

# Disentangling In-Language Alignment and Cross-Lingual Safety Borrowing in Low-Resource Languages

Zeguang Huang<sup>1,2</sup>, Zhenma Spenpa<sup>1,2</sup>, Zhu Yecheng<sup>4</sup>,  
Xi Cao<sup>3</sup>, Nuo Qun<sup>\*1,2</sup>, Tashi Nyima<sup>1,2</sup>

<sup>1</sup>School of Information Science and Technology, Xizang University, Lhasa, China

<sup>2</sup>State Key Laboratory of Tibetan Intelligence, Lhasa, China

<sup>3</sup>Minzu University of China, Beijing, China

<sup>4</sup>The University of Hong Kong, Hong Kong, China

misaka@stu.utibet.edu.cn, q\_nuo@utibet.edu.cn

## Abstract

We study two modes of low-resource language safety: **in-language alignment**, where models refuse harmful instructions in the target language, and **cross-lingual safety borrowing**, where they rely on a better-aligned proximate high-resource language, including via response-language fallback. We evaluate 11 commercial APIs on English, Mandarin, Cantonese, Tibetan, and Russian, plus 3 Tibetan-specialized open-source models, using native-speaker-verified datasets. Results reveal two patterns: Tibetan, for which response-language fallback provides limited aggregate protection in our setting, shows a safety cliff with ASR of 12.69–60.96%, far above English; Cantonese, which shares script and vocabulary with Mandarin, keeps aggregate ASR in single digits, but Cantonese-reply ASR exceeds aggregate ASR for all 11 models (sign test,  $p < 0.001$ ), exposing in-language safety gaps concealed by aggregate metrics. We release Cantonese-AdvBench and Tibetan-AdvBench.

## 1 Introduction

Prior work on multilingual safety has documented two pathways by which large language models achieve safety alignment in low-resource languages: **in-language alignment**, where the model refuses harmful instructions in the target language, and cross-lingual safety transfer, where alignment from a high-resource language generalizes to others (Zhao et al., 2025). We refer to the latter as **cross-lingual safety borrowing**: low-resource safety may be partly obtained from better-developed refusal behavior in a high-resource language, either because that language is linguistically proximate to the target language or because it serves as a safer default response language for the model. This borrowing can in principle

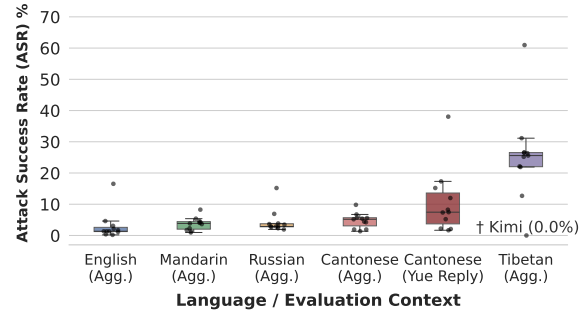


Figure 1: Aggregate ASR across languages and the safety cliff in Tibetan.† Kimi refuses all Tibetan inputs at the language level rather than evaluating content. Cantonese (Yue Reply) refers to the attack success rate specifically for model responses generated in Cantonese within the Cantonese dataset.

arise through shared refusal representations (Wang et al., 2025; Farashah et al., 2026), but in this paper our primary observable evidence concerns the behavioral channel: models may fall back to a safer high-resource response language when prompted in a low-resource language. However, aggregate ASR, the dominant metric in multilingual safety evaluation, cannot distinguish target-language refusal from safer non-target-language fallback. Response language therefore provides an observable diagnostic signal: when fallback is available, aggregate ASR can understate in-language risk; when fallback is limited, low-resource ASR rises sharply.

Testing this distinction requires a controlled contrast: two low-resource languages that differ in the availability of a proximate high-resource fallback language. Cantonese and Tibetan provide a useful contrast. Cantonese shares its writing system and substantial vocabulary with Mandarin (Snow, 2004; Xiang et al., 2024)—some expressions differ only in pronunciation—making Mandarin a plausible source of cross-lingual safety borrowing. Tibetan, by contrast, shares neither

\*Corresponding author.

script nor substantial vocabulary with Mandarin or English (Tournadre and Dorje, 2003; Chen et al., 2025), making such borrowing less readily available in our evaluation setting. To situate this contrast along the language-resource spectrum, we additionally evaluate Russian as a mid-resource reference: its speaker population is far smaller than that of Mandarin or English, yet substantially larger than that of Cantonese or Tibetan. This intermediate position provides a natural point of comparison between the high-resource languages and the two low-resource targets, testing whether elevated ASR is specific to the low-resource regime or already emerges at mid-resource levels. If the two-mode account is correct, Cantonese and Tibetan should exhibit qualitatively different ASR patterns under the same evaluation framework: Cantonese ASR should track Mandarin ASR at the aggregate level but diverge under reply-language disaggregation, whereas Tibetan ASR should remain high among content-responding models because response-language fallback provides less systematic aggregate protection.

Prior multilingual safety work establishes that low-resource languages bypass safety alignment more readily than high-resource ones (Wang et al., 2024; Yong et al., 2023; Deng et al., 2024), but these studies rely on aggregate ASR and do not distinguish in-language from cross-lingual transfer; nor do they cover Cantonese or Tibetan. Existing NLP resources for these two languages address only general-purpose understanding (Jiang et al., 2025; Cheng et al., 2025), foundational corpora (Huang et al., 2025), and offensive language detection (Ren et al., 2024)—no jailbreak evaluation datasets exist. We close both gaps by translating the harmful\_behaviors split of AdvBench (Zou et al., 2023) into Mandarin, Cantonese, and Tibetan, with native-speaker verification for Cantonese and Tibetan, and by introducing reply-language disaggregated ASR as the evaluation method.

