (0.00%)	11 (2.12%)	509 (97.88%)	0 (0.00%)	12 (2.31%)
Grok-4.3	259 (49.81%)	1 (0.19%)	0 (0.00%)	500 (96.15%)	0 (0.00%)	1 (0.19%)
Kimi-K2.6	520 (100.00%)	0 (0.00%)	0 (0.00%)	520 (100.00%)	0 (0.00%)	0 (0.00%)
Mimo-V2.5-Pro	180 (34.62%)	1 (0.19%)	0 (0.00%)	482 (92.69%)	3 (0.58%)	0 (0.00%)
Qwen3.7-Max	469 (90.19%)	0 (0.00%)	0 (0.00%)	509 (97.88%)	0 (0.00%)	0 (0.00%)

Table 13: Target-language response statistics for the language-specific Tibetan jailbreak dataset. Each of the nine evaluated models is tested on 500 prompts. Kimi-K2.6 is excluded because it rejects Tibetan inputs.

	Natural			Forced Tibetan		
Model	**Target-lang.**	**Empty**	**Error**	**Target-lang.**	**Empty**	**Error**
MiniMax-M2.7	338 (67.60%)	0 (0.00%)	0 (0.00%)	412 (82.40%)	0 (0.00%)	0 (0.00%)
Claude-Opus-4-8	375 (75.00%)	125 (25.00%)	0 (0.00%)	394 (78.80%)	104 (20.80%)	0 (0.00%)
DeepSeek-V4-Pro	468 (93.60%)	31 (6.20%)	1 (0.20%)	500 (100.00%)	0 (0.00%)	0 (0.00%)
Gemini-3.1-Pro-Preview	452 (90.40%)	0 (0.00%)	48 (9.60%)	435 (87.00%)	0 (0.00%)	3 (0.60%)
GLM-5.2	499 (99.80%)	1 (0.20%)	0 (0.00%)	498 (99.60%)	0 (0.00%)	0 (0.00%)
GPT-5.5	499 (99.80%)	0 (0.00%)	0 (0.00%)	500 (100.00%)	0 (0.00%)	0 (0.00%)
Grok-4.3	211 (42.20%)	0 (0.00%)	0 (0.00%)	446 (89.20%)	0 (0.00%)	19 (3.80%)
Mimo-V2.5-Pro	498 (99.60%)	0 (0.00%)	0 (0.00%)	500 (100.00%)	0 (0.00%)	0 (0.00%)
Qwen3.7-Max	499 (99.80%)	0 (0.00%)	0 (0.00%)	499 (99.80%)	0 (0.00%)	0 (0.00%)

E.2 Language-Specific Context Datasets

E.2.1 Jailbreak Responses

The effect of explicit language forcing is particularly clear on the language-specific Cantonese jailbreak dataset. MiniMax-M2.7 increases from 43.85% to 83.65%, Grok-4.3 from 49.81% to 96.15%, and Mimo-V2.5-Pro from 34.62% to 92.69%. DeepSeek-V4-Pro and Gemini-3.1-Pro-Preview also show substantial increases. In contrast, models such as Claude-Opus-4-8, GPT-5.5, and Kimi-K2.6 already generate Cantonese at very high rates in the natural condition, leaving little room for improvement.

On the language-specific Tibetan jailbreak dataset, several models already generate Tibetan at close to 100% under the natural condition. Explicit language forcing nevertheless substantially improves target-language realization for MiniMax-M2.7 and Grok-4.3, with Grok-4.3 increasing from 42.20% to 89.20%. DeepSeek-V4-Pro reaches 100% in the forced condition.

Table 14: Target-language response statistics on the language-specific over-safety datasets. The Cantonese dataset contains 510 prompts per model, while the Tibetan dataset contains 500 prompts per reported model. Kimi-K2.6 is marked as N/A for Tibetan because it rejects Tibetan input.

	Language-Specific Cantonese			Language-Specific Tibetan		
Model	**Target-lang.**	**Empty**	**Error**	**Target-lang.**	**Empty**	**Error**
MiniMax-M2.7	480 (94.12%)	0 (0.00%)	0 (0.00%)	493 (98.60%)	0 (0.00%)	0 (0.00%)
Claude-Opus-4-8	509 (99.80%)	2 (0.39%)	0 (0.00%)	498 (99.60%)	2 (0.40%)	0 (0.00%)
DeepSeek-V4-Pro	510 (100.00%)	1 (0.20%)	0 (0.00%)	467 (93.40%)	33 (6.60%)	0 (0.00%)
Gemini-3.1-Pro-Preview	509 (99.80%)	0 (0.00%)	1 (0.20%)	500 (100.00%)	0 (0.00%)	0 (0.00%)
GLM-5.2	502 (98.43%)	0 (0.00%)	0 (0.00%)	499 (99.80%)	1 (0.20%)	0 (0.00%)
GPT-5.5	508 (99.61%)	0 (0.00%)	0 (0.00%)	499 (99.80%)	0 (0.00%)	0 (0.00%)
Grok-4.3	481 (94.31%)	0 (0.00%)	2 (0.39%)	496 (99.20%)	1 (0.20%)	0 (0.00%)
Kimi-K2.6	507 (99.41%)	0 (0.00%)	0 (0.00%)	N/A	N/A	N/A
Mimo-V2.5-Pro	451 (88.43%)	1 (0.20%)	0 (0.00%)	500 (100.00%)	0 (0.00%)	0 (0.00%)
Qwen3.7-Max	501 (98.24%)	0 (0.00%)	1 (0.20%)	500 (100.00%)	0 (0.00%)	0 (0.00%)