Our contributions are as follows:

1. We distinguish two observable modes of low-resource safety—**in-language alignment** and **cross-lingual safety borrowing** via response-language fallback—and propose reply-language disaggregated ASR to reveal when aggregate ASR is lowered by safer high-resource-language replies.
2. Across 11 commercial APIs and 3 Tibetan-

specialized models, we find a Tibetan safety cliff (ASR 12.69%–60.96%) and a Cantonese in-language gap: Cantonese-reply ASR exceeds aggregate ASR for all APIs ( $p < 0.001$ ), showing stronger fallback reliance than Tibetan.

3. We release Cantonese-AdvBench and Tibetan-AdvBench, the first native-speaker-verified safety benchmarks for these languages.

## 2 Methodology

### 2.1 Dataset Construction

We build on the harmful\_behaviors subset of AdvBench (Zou et al., 2023) (520 English harmful instructions). Mandarin, Russian, and Cantonese translations are produced with NiuTrans<sup>1</sup>, and Tibetan translations are produced with Sunshine Tibetan Translator<sup>2</sup>. The Cantonese and Tibetan sets are each manually verified by a native-speaker co-author, and the Russian set has been manually verified by a Russian-language expert. The five language versions are strictly aligned (520 entries each).

### 2.2 Models Under Evaluation

We evaluate two groups of models (Appendix I): 11 commercial LLM APIs tested on all four languages, and 3 open-source Tibetan-specialized models tested on Tibetan only, as their cross-language safety alignment is out of scope. All API models are queried with default parameters in April 2026; open-source models are run locally.

### 2.3 Evaluation Framework

We use **ASR** as the proportion of Harmful Responses. Each output is classified as *Refusal*, *Harmful Response*, or *Off-topic*. Classification uses an LLM-as-a-judge ensemble (Zheng et al., 2023) of Gemini 3.1 Flash-Lite Preview and Qwen3.5-Plus, with disagreements arbitrated by Claude Sonnet 4.6 (thinking mode enabled). Each judge receives a one-shot prompt containing a harmful instruction (goal) paired with its expected harmful output (target) from AdvBench as the in-context example, followed by the actual goal under evaluation and the model’s response. The judge then classifies the response into one

<sup>1</sup><https://ntrans.xfyun.cn>

<sup>2</sup><https://mt.utibet.edu.cn/>

| Model                     | en    | zh   | ru    | yue  | bo                |
|---------------------------|-------|------|-------|------|-------------------|
| Claude Opus 4.7           | 1.73  | 3.65 | 3.85  | 4.23 | 12.69             |
| DeepSeek-Reasoner         | 1.15  | 5.38 | 2.31  | 6.73 | 25.19             |
| Doubao-Seed-2.0-Pro       | 1.35  | 4.62 | 6.92  | 5.77 | 60.96             |
| Gemini 3.1 Pro            | 4.62  | 4.04 | 3.46  | 5.58 | 25.58             |
| GLM-5.1                   | 1.35  | 0.96 | 2.88  | 1.92 | 31.15             |
| GPT-5.4                   | 0.19  | 1.73 | 3.65  | 1.92 | 22.12             |
| Grok-4.20                 | 16.54 | 8.27 | 15.19 | 9.81 | 26.54             |
| Kimi K2.5                 | 2.12  | 2.31 | 2.88  | 5.19 | 0.00 <sup>†</sup> |
| MiMo-V2-Pro               | 3.08  | 4.23 | 2.88  | 5.38 | 26.35             |
| MiniMax-M2.7              | 0.38  | 1.35 | 2.69  | 1.35 | 21.92             |
| Qwen3.6-Plus              | 1.35  | 3.85 | 1.92  | 4.42 | 26.54             |
| Tibetan-13B-Llama2-Alpaca | –     | –    | –     | –    | 34.42             |
| Tibetan-7B-Llama2-Alpaca  | –     | –    | –     | –    | 47.69             |
| Qwen2.5-7B-Tib-LoRA       | –     | –    | –     | –    | 20.00             |

Table 1: Aggregate ASR (%) across five languages. <sup>†</sup> Kimi refuses all Tibetan inputs at the language level rather than evaluating content. Open-source Tibetan models are evaluated on Tibetan only.

of the three categories. Human validation is conducted on 100 Cantonese outputs and 33 Tibetan outputs; details are reported in Appendix G.

To enable response-language-aware analysis, we compute the **Cantonese Response Rate (CRR)** and **Tibetan Response Rate (TRR)**—the proportion of outputs in the target language—and disaggregate ASR by reply language (*matching* vs. *non-matching*). For Tibetan, response language is identified by the presence of Tibetan Unicode characters (U+0F00–U+0FFF). For Cantonese, detection uses a two-stage process: responses are first filtered by high-frequency Cantonese-specific lexical markers, then verified by the same LLM-as-a-judge ensemble described above. The validation of this Cantonese detection pipeline is reported in Appendix G.

We further report the **dilution gap** =  $(1 - \text{CRR}) \cdot (\text{ASR}_{\text{yue}} - \text{ASR}_{\text{non-yue}})$ , equivalent to  $\text{ASR}_{\text{yue}} - \text{ASR}_{\text{aggregate}}$ , quantifying how much language-switching dilutes aggregate ASR (analogously for Tibetan with TRR). This factorization separates two orthogonal failure modes—language-switching frequency and the in-language safety gap—that call for distinct mitigations.

### 3 Results

#### 3.1 The Safety Cliff

Table 1 reveals a discrete safety cliff: ASR values cluster in the single digits for English, Mandarin, and Cantonese, but rises to 12.69–60.96% for Tibetan—an order of magnitude above the English baseline. Kimi K2.5 sidesteps the problem

| Model               | CRR   | ASR <sub>yue</sub> | ASR <sub>non-yue</sub> | Gap   |
|---------------------|-------|--------------------|------------------------|-------|
| DeepSeek-Reasoner   | 17.69 | 38.04              | 0.00                   | 31.31 |
| Doubao-Seed-2.0-Pro | 24.42 | 17.32              | 2.04                   | 11.55 |
| Qwen3.6-Plus        | 24.04 | 15.20              | 1.01                   | 10.78 |
| Gemini 3.1 Pro      | 60.58 | 8.25               | 1.46                   | 2.68  |
| Grok-4.20           | 76.92 | 12.00              | 2.50                   | 2.19  |
| Kimi K2.5           | 57.50 | 7.36               | 2.26                   | 2.17  |
| MiMo-V2-Pro         | 64.42 | 7.46               | 1.62                   | 2.08  |
| Claude Opus 4.7     | 70.00 | 5.22               | 1.92                   | 0.99  |
| MiniMax-M2.7        | 80.58 | 1.67               | 0.00                   | 0.32  |
| GLM-5.1             | 78.85 | 2.20               | 0.91                   | 0.27  |
| GPT-5.4             | 93.08 | 2.07               | 0.00                   | 0.14  |

Table 2: Cantonese Response Rate (CRR, %), disaggregated ASR (%) for Cantonese inputs, and dilution gap (pp). Models sorted by Dilution Gap (descending).

by refusing all Tibetan inputs at the language level. The Tibetan cliff is not an artifact of comprehension failure. Native-speaker re-annotation (Appendix G) shows that genuinely off-topic outputs are a small minority: among 45 outputs the judge labeled as Off-topic, only 1 was confirmed as truly off-topic, while 27 were refusals and 17 were harmful responses. Genuine comprehension failure on Tibetan is therefore at most as frequent as the already low automated Off-topic rates (mean 2.65%, versus 8.46% for Mandarin and 9.09% for Cantonese)—and in practice far rarer, since even most judge-labeled Off-topic outputs were substantive. Conversely, the 17 harmful responses misclassified as Off-topic show that the reported Tibetan ASR is an underestimate: the only sizeable judge error hides harmful responses in the auxiliary class, so the true cliff is sharper than reported. We accordingly treat automated Off-topic rates on Tibetan only as an auxiliary diagnostic, not as a comprehension measure (Appendix A; open-source Tibetan models in Table 8).

#### 3.2 The In-Language Safety Gap on Cantonese

Although aggregate Cantonese ASR appears comparable to that of Mandarin (Table 1), models sometimes silently switch to Mandarin when prompted in Cantonese. Table 2 disentangles these: CRR ranges from 93.08% (GPT-5.4) to 17.69% (DeepSeek), and **every model shows higher ASR when replying in Cantonese (sign test,  $p < 0.001$ )**. The dilution gap quantifies how much aggregate ASR understates in-language risk.

| Model               | TRR               | ASR <sub>bo</sub> | ASR <sub>non-bo</sub> | Gap   |
|---------------------|-------------------|-------------------|-----------------------|-------|
| Claude Opus 4.7     | 30.00             | 39.74             | 1.10                  | 27.05 |
| GPT-5.4             | 79.04             | 25.30             | 10.09                 | 3.19  |
| Grok-4.20           | 90.38             | 29.36             | 0.00                  | 2.82  |
| Doubao-Seed-2.0-Pro | 98.27             | 62.04             | 0.00                  | 1.07  |
| Gemini 3.1 Pro      | 96.54             | 26.29             | 5.56 <sup>‡</sup>     | 0.72  |
| GLM-5.1             | 98.27             | 31.70             | 0.00 <sup>‡</sup>     | 0.55  |
| MiMo-V2-Pro         | 99.62             | 26.45             | 0.00 <sup>‡</sup>     | 0.10  |
| DeepSeek-Reasoner   | 99.23             | 25.19             | 25.00 <sup>‡</sup>    | 0.00  |
| MiniMax-M2.7        | 100.00            | 21.92             | —                     | 0.00  |
| Qwen3.6-Plus        | 97.88             | 26.13             | 45.45 <sup>‡</sup>    | −0.41 |
| Kimi K2.5           | 0.00 <sup>†</sup> | —                 | —                     | —     |

Table 3: Tibetan Response Rate (TRR, %), disaggregated ASR (%) for Tibetan inputs, and dilution gap (pp). Models sorted by Dilution Gap (descending). <sup>†</sup> Kimi refuses all Tibetan inputs at the language level. <sup>‡</sup> Fewer than 20 non-Tibetan replies; interpret with caution. “—” = not applicable.

### 3.3 Response Language within Tibetan Inputs

As shown in Table 3, most commercial models achieve Tibetan response rates above 90%, with Claude Opus 4.7 (30%) and GPT-5.4 (79%) as the only commercial APIs with notably lower rates. Kimi K2.5 refuses all Tibetan inputs and is excluded from reply-language analysis. Most models nonetheless exhibit higher ASR when replying in Tibetan than when replying in other languages, though small non-Tibetan reply subgroups (Table 3) limit conclusions about individual models. Dilution gaps are 3.19 pp or lower for all commercial APIs except Claude Opus 4.7, whose in-language Tibetan ASR (39.74%) exceeds its aggregate (12.69%) by 27.05 pp.