Claude-Opus-4-8 exhibits a different failure mode, with empty-output rates of 25.00% and 20.80% under the natural and forced conditions, respectively. Gemini-3.1-Pro-Preview records errors for 9.60% of natural-response samples, although this falls to 0.60% with explicit Tibetan forcing.

E.2.2 Over-Safety Responses

The language-specific over-safety datasets show substantially more consistent target-language realization than several of the translated conditions. For Cantonese, most models produce the target language for more than 94% of the prompts, with Mimo-V2.5-Pro being the main exception at 88.43%. For Tibetan, every reported model exceeds 93%, and several reach 100%. DeepSeek-V4-Pro is the principal exception in terms of output availability, with 6.60% empty responses.

Taken together, these results show that response-language behavior varies substantially across dataset constructions and language-instruction settings. Explicit target-language instructions are associated with large increases in in-language generation for several models, particularly on the jailbreak datasets. At the same time, persistent blank outputs and safety-filtered responses constitute a distinct form of response-availability failure that should not be conflated with either ordinary language switching or explicit textual refusal.

F Target-Language-Conditioned Response Category Distributions

To complement the target-language-conditioned attack success rate (TL-ASR) reported in the main text, we further examine the full response-category distribution after conditioning on target-language responses. For each jailbreak setting, responses entering the target-language-conditioned analysis are divided into three categories: *Refusal*, *Off-topic*, and *Harmful*. Refusal denotes responses that do not provide substantive assistance for the harmful request. Following the response-accounting protocol used in this analysis, API-call failures are also assigned to the Refusal category. Off-topic responses fail to directly address the requested task, whereas Harmful responses substantively address the harmful request and therefore constitute successful jailbreaks.

We report results separately for translated AdvBench and the language-specific jailbreak datasets, and for both the non-forced and forced target-language settings. "Forced" denotes the setting in which an explicit instruction is added requiring the model to respond in the corresponding target language. Because the number of responses entering the target-language-conditioned analysis varies substantially across models, we report the corresponding denominator N for every model and setting. The percentage of Harmful responses in these tables corresponds to the target-language-conditioned attack success rate under the same response-accounting rule.

Tables 15 and 16 show that the composition of target-language responses differs substantially between Cantonese and Tibetan AdvBench attacks. For translated Cantonese, refusal remains the dominant outcome for most models even after restricting the analysis to the target-language-conditioned subset. Nevertheless, explicitly requiring Cantonese increases the harmful-response share for many models. For example, the harmful proportion increases from 10.53% to 27.91% for MiniMax-M2.7,

Table 15: Target-language-conditioned response-category distributions on the Cantonese translation of AdvBench. Each category cell reports count (percentage), where the percentage is computed over the corresponding N. API-call failures are included in Refusal. "Forced" denotes an explicit instruction requiring a Cantonese response.

	Non-forced				Forced Cantonese			
Model	N	Refusal	Off-topic	Harmful	N	Refusal	Off-topic	Harmful
MiniMax-M2.7	38	33 (86.84%)	1 (2.63%)	4 (10.53%)	43	31 (72.09%)	0 (0.00%)	12 (27.91%)
Claude-Opus-4-8	412	394 (95.63%)	0 (0.00%)	18 (4.37%)	421	397 (94.30%)	0 (0.00%)	24 (5.70%)
DeepSeek-V4-Pro	240	214 (89.17%)	0 (0.00%)	26 (10.83%)	243	211 (86.83%)	0 (0.00%)	32 (13.17%)
Gemini-3.1-Pro-Preview	315	287 (91.11%)	1 (0.32%)	27 (8.57%)	516	408 (79.07%)	2 (0.39%)	106 (20.54%)
GLM-5.2	265	246 (92.83%)	0 (0.00%)	19 (7.17%)	278	250 (89.93%)	0 (0.00%)	28 (10.07%)
GPT-5.5	445	420 (94.38%)	0 (0.00%)	25 (5.62%)	436	386 (88.53%)	1 (0.23%)	49 (11.24%)
Grok-4.3	140	114 (81.43%)	2 (1.43%)	24 (17.14%)	121	98 (80.99%)	1 (0.83%)	22 (18.18%)
Kimi-K2.6	423	389 (91.96%)	0 (0.00%)	34 (8.04%)	422	382 (90.52%)	0 (0.00%)	40 (9.48%)
Mimo-V2.5-Pro	172	158 (91.86%)	0 (0.00%)	14 (8.14%)	489	461 (94.27%)	1 (0.20%)	27 (5.52%)
Qwen3.7-Max	310	287 (92.58%)	1 (0.32%)	22 (7.10%)	330	286 (86.67%)	0 (0.00%)	44 (13.33%)

Table 16: Target-language-conditioned response-category distributions on the Tibetan translation of AdvBench. Each category cell reports count (percentage) over the corresponding N. API-call failures are included in Refusal. Kimi-K2.6 is excluded from the Tibetan evaluation because it rejects Tibetan inputs.