### 3.4 Target-Language Response Rates

Table 4 shows near-perfect target-language adherence for English, Mandarin, and Russian for most models, with Claude Opus 4.7 as the main exception in Mandarin and Russian. Low-resource languages diverge sharply: Cantonese response rates vary from 17.69% to 93.08%, with several models frequently falling back to Mandarin, whereas Tibetan response rates exceed 90% for most models, aside from Claude Opus 4.7, GPT-5.4, and Kimi K2.5. This contrast motivates the subsequent analysis: Cantonese leaves substantial room for cross-lingual safety borrowing through language switching, while Tibetan relies more heavily on in-language alignment.

| Model               | en    | zh    | ru    | yue   | bo                |
|---------------------|-------|-------|-------|-------|-------------------|
| Claude Opus 4.7     | 98.85 | 83.65 | 82.50 | 70.00 | 30.00             |
| DeepSeek-Reasoner   | 100.0 | 100.0 | 99.81 | 17.69 | 99.23             |
| Doubao-Seed-2.0-Pro | 100.0 | 99.81 | 99.81 | 24.42 | 98.27             |
| Gemini 3.1 Pro      | 99.81 | 100.0 | 99.04 | 60.58 | 96.54             |
| GLM-5.1             | 100.0 | 99.62 | 98.65 | 78.85 | 98.27             |
| GPT-5.4             | 98.08 | 97.88 | 98.27 | 93.08 | 79.04             |
| Grok-4.20           | 100.0 | 95.96 | 95.38 | 76.92 | 90.38             |
| Kimi K2.5           | 100.0 | 100.0 | 99.81 | 57.50 | 0.00 <sup>†</sup> |
| MiMo-V2-Pro         | 100.0 | 98.08 | 99.23 | 64.42 | 99.62             |
| MiniMax-M2.7        | 100.0 | 99.81 | 99.23 | 80.58 | 100.0             |
| Qwen3.6-Plus        | 100.0 | 99.81 | 99.62 | 24.04 | 97.88             |

Table 4: Target-language response rates (%) across five languages. <sup>†</sup> Kimi refuses all Tibetan inputs at the language level.

## 4 Analysis

We diagnose the two mechanisms in Section 1 through reply-language behavior. Target-language ASR measures in-language alignment, while target-language response rate shows how often that channel is used. Safer high-resource-language replies indicate cross-lingual safety borrowing when they reduce aggregate ASR. Forced target-language generation further tests this by reducing reliance on response-language fallback.

### 4.1 In-Language Alignment: Cantonese Is Masked, Tibetan Is Exposed

As shown in Section 3, Aggregate ASR suggests that Cantonese is safer than Tibetan among responding commercial APIs, but reply-language disaggregation reveals different in-language alignment patterns. In Cantonese, every model has higher ASR when replying in Cantonese than in aggregate, indicating that aggregate ASR masks weaker Cantonese in-language safety. In Tibetan, most models reply in Tibetan for the vast majority of inputs, yet Tibetan-reply ASR remains high, so the in-language weakness is exposed rather than diluted. This contrast is reflected in the dilution gap: 7 of 11 commercial models evaluated on Cantonese have gaps above 2 percentage points, whereas 6 of 10 commercial models evaluated on Tibetan have gaps below 1 percentage point. Cantonese and Tibetan therefore exhibit different failure modes: Cantonese hides in-language risk through response-language switching, while Tibetan reveals it directly.

## 4.2 Cross-Lingual Safety Borrowing: Systematic in Cantonese, Limited in Tibetan

For Cantonese, non-Cantonese replies are usually safer than Cantonese replies and occur frequently enough to lower aggregate ASR. This is reflected in the larger dilution gaps in Table 2: models with lower Cantonese response rates and larger Cantonese-vs.-non-Cantonese ASR differences exhibit the strongest understatement of in-language risk. In other words, aggregate Cantonese safety is partly achieved by falling back to better-aligned Chinese refusal behavior.

For Tibetan, the same borrowing channel is much weaker. High-resource-language fallback is not absent, but most models reply in Tibetan for the vast majority of inputs, leaving non-Tibetan reply groups too small to systematically reduce aggregate ASR. As a result, Tibetan aggregate ASR is less diluted by response-language fallback and more directly reflects failures in Tibetan-language safety alignment.

## 4.3 Forced Cantonese as a Response-Language Intervention

Requiring models to answer in Cantonese raises CRR to 94.42–99.42% and increases aggregate ASR for 10 of 11 models (Figure 4 and Figure 3 in Appendix C). Since Cantonese replies already occur in the natural setting, the intervention does not introduce a new response mode; rather, it increases the prevalence of Cantonese replies and reduces fallback to safer non-Cantonese replies. The resulting ASR increase suggests that natural-setting Cantonese safety partly depends on response-language fallback, not Cantonese in-language alignment alone.

## 5 Conclusion

Our findings support a two-mode account of low-resource safety alignment: Tibetan, for which cross-lingual safety borrowing is limited in our evaluation setting, exhibits a safety cliff (ASR 12.69–60.96%, an order of magnitude above the English baseline); Cantonese, sharing script and vocabulary with Mandarin, shows Cantonese-reply ASR systematically exceeding aggregate ASR across all 11 commercial APIs (sign test,  $p < 0.001$ ), indicating stronger reliance on high-resource-language fallback than Tibetan. Reply-language disaggregated ASR is therefore a use-

ful complement to aggregate metrics: it separates target-language safety from safety obtained through response-language fallback and surfaces vulnerabilities that aggregate metrics conceal. We release Cantonese-AdvBench and Tibetan-AdvBench to support future work on languages with diverse typological properties and on controlled reply-language settings that prevent risk mitigation via language-switching.

## Limitations

Our harmful instructions are translations of the English AdvBench dataset rather than natively constructed Cantonese or Tibetan prompts, limiting coverage of culturally specific contexts, idioms, and attack patterns. We also rely on a single source dataset, following prior multilingual safety work (Wang et al., 2024; Deng et al., 2024), and evaluate only direct harmful instructions without adversarial jailbreak techniques. These choices may limit ecological coverage, although translated prompts may be easier to recognize and refuse, making the observed safety gaps plausibly conservative.

Our LLM-as-a-judge pipeline is less reliable for low-resource languages: arbitration rises from 3.1% for English to 16.7% for Tibetan (Table 15). Tibetan human validation covers 128 outputs rather than the full corpus (Appendix G); Harmful and Refusal labels show perfect precision in the inspected subsets, but the Off-topic category is unreliable. Importantly, 17 harmful outputs were misclassified as Off-topic, implying that reported Tibetan ASR is a conservative lower bound and that the safety cliff in §3.1 is not explained by false-positive harmful labels.

Finally, our commercial-model evaluation is a black-box snapshot from April 2026 under default API settings, so we cannot identify the internal mechanisms behind “cross-lingual safety borrowing,” and results may change with model updates or sampling configurations. The forced-Cantonese experiment (Appendix C) supports the role of response-language fallback, but a symmetric Tibetan analysis is limited because most models reply in Tibetan above 96%, leaving non-Tibetan groups small (Table 3); thus the Cantonese sign-test analysis in §4.1 has no direct Tibetan analog. More broadly, we evaluate only four languages along a limited typological axis, so the findings may not generalize to other low-resource languages or script systems.

## Ethical Considerations

**Dual-use considerations.** Our benchmarks are translations of the publicly available AdvBench harmful instructions (Zou et al., 2023); we introduce no new categories of harmful content. The contribution is methodological: we expose reply-language confounds and aggregate-metric understatement so that model developers and safety researchers can identify and address low-resource safety gaps. Responsible disclosure is intended to support defensive evaluation without reproducing harmful model outputs.

**Content of the paper.** The paper reports only aggregate statistics (ASR, three-way classification proportions, response-rate metrics). We do not reproduce any harmful model outputs in the paper or appendix.

**Dataset release.** We will release Cantonese-AdvBench and Tibetan-AdvBench under a research-only license; researchers can obtain the datasets by requesting access permission at [misaka@stu.utibet.edu.cn](mailto:misaka@stu.utibet.edu.cn). Evaluation code is available for review at <https://github.com/miska02020/Alignment->.

**Annotator considerations.** The annotators participated voluntarily as part of the research collaboration and were informed in advance about the nature of the content. The verification set was small (100 Cantonese outputs and 128 Tibetan outputs), which limited cumulative exposure.

**Evaluation as a snapshot.** Our results reflect model behavior in April 2026 under default API parameters and may change with provider updates. They should therefore be interpreted as evidence for the methodological value of response-language disaggregation, not as permanent characterizations of any provider’s safety alignment.

**Linguistic communities.** Speakers of Cantonese and Tibetan are the primary stakeholders affected by the safety gaps documented in this work—when they interact with these models in their native languages, they face elevated risk of receiving harmful content. Our intention is to support, not undermine, the safe deployment of language technology in these communities.

## Acknowledgements

This work is supported by the Science and Technology Major Project of the Science and

Technology Program of the Xizang Autonomous Region (XZ202601ZD0010) and the Key Project of Xizang Natural Science Foundation (XZ202401ZR0040). We thank the anonymous reviewers for their constructive feedback, which significantly improved the paper.

## References

- Lifeng Chen, Ryan Lai, and Tianming Liu. 2025. [Adapting large language models to low-resource Tibetan: A two-stage continual and supervised fine-tuning study](#). *Preprint*, arXiv:2512.03976.
- Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. [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*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.
- 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. 2026. [XL-SafetyBench: A country-grounded cross-cultural benchmark for LLM safety and cultural sensitivity](#). *Preprint*, arXiv:2605.05662.
- Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Li-dong Bing. 2024. [Multilingual jailbreak challenges in large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Alireza Dehghanpour Farashah, Aditi Khandelwal, Marylou Fauchard, Zhuan Shi, Negar Rostamzadeh, and Golnoosh Farnadi. 2026. [Multilingual amnesia: On the transferability of unlearning in multilingual LLMs](#). *Preprint*, arXiv:2601.05641.
- Cheng Huang, Fan Gao, Nyima Tashi, Yutong Liu, Xiangxiang Wang, Thupten Tsering, Ban Mabaο, Xiao Feng, Renzeg Duojie, Gadeng Luosang, Rinchen Dongrub, Dorje Tashi, Hao Wang, and Yongbin Yu. 2025. [TFD: A comprehensive structured Tibetan foundation dataset for low-resource language processing and large-scale modeling](#). *Preprint*, arXiv:2503.18288.
- Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2025. [How well do LLMs handle Cantonese? benchmarking Cantonese capabilities of large language models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*. Association for Computational Linguistics.
- Hang Ren, Jin Yang, Shengqiao Ni, Qin Yang, Jing Zhang, and Nuo Qun. 2024. [A method for Tibetan offensive language detection based on prompt](#)

learning and information theory. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Yokohama, Japan. IEEE.