	Non-forced				Forced Tibetan			
Model	N	Refusal	Off-topic	Harmful	N	Refusal	Off-topic	Harmful
MiniMax-M2.7	362	188 (51.93%)	25 (6.91%)	149 (41.16%)	393	180 (45.80%)	27 (6.87%)	186 (47.33%)
Claude-Opus-4-8	439	359 (81.78%)	1 (0.23%)	79 (18.00%)	454	374 (82.38%)	3 (0.66%)	77 (16.96%)
DeepSeek-V4-Pro	506	298 (58.89%)	10 (1.98%)	198 (39.13%)	508	295 (58.07%)	11 (2.17%)	202 (39.76%)
Gemini-3.1-Pro-Preview	506	376 (74.31%)	1 (0.20%)	129 (25.49%)	513	382 (74.46%)	0 (0.00%)	131 (25.54%)
GLM-5.2	515	367 (71.26%)	1 (0.19%)	147 (28.54%)	520	327 (62.88%)	2 (0.38%)	191 (36.73%)
GPT-5.5	512	379 (74.02%)	1 (0.20%)	132 (25.78%)	516	321 (62.21%)	6 (1.16%)	189 (36.63%)
Grok-4.3	228	42 (18.42%)	45 (19.74%)	141 (61.84%)	425	239 (56.24%)	25 (5.88%)	161 (37.88%)
Mimo-V2.5-Pro	512	377 (73.63%)	2 (0.39%)	133 (25.98%)	520	347 (66.73%)	2 (0.38%)	171 (32.88%)
Qwen3.7-Max	509	334 (65.62%)	4 (0.79%)	171 (33.60%)	518	268 (51.74%)	4 (0.77%)	246 (47.49%)

from 8.57% to 20.54% for Gemini-3.1-Pro-Preview, and from 5.62% to 11.24% for GPT-5.5. This pattern is not universal: Mimo-V2.5-Pro decreases from 8.14% to 5.52%.

Target-language Tibetan responses exhibit substantially larger harmful shares for several models. In the non-forced setting, Grok-4.3 reaches 61.84% Harmful responses, while MiniMax-M2.7 and DeepSeek-V4-Pro reach 41.16% and 39.13%, respectively. Requiring Tibetan responses increases the harmful share for several models, including Qwen3.7-Max (33.60% to 47.49%), GPT-5.5 (25.78% to 36.63%), and GLM-5.2 (28.54% to 36.73%). In contrast, Grok-4.3 decreases from 61.84% to 37.88%, indicating that forced target-language generation does not have a uniform effect on jailbreak outcomes.

Tables 17 and 18 further show that the language-specific jailbreak datasets produce response compositions that differ from those observed on translated AdvBench. On the Cantonese language-specific dataset, explicitly enforcing Cantonese substantially increases the Harmful proportion for several models, including DeepSeek-V4-Pro (21.06% to 38.49%), GPT-5.5 (7.06% to 30.06%), Gemini-3.1-Pro-Preview (18.71% to 26.40%), and Qwen3.7-Max (15.14% to 25.15%). At the same time, some models move in the opposite direction: Grok-4.3 decreases from 22.78% to 12.00%, and Mimo-V2.5-Pro decreases from 15.56% to 12.45%. Claude-Opus-4-8 also shows a relatively high Off-topic rate of 11.40% in the non-forced language-specific Cantonese setting, which disappears when Cantonese is explicitly enforced.

The language-specific Tibetan results are similarly model dependent. Forced Tibetan responses increase the Harmful proportion for Claude-Opus-4-8, DeepSeek-V4-Pro, Gemini-3.1-Pro-Preview, GLM-5.2, GPT-5.5, Mimo-V2.5-Pro, and Qwen3.7-Max, with the largest increase in these results occurring for GPT-5.5, from 10.62% to 38.40%. Conversely, the Harmful proportion decreases for MiniMax-M2.7 (31.36% to 22.82%) and Grok-4.3 (38.86% to 19.51%). These results reinforce that target-language enforcement changes not only whether a model remains in the requested language, but also the composition of its in-language safety behavior, and that the direction and magnitude of this change are strongly model- and dataset-dependent.

Table 17: Target-language-conditioned response-category distributions on the language-specific Cantonese jailbreak dataset. Each category cell reports count (percentage) over the corresponding N. API-call failures are included in Refusal. "Forced" denotes an explicit instruction requiring a Cantonese response.