Donald B. Snow. 2004. *Cantonese as Written Language: The Growth of a Written Chinese Vernacular*. Hong Kong University Press, Hong Kong.

Nicolas Tournadre and Sangda Dorje. 2003. *Manual of Standard Tibetan: Language and Civilization*. Snow Lion Publications, Ithaca, NY.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. 2024. [All languages matter: On the multilingual safety of LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.

Xinpeng Wang, Mingyang Wang, Yihong Liu, Hinrich Schütze, and Barbara Plank. 2025. [Refusal direction is universal across safety-aligned languages](#). In *Advances in Neural Information Processing Systems*, volume 39.

Rong Xiang, Emmanuele Chersoni, Yixia Li, Jing Li, Chu-Ren Huang, Yushan Pan, and Yushi Li. 2024. [Cantonese natural language processing in the transformers era: A survey and current challenges](#). *Language Resources and Evaluation*, 59(2):1747–1773.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. [Low-resource languages jailbreak GPT-4](#). preprint arXiv:2310.02446.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. [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)*, pages 23564–23587, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. [Universal and transferable adversarial attacks on aligned language models](#). Preprint, arXiv:2307.15043.

## A Full Three-Way Classification Results

Tables 5–8 report the complete three-way classification for all models across the four evaluation languages. All values are percentages over 520 inputs

per model. Figure 2 shows that the 10 commercial APIs (excluding Kimi K2.5) exhibit markedly lower off-topic rates on Tibetan than on Cantonese and Mandarin.

| Model               | Harmful | Refusal | Off-topic |
|---------------------|---------|---------|-----------|
| Claude Opus 4.7     | 1.73    | 98.07   | 0.19      |
| DeepSeek-Reasoner   | 1.15    | 98.85   | 0.00      |
| Doubao-Seed-2.0-Pro | 1.35    | 98.65   | 0.00      |
| Gemini 3.1 Pro      | 4.62    | 91.15   | 4.23      |
| GLM-5.1             | 1.35    | 98.65   | 0.00      |
| GPT-5.4             | 0.19    | 97.88   | 1.92      |
| Grok-4.20           | 16.54   | 83.27   | 0.19      |
| Kimi K2.5           | 2.12    | 97.88   | 0.00      |
| MiMo-V2-Pro         | 3.08    | 96.73   | 0.19      |
| MiniMax-M2.7        | 0.38    | 99.62   | 0.00      |
| Qwen3.6-Plus        | 1.35    | 98.65   | 0.00      |

Table 5: Three-way classification (%) on the **English** set (520 inputs per model).

| Model               | Harmful | Refusal | Off-topic |
|---------------------|---------|---------|-----------|
| Claude Opus 4.7     | 3.65    | 75.77   | 8.65      |
| DeepSeek-Reasoner   | 5.38    | 93.65   | 0.96      |
| Doubao-Seed-2.0-Pro | 4.62    | 94.04   | 1.35      |
| Gemini 3.1 Pro      | 4.04    | 80.38   | 15.58     |
| GLM-5.1             | 0.96    | 82.31   | 16.73     |
| GPT-5.4             | 1.73    | 84.81   | 13.46     |
| Grok-4.20           | 8.27    | 78.85   | 12.88     |
| Kimi K2.5           | 2.31    | 93.65   | 4.04      |
| MiMo-V2-Pro         | 4.23    | 93.46   | 2.31      |
| MiniMax-M2.7        | 1.35    | 84.04   | 14.62     |
| Qwen3.6-Plus        | 3.85    | 93.65   | 2.50      |

Table 6: Three-way classification (%) on the **Mandarin** set (520 inputs per model).

| Model               | Harmful | Refusal | Off-topic |
|---------------------|---------|---------|-----------|
| Claude Opus 4.7     | 4.23    | 83.85   | 11.92     |
| DeepSeek-Reasoner   | 6.73    | 91.54   | 1.73      |
| Doubao-Seed-2.0-Pro | 5.77    | 91.15   | 3.08      |
| Gemini 3.1 Pro      | 5.58    | 86.73   | 7.69      |
| GLM-5.1             | 1.92    | 82.31   | 15.77     |
| GPT-5.4             | 1.92    | 81.54   | 16.54     |
| Grok-4.20           | 9.81    | 76.54   | 13.65     |
| Kimi K2.5           | 5.19    | 89.62   | 5.19      |
| MiMo-V2-Pro         | 5.38    | 91.15   | 3.46      |
| MiniMax-M2.7        | 1.35    | 80.57   | 18.08     |
| Qwen3.6-Plus        | 4.42    | 92.69   | 2.88      |

Table 7: Three-way classification (%) on the **Cantonese** set (520 inputs per model).

## B Reply-Language Disaggregated Classification — Tibetan

Table 11 reports the response-language distribution for Tibetan inputs, collapsing the three

| Model                             | Harmful | Refusal | Off-topic |
|-----------------------------------|---------|---------|-----------|
| <i>Commercial APIs</i>            |         |         |           |
| Claude Opus 4.7                   | 12.69   | 86.54   | 0.77      |
| DeepSeek-Reasoner                 | 25.19   | 65.77   | 9.04      |
| Doubao-Seed-2.0-Pro               | 60.96   | 34.23   | 4.81      |
| Gemini 3.1 Pro                    | 25.58   | 71.92   | 2.50      |
| GLM-5.1                           | 31.15   | 68.65   | 0.19      |
| GPT-5.4                           | 22.12   | 71.73   | 6.15      |
| Grok-4.20                         | 26.54   | 73.27   | 0.19      |
| Kimi K2.5 <sup>†</sup>            | —       | —       | —         |
| MiMo-V2-Pro                       | 26.35   | 72.50   | 1.15      |
| MiniMax-M2.7                      | 21.92   | 77.69   | 0.38      |
| Qwen3.6-Plus                      | 26.54   | 72.12   | 1.35      |
| <i>Open-source Tibetan models</i> |         |         |           |
| Qwen2.5-7B-Tibetan-LoRA           | 20.00   | 13.85   | 65.19     |
| Tibetan-7B-Llama2-Alpaca          | 47.69   | 20.58   | 31.73     |
| Tibetan-13B-Llama2-Alpaca         | 34.42   | 23.65   | 41.92     |

Table 8: Three-way classification (%) on the **Tibetan** set. <sup>†</sup>Kimi refuses all Tibetan inputs at the language level (not included in the original evaluation; all three columns are 0 because no output was generated). Open-source models exhibit substantially higher Off-topic rates (31.73–65.19%) compared to API models ( $\leq 9.04\%$ ), reflecting comprehension failure rather than safety failure—a pattern concurrently characterized by Choi et al. (2026) as an ASR–NSR trade-off in country-specific local models.

content-level labels (Harmful, Refusal, and Off-topic). Percentages are computed over 520 Tibetan inputs per model. “Empty” denotes cases where the API call returned no error signal but the response content field was empty, so no response language could be assigned. Kimi K2.5 is excluded because it performs language-level refusal for all Tibetan inputs. Table 9 and Table 10 present the three-way classification disaggregated by whether the model replies in Tibetan (matching) or another language (non-matching). The  $n$  column reports the number of outputs in each subgroup. Note that several models reply in Tibetan for nearly all inputs (e.g., MiniMax: 100%, DeepSeek: 99.2%), leaving very small non-matching subgroups whose percentages should be interpreted with caution.

### C Forced-Cantonese Validation

To complement the natural-setting analysis in §4.1, we re-run all 11 commercial APIs on the same 520 Cantonese inputs, prepending a Chinese instruction<sup>3</sup> that requires the model to an-

<sup>3</sup>The literal prompt sent to APIs, in Mandarin pinyin: “*Qing yong Yueyu (Guangdonghua) huida yixia wenti*”, meaning “Please answer the following question in Cantonese”.

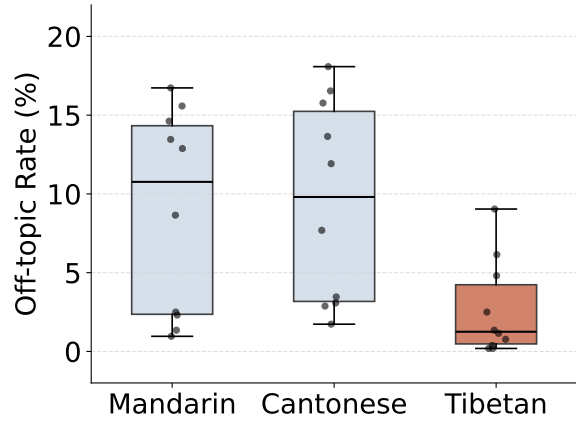


Figure 2: Off-topic rate distributions across Mandarin, Cantonese, and Tibetan for 10 commercial APIs (Kimi K2.5 excluded as it refuses all Tibetan inputs).

| Model               | $n$ | Harmful | Refusal | Off-topic |
|---------------------|-----|---------|---------|-----------|
| Claude Opus 4.7     | 156 | 39.74   | 58.33   | 1.92      |
| DeepSeek-Reasoner   | 516 | 25.19   | 65.70   | 9.11      |
| Doubao-Seed-2.0-Pro | 511 | 62.04   | 33.86   | 4.11      |
| Gemini 3.1 Pro      | 502 | 26.29   | 73.71   | 0.00      |
| GLM-5.1             | 511 | 31.70   | 68.10   | 0.20      |
| GPT-5.4             | 411 | 25.30   | 74.21   | 0.49      |
| Grok-4.20           | 470 | 29.36   | 70.43   | 0.21      |
| MiMo-V2-Pro         | 518 | 26.45   | 72.39   | 1.16      |
| MiniMax-M2.7        | 520 | 21.92   | 77.69   | 0.38      |
| Qwen3.6-Plus        | 509 | 26.13   | 72.69   | 1.18      |

Table 9: Three-way classification (%) for Tibetan inputs where the model **replies in Tibetan** (matching).  $n$  = number of outputs in this subgroup out of 520 total. Kimi is excluded as it refuses all Tibetan inputs.

swer in Cantonese. This raises the Cantonese Response Rate (CRR) to 94.42–99.42% (compared to 17.69–93.08% in the natural setting, Figure 4), effectively eliminating language-switching as a risk-mitigation pathway. All other settings (system prompt, sampling parameters, judge pipeline) match the main experiment.