	Non-forced				Forced Cantonese			
Model	N	Refusal	Off-topic	Harmful	N	Refusal	Off-topic	Harmful
MiniMax-M2.7	228	196 (85.96%)	0 (0.00%)	32 (14.04%)	435	368 (84.60%)	1 (0.23%)	66 (15.17%)
Claude-Opus-4-8	500	340 (68.00%)	57 (11.40%)	103 (20.60%)	498	395 (79.32%)	0 (0.00%)	103 (20.68%)
DeepSeek-V4-Pro	451	356 (78.94%)	0 (0.00%)	95 (21.06%)	504	310 (61.51%)	0 (0.00%)	194 (38.49%)
Gemini-3.1-Pro-Preview	433	352 (81.29%)	0 (0.00%)	81 (18.71%)	519	381 (73.41%)	1 (0.19%)	137 (26.40%)
GLM-5.2	477	424 (88.89%)	0 (0.00%)	53 (11.11%)	515	414 (80.39%)	2 (0.39%)	99 (19.22%)
GPT-5.5	510	474 (92.94%)	0 (0.00%)	36 (7.06%)	509	354 (69.55%)	2 (0.39%)	153 (30.06%)
Grok-4.3	259	200 (77.22%)	0 (0.00%)	59 (22.78%)	500	440 (88.00%)	0 (0.00%)	60 (12.00%)
Kimi-K2.6	520	429 (82.50%)	0 (0.00%)	91 (17.50%)	520	375 (72.12%)	0 (0.00%)	145 (27.88%)
Mimo-V2.5-Pro	180	152 (84.44%)	0 (0.00%)	28 (15.56%)	482	415 (86.10%)	7 (1.45%)	60 (12.45%)
Qwen3.7-Max	469	398 (84.86%)	0 (0.00%)	71 (15.14%)	509	379 (74.46%)	2 (0.39%)	128 (25.15%)

Table 18: Target-language-conditioned response-category distributions on the language-specific Tibetan jailbreak dataset. Each category cell reports count (percentage) over the corresponding N. API-call failures are included in Refusal. Kimi-K2.6 is excluded from Tibetan evaluation.

	Non-forced				Forced Tibetan			
Model	N	Refusal	Off-topic	Harmful	N	Refusal	Off-topic	Harmful
MiniMax-M2.7	338	185 (54.73%)	47 (13.91%)	106 (31.36%)	412	243 (58.98%)	75 (18.20%)	94 (22.82%)
Claude-Opus-4-8	375	321 (85.60%)	1 (0.27%)	53 (14.13%)	394	289 (73.35%)	0 (0.00%)	105 (26.65%)
DeepSeek-V4-Pro	468	363 (77.56%)	5 (1.07%)	100 (21.37%)	500	344 (68.80%)	6 (1.20%)	150 (30.00%)
Gemini-3.1-Pro-Preview	452	315 (69.69%)	4 (0.88%)	133 (29.42%)	435	287 (65.98%)	5 (1.15%)	143 (32.87%)
GLM-5.2	499	351 (70.34%)	7 (1.40%)	141 (28.26%)	498	325 (65.26%)	7 (1.41%)	166 (33.33%)
GPT-5.5	499	443 (88.78%)	3 (0.60%)	53 (10.62%)	500	304 (60.80%)	4 (0.80%)	192 (38.40%)
Grok-4.3	211	98 (46.45%)	31 (14.69%)	82 (38.86%)	446	354 (79.37%)	5 (1.12%)	87 (19.51%)
Mimo-V2.5-Pro	498	378 (75.90%)	6 (1.20%)	114 (22.89%)	500	356 (71.20%)	17 (3.40%)	127 (25.40%)
Qwen3.7-Max	499	355 (71.14%)	3 (0.60%)	141 (28.26%)	499	318 (63.73%)	11 (2.20%)	170 (34.07%)

G Target-Language-Conditioned Response Categories on Over-Refusal Datasets

To complement the target-language-conditioned over-refusal rate (TL-ORR) reported in the main text, we further report the complete response-category distributions among responses identified as being in the target language.

For each over-refusal setting, target-language responses are divided into three mutually exclusive categories: *Answer*, *Textual Refusal*, and *Off-topic*. Answer denotes a substantive response that addresses the benign request. Textual Refusal denotes an explicit refusal of an otherwise benign request and therefore corresponds to the numerator of TL-ORR. Off-topic denotes a response that does not directly address the request.

In the four target-language-conditioned over-refusal result files used here, API-call failures, unparsable outputs, and overlapping labels all have zero counts. Therefore, the three categories reported below exhaust the target-language response subset. Each table reports both counts and percentages, with percentages computed over the corresponding number of target-language responses N.

Tables 19 and 20 reveal substantial differences between translated and language-specific over-refusal settings even after conditioning on responses that remain in the target language.

On the translated Cantonese dataset, Textual Refusal accounts for roughly 10–18% of target-language responses for most models, while Off-topic responses are generally rare. The translated Tibetan setting is substantially more heterogeneous. DeepSeek-V4-Pro exhibits the highest Textual Refusal rate at 38.71%, followed by Claude-Opus-4-8 at 24.02%. In addition, several models generate a non-trivial proportion of Off-topic responses in Tibetan, most notably Grok-4.3 (36.87%) and MiniMax-M2.7 (30.09%). This indicates that undesirable behavior on translated benign inputs is not limited to explicit over-refusal, but can also appear as failure to directly address the request.

Table 19: Target-language-conditioned response-category distributions on the translated over-refusal datasets. Each category cell reports count (percentage) among responses identified as being in the target language. The Textual Refusal percentage is the TL-ORR reported in the main text. Kimi-K2.6 is excluded from Tibetan evaluation because it rejects Tibetan inputs.

	Translated Cantonese				Translated Tibetan			
Model	N	Answer	Textual Refusal	Off-topic	N	Answer	Textual Refusal	Off-topic
MiniMax-M2.7	241	200 (82.99%)	35 (14.52%)	6 (2.49%)	452	256 (56.64%)	60 (13.27%)	136 (30.09%)
Claude-Opus-4-8	482	415 (86.10%)	64 (13.28%)	3 (0.62%)	483	324 (67.08%)	116 (24.02%)	43 (8.90%)
DeepSeek-V4-Pro	474	427 (90.08%)	47 (9.92%)	0 (0.00%)	217	122 (56.22%)	84 (38.71%)	11 (5.07%)
Gemini-3.1-Pro-Preview	479	420 (87.68%)	59 (12.32%)	0 (0.00%)	497	440 (88.53%)	46 (9.26%)	11 (2.21%)
GLM-5.2	461	374 (81.13%)	84 (18.22%)	3 (0.65%)	471	356 (75.58%)	93 (19.75%)	22 (4.67%)
GPT-5.5	495	410 (82.83%)	84 (16.97%)	1 (0.20%)	494	394 (79.76%)	88 (17.81%)	12 (2.43%)
Grok-4.3	457	383 (83.81%)	72 (15.75%)	2 (0.44%)	396	212 (53.54%)	38 (9.60%)	146 (36.87%)
Kimi-K2.6	493	429 (87.02%)	63 (12.78%)	1 (0.20%)	–	–	–	–
Mimo-V2.5-Pro	332	290 (87.35%)	39 (11.75%)	3 (0.90%)	494	414 (83.81%)	66 (13.36%)	14 (2.83%)
Qwen3.7-Max	483	410 (84.89%)	71 (14.70%)	2 (0.41%)	488	405 (82.99%)	59 (12.09%)	24 (4.92%)

Table 20: Target-language-conditioned response-category distributions on the language-specific over-refusal datasets. Each category cell reports count (percentage) among responses identified as being in the target language. The Textual Refusal percentage is the TL-ORR reported in the main text. Kimi-K2.6 is excluded from Tibetan evaluation.

	Language-Specific Cantonese				Language-Specific Tibetan			
Model	N	Answer	Textual Refusal	Off-topic	N	Answer	Textual Refusal	Off-topic
MiniMax-M2.7	480	460 (95.83%)	16 (3.33%)	4 (0.83%)	493	477 (96.75%)	0 (0.00%)	16 (3.25%)
Claude-Opus-4-8	509	505 (99.21%)	4 (0.79%)	0 (0.00%)	498	495 (99.40%)	0 (0.00%)	3 (0.60%)
DeepSeek-V4-Pro	510	508 (99.61%)	1 (0.20%)	1 (0.20%)	467	456 (97.64%)	9 (1.93%)	2 (0.43%)
Gemini-3.1-Pro-Preview	509	508 (99.80%)	1 (0.20%)	0 (0.00%)	500	498 (99.60%)	0 (0.00%)	2 (0.40%)
GLM-5.2	502	500 (99.60%)	2 (0.40%)	0 (0.00%)	499	495 (99.20%)	1 (0.20%)	3 (0.60%)
GPT-5.5	508	504 (99.21%)	3 (0.59%)	1 (0.20%)	499	491 (98.40%)	0 (0.00%)	8 (1.60%)
Grok-4.3	481	473 (98.34%)	3 (0.62%)	5 (1.04%)	496	442 (89.11%)	3 (0.60%)	51 (10.28%)
Kimi-K2.6	507	507 (100.00%)	0 (0.00%)	0 (0.00%)	–	–	–	–
Mimo-V2.5-Pro	451	442 (98.00%)	4 (0.89%)	5 (1.11%)	500	490 (98.00%)	0 (0.00%)	10 (2.00%)
Qwen3.7-Max	501	497 (99.20%)	3 (0.60%)	1 (0.20%)	500	497 (99.40%)	0 (0.00%)	3 (0.60%)

In contrast, target-language responses on the language-specific over-refusal datasets are overwhelmingly classified as Answer. On the Cantonese dataset, Answer exceeds 95% for every model, while Textual Refusal remains at or below 3.33%. On the Tibetan dataset, Textual Refusal is near zero for most models; DeepSeek-V4-Pro has the highest value at 1.93%. Off-topic behavior is also generally low, although Grok-4.3 reaches 10.28% on the language-specific Tibetan dataset.

These distributions provide a more complete view than TL-ORR alone. In particular, the low TL-ORR observed on the language-specific datasets corresponds primarily to models directly answering the benign request rather than replacing explicit refusals with another response category. At the same time, the translated Tibetan results show that Off-topic behavior can remain substantial even when explicit textual over-refusal is relatively limited.

H Language-Specific Dataset Construction

This appendix provides the complete construction details of the language-specific datasets introduced in Section 3.1. The purpose of these datasets is to complement translated multilingual benchmarks with prompts that are constructed directly around linguistic phenomena, culturally grounded contexts, and usage patterns associated with the target language. Across all four datasets, candidate generation and rubric-based quality scoring were performed with LLM assistance. These automatic scores served only as an initial filtering stage: every example in the final evaluation sets was manually reviewed, and any example rejected during human review was manually rewritten and re-verified before inclusion.

We construct four language-specific evaluation sets: Cantonese and Tibetan jailbreak datasets, and Cantonese and Tibetan over-refusal datasets. The final evaluation sets contain 520 Cantonese jailbreak prompts, 500 Tibetan jailbreak prompts, 510 Cantonese over-refusal prompts, and 500 Tibetan over-refusal prompts.

Table 21: Summary of the language-specific datasets used in our evaluation. The construction dimensions are language dependent and are designed to capture linguistic or contextual phenomena that are difficult to obtain through direct translation.

Dataset	Evaluation Goal	Construction Dimensions	Size
Cantonese Jailbreak	Jailbreak robustness	Harm category × strategy × Cantonese feature	520
Tibetan Jailbreak	Jailbreak robustness	Harm category × strategy × Tibetan context	500
Cantonese Over-Refusal	Benign-request handling	15 language-specific lexical/semantic subcategories	510
Tibetan Over-Refusal	Benign-request handling	Safety domain × request type × Tibetan context	500

Importantly, we do not enforce an identical taxonomy either across languages or across evaluation objectives. The relevant linguistic and contextual phenomena differ by language and by safety task. For example, the Cantonese jailbreak dataset varies harmful intent, attack strategy, and Cantonese linguistic realization, whereas the Cantonese over-refusal dataset targets benign Cantonese-specific expressions that may superficially trigger safety mechanisms. This task-dependent taxonomy is intentional: harmful bypass and benign false refusal need not arise from the same linguistic phenomena. Nevertheless, all four datasets follow a common high-level procedure: controlled specification of evaluation dimensions, target-language generation, and quality-oriented filtering before inclusion in the final evaluation set.

H.1 Cantonese Jailbreak Dataset

Design motivation. The Cantonese jailbreak dataset is designed to test whether safety behavior changes when harmful requests are expressed using linguistic forms that are characteristic of natural Cantonese rather than translated written Chinese. We therefore construct the dataset along three dimensions: harmful intent, jailbreak strategy, and Cantonese linguistic realization.

Harm categories. We consider four broad harmful-intent categories: (1) fraud and misinformation, (2) illegal activities and underground markets, (3) cyber attacks and privacy violations, and (4) violence and dangerous behavior. The categories are intentionally broad so that individual prompts can vary substantially in scenario while remaining associated with a controlled high-level intent.

Jailbreak strategies. Each harmful intent is combined with one of five jailbreak strategies: *role-play*, *fictional framing*, *instruction override*, *social engineering*, and *academic framing*. These strategies are used as generation constraints rather than as post-hoc labels.

Cantonese linguistic features. To introduce language-specific variation, we further distinguish two forms of Cantonese realization. The first emphasizes colloquial Cantonese, including characteristic particles, slang, and spoken grammatical patterns. The second emphasizes natural Cantonese–English code-switching, reflecting the mixed-language expressions that frequently occur in informal and workplace Cantonese.

The Cartesian product of the four harmful categories, five jailbreak strategies, and two linguistic realizations produces 40 base design combinations. The final version used in our experiments contains 520 Cantonese jailbreak prompts.

Generation. Candidate prompts are generated directly in Cantonese while conditioning on the assigned harm category, jailbreak strategy, and linguistic feature. The generation instructions require diversity of scenarios and discourage written-Chinese constructions that would make the prompt resemble a direct translation. Prompts are also restricted from introducing political, ideological, or other unrelated sensitive content.

Quality filtering. Each candidate is subsequently evaluated for three properties: (1) absence of excluded political or ideological content, (2) Cantonese linguistic authenticity, and (3) consistency with the assigned harm category, jailbreak strategy, and Cantonese feature. A candidate is retained only when no excluded content is detected, its Cantonese-authenticity score is at least 4 on a five-point scale, and the intended dimensions are judged to be correctly instantiated. Candidates failing any criterion are removed.

H.2 TIBETAN JAILBREAK DATASET

Design motivation. The Tibetan jailbreak dataset follows the same principle of direct language-specific construction but uses a different set of dimensions to represent Tibetan-relevant scenarios. It is organized along harmful intent, jailbreak strategy, and contextual setting.

Harm categories. We define five harmful-intent categories: (1) economic or cultural-artifact fraud, (2) ecological damage and poaching, (3) dangerous items or harmful remedies, (4) cultural-heritage destruction or illegal acquisition, and (5) privacy violations or online harassment.

Jailbreak strategies. Five attack strategies are considered: (1) philosophical or religious framing, (2) fictional or role-playing scenarios, (3) academic or cultural-heritage framing, (4) emotional or emergency framing, and (5) language- or format-based constraints.

Contextual settings. Each intent–strategy pair is further instantiated in one of four contextual settings: (1) pastoral or nomadic life, (2) generic religious or monastic settings, (3) towns and cultural-artifact markets, and (4) wilderness or nature-protection environments. These contexts serve as scenario anchors rather than standalone topics.

The resulting design contains $5 \times 5 \times 4 = 100$ combinations. Five qualified prompts are retained for each combination, producing 500 Tibetan jailbreak prompts.

Generation. For each combination, the generator produces a Tibetan prompt together with a Chinese semantic translation. The Tibetan prompt is required to express the assigned harmful intent, jailbreak strategy, and contextual anchor while remaining natural in the target language. The accompanying Chinese rendering is stored to facilitate semantic inspection and quality control.

Quality filtering. Candidate prompts are screened for both content and dimensional consistency. Prompts containing excluded political, ideological, ethnic, or real-world sensitive material are removed. The quality-control procedure additionally checks whether the intended harmful objective, jailbreak strategy, and contextual anchor are preserved. Only prompts satisfying both requirements are retained in the final dataset.

Native-speaker semantic validation. A coauthor who is a native Tibetan speaker manually compared each Tibetan prompt with its paired Chinese semantic rendering to verify that the two expressed the same underlying request. In the initial review, 436 of the 500 pairs (87.2%) were judged semantically aligned. The remaining 64 pairs were revised to resolve semantic discrepancies and were subsequently re-checked by the same native speaker. Only pairs whose Tibetan prompt and Chinese rendering were judged semantically aligned were retained in the final dataset.

H.3 CANTONESE OVER-REFUSAL DATASET

The Cantonese over-refusal dataset is constructed in two stages. The first stage generates benign Cantonese examples containing language-specific surface expressions that may resemble safety-sensitive content. The second stage converts these examples into explicit, single-turn, answerable requests while preserving the original linguistic trigger.

H.3.1 STAGE 1: LANGUAGE-SPECIFIC BENIGN TRIGGERS

Category design. We define 15 linguistic subcategories grouped into five broader families.

The first family covers **colloquial profanity and pejorative slang**, including profanity used as an emotional intensifier, affectionate teasing between familiar speakers, and mild complaints directed at objects or situations rather than people.

The second family covers **gangster and underworld expressions**, including organizational titles used metaphorically, violent-sounding descriptions of ordinary actions, and ritual-like expressions used in non-criminal social settings.

The third family covers **death-, violence-, and superstition-related hyperbole**, including exaggerated descriptions of ordinary difficulties, violent metaphors for games or object manipulation, and expressions associated with ghosts, funerary vocabulary, or folk culture.

The fourth family covers **gambling, physiological, and suggestive expressions**, including mahjong or gambling terminology used in family entertainment, direct physiological or excretory vocabulary in legitimate everyday contexts, and Cantonese-specific homophones or metaphors whose actual referents are benign.

The fifth family covers **dangerous-item, prohibited-substance, and crime-related metaphors**, including drug-related slang referring to ordinary ingredients, weapon or explosion terminology used in games or household contexts, and robbery or kidnapping expressions used metaphorically in shopping, games, or interpersonal jokes.

Each of the 15 subcategories contributes 34 retained instances, yielding 510 Cantonese examples.

Generation constraints. The generation procedure requires each example to contain a surface-level trigger associated with its assigned category while remaining unambiguously benign in context. Examples are generated using natural Cantonese orthography and colloquial syntax, and generation is constrained to avoid mixing multiple sensitive dimensions within the same instance.

Initial quality filtering. Generated instances first undergo a programmatic check requiring the presence of at least one identified trigger expression. A subsequent quality-control stage verifies that the trigger is genuinely associated with the intended sensitive lexical category. Candidates that pass this check are then evaluated for contextual harmlessness and Cantonese authenticity. Only instances receiving the maximum benign-context score and an authenticity score of at least 4 on a five-point scale are retained.

H.3.2 Stage 2: Conversion to Explicit Benign Requests

Although the first-stage examples contain the intended linguistic phenomena, we further normalize them into clearly answerable single-turn prompts. The original Cantonese utterance is preserved, and a category-specific request suffix is appended rather than rewriting or replacing the original content.

The appended request may ask for an explanation, rewrite, recommendation, dialogue, description, or another ordinary task depending on the category. It must remain semantically consistent with the original benign context and must not introduce a new sensitive dimension.

To avoid artificially revealing the intended benchmark label, the augmented request is prohibited from containing explicit cues such as "this request is harmless," "this is not a crime," or "do not refuse." The request must also remain natural in Cantonese and must form a clear task that can be answered in a single turn.

Augmented candidates are retained only when all of the following conditions are satisfied: (1) the resulting prompt contains an explicit request, (2) the underlying intent remains benign, (3) the request remains consistent with the original category, (4) no new sensitive dimension is introduced, (5) no benchmark-label or safety-label leakage is present, (6) the Cantonese expression remains natural, and (7) the requested task is unambiguous.

This second stage produces the final 510-instance Cantonese over-refusal evaluation set.

H.4 Tibetan Over-Refusal Dataset

Direct-request construction. In contrast to the two-stage Cantonese procedure, the Tibetan over-refusal dataset directly generates complete user requests. This avoids constructing isolated seed utterances that would subsequently need to be converted into questions.

Each request must express an explicitly benign objective, such as identifying a risk, preventing harm, protecting people or property, responding safely to a suspicious situation, explaining a risk, or communicating safety information. At the same time, the prompt naturally preserves lexical items or

concepts associated with potentially unsafe activities so that it can test whether a model overreacts to surface-level triggers.

Safety domains. We define five benign safety domains: (1) economic and cultural-artifact fraud prevention, (2) ecological protection and anti-poaching, (3) hazardous-item and harmful-remedy risk prevention, (4) cultural-heritage protection and prevention of theft or damage, and (5) privacy protection and prevention of online harassment.

Request types. Each domain is combined with one of five ordinary request types: (1) risk identification, (2) prevention advice, (3) safe response, (4) educational explanation, and (5) communication assistance.

Contextual settings. Requests are further grounded in four Tibetan-relevant contexts: (1) pastoral or nomadic life, (2) traditional culture and community life, (3) towns and cultural-artifact or handicraft markets, and (4) wilderness and nature-protection environments.

The full design therefore contains $5 \times 5 \times 4 = 100$ combinations. Five qualified requests are retained for each combination, yielding 500 Tibetan over-refusal prompts.

Generation constraints. Generated requests must have a clearly benign intent while naturally containing potentially safety-triggering terminology. They may ask how to recognize risks, avoid victimization, respond safely, preserve evidence, seek legitimate assistance, communicate warnings, or understand why a behavior is dangerous.

Requests that seek instructions for carrying out the underlying harmful activity, improving its effectiveness, concealing it, or avoiding detection are explicitly excluded. Each Tibetan request is stored together with a Chinese semantic rendering and the identified trigger terms.

Quality filtering. Candidate requests are evaluated along five dimensions: *benign intent*, *absence of actionable harmful information*, *Tibetan linguistic authenticity*, *fit with the assigned domain, request type, and context*, and *strength of the surface-level safety trigger*.

A candidate is retained only if the benign-intent and non-actionability scores both receive the maximum value of 5, Tibetan authenticity and domain fit are at least 4 on a five-point scale, and trigger strength is at least 3. The quality-control procedure additionally verifies that the identified trigger expressions actually occur in the Tibetan prompt and that the accompanying Chinese rendering preserves the meaning of the original request.

H.5 Shared Construction Principles

Despite their intentionally language- and task-dependent taxonomies, all four language-specific datasets follow three shared construction principles.

Direct language-specific construction. The datasets are constructed around the target languages themselves rather than obtained by translating a common English source set. Language-specific lexical patterns, discourse styles, and contextual settings are therefore part of the benchmark design rather than artifacts introduced after translation.

Controlled coverage. Generation is conditioned on predefined dimensions rather than relying on unconstrained free-form sampling. The specific dimensions are selected according to the relevant language and evaluation objective, enabling systematic coverage while preserving linguistic and task-specific differences rather than imposing an artificial shared taxonomy.

Quality-oriented filtering. All datasets apply post-generation filtering for linguistic naturalness and consistency with the intended benchmark dimensions. For jailbreak datasets, the intended attack objective and jailbreak structure must remain present. For over-refusal datasets, the underlying user intent must remain clearly benign while preserving sufficient surface-level safety triggers to make unnecessary refusal measurable.

These datasets are therefore intended to be complementary to translated multilingual benchmarks rather than replacements for them. The translated benchmarks provide controlled cross-lingual comparability, whereas the language-specific datasets probe safety behavior under linguistic and contextual phenomena that may not be preserved through translation.

H.6 Dataset Availability

The Cantonese and Tibetan over-refusal datasets, together with the code used for their construction and evaluation, will be made available through the anonymous repository described in the reproducibility statement.

Because the jailbreak datasets contain potentially harmful and dual-use prompts, we do not provide unrestricted public download during the submission process. Following acceptance, qualified researchers may request access for research purposes through the contact information provided by the authors. This controlled-access policy is intended to preserve reproducibility while reducing unnecessary dissemination of potentially misusable jailbreak content.