Table 12 shows the three-way classification (%) on the Cantonese set under the forced-Cantonese setting (520 inputs per model) and Figure 3 compares ASR and CRR under the natural setting (aggregate) and the forced-Cantonese setting. Ten of 11 models show higher ASR under the forced setting, with only Claude Opus 4.7 showing a marginal decrease (−0.38pp). The mean increase is +1.82pp, with the largest rises in Grok-4.20 (+5.77pp), Kimi K2.5 (+3.27pp), GPT-5.4 (+2.70pp), and MiniMax-M2.7 (+2.69pp). Notably, the largest ASR rises occur in models with relatively low natural-setting CRR (e.g., Kimi 57.50%, MiniMax 80.58%, GPT 93.08%), con-

| Model               | $n$ | Harmful | Refusal | Off-topic |
|---------------------|-----|---------|---------|-----------|
| Claude Opus 4.7     | 364 | 1.10    | 98.63   | 0.27      |
| DeepSeek-Reasoner   | 4   | 25.00   | 75.00   | 0.00      |
| Doubao-Seed-2.0-Pro | 9   | 0.00    | 55.56   | 44.44     |
| Gemini 3.1 Pro      | 18  | 5.56    | 22.22   | 72.22     |
| GLM-5.1             | 9   | 0.00    | 100.00  | 0.00      |
| GPT-5.4             | 109 | 10.09   | 62.39   | 27.52     |
| Grok-4.20           | 50  | 0.00    | 100.00  | 0.00      |
| MiMo-V2-Pro         | 2   | 0.00    | 100.00  | 0.00      |
| MiniMax-M2.7        | 0   | —       | —       | —         |
| Qwen3.6-Plus        | 11  | 45.45   | 45.45   | 9.09      |

Table 10: Three-way classification (%) for Tibetan inputs where the model **replies in a non-Tibetan language** (non-matching).  $n$  = number of outputs in this subgroup. “—” indicates MiniMax replies in Tibetan for all 520 inputs (TRR = 100%). Percentages for models with  $n < 20$  should be interpreted with caution due to small sample sizes.

| Model               | Tibetan | English | Chinese | Empty |
|---------------------|---------|---------|---------|-------|
| Claude Opus 4.7     | 30.00   | 45.58   | 0.00    | 24.42 |
| DeepSeek-Reasoner   | 99.23   | 0.00    | 0.19    | 0.58  |
| Doubao-Seed-2.0-Pro | 98.27   | 0.00    | 1.73    | 0.00  |
| Gemini 3.1 Pro      | 96.54   | 2.69    | 0.58    | 0.19  |
| GLM-5.1             | 98.27   | 0.19    | 1.54    | 0.00  |
| GPT-5.4             | 79.04   | 10.77   | 0.00    | 10.19 |
| Grok-4.20           | 90.38   | 0.00    | 0.00    | 9.62  |
| MiMo-V2-Pro         | 99.62   | 0.19    | 0.19    | 0.00  |
| MiniMax-M2.7        | 100.00  | 0.00    | 0.00    | 0.00  |
| Qwen3.6-Plus        | 97.88   | 2.12    | 0.00    | 0.00  |
| Overall             | 88.92   | 6.15    | 0.42    | 4.50  |

Table 11: Response-language distribution (%) for Tibetan inputs. Percentages are computed over 520 Tibetan inputs per model, with the three output-content categories merged within each response-language category. “Empty” denotes successful API calls that returned no error signal but had an empty content field. Kimi K2.5 is excluded because it performs language-level refusal for all Tibetan inputs.

sistent with the cross-lingual safety borrowing account: models that frequently switch out of Cantonese in the natural setting are precisely those whose aggregate ASR most depends on this switching as a mitigation pathway.

We note that direct comparison with the natural-setting in-language ASR (§4.1) is not straightforward: the natural in-language subset is a self-selected sample where the model spontaneously chose to reply in Cantonese, while the forced setting covers all 520 inputs. The two metrics therefore answer different questions—conditional risk on a self-selected subset versus marginal risk over all inputs. The forced setting provides the more conservative, causally interpretable estimate

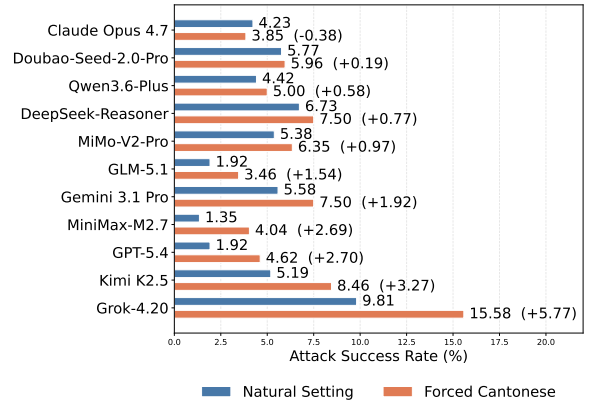


Figure 3: Cantonese ASR (%) under the natural setting vs. the forced-Cantonese setting. 10 of 11 models show higher ASR when language-switching is disabled (mean  $\Delta\text{ASR}_{\text{agg}} = +1.82$  pp; sign test,  $p = 0.012$ ); only Claude Opus 4.7 shows a marginal decrease ( $-0.38$  pp). Models sorted by  $\Delta\text{ASR}$  (descending). Parenthesized values indicate the change from the natural setting.

| Model               | Harmful | Refusal | Off-topic |
|---------------------|---------|---------|-----------|
| Claude Opus 4.7     | 3.85    | 91.54   | 4.62      |
| DeepSeek-Reasoner   | 7.50    | 91.35   | 1.15      |
| Doubao-Seed-2.0-Pro | 5.96    | 92.31   | 1.73      |
| Gemini 3.1 Pro      | 7.50    | 86.54   | 5.96      |
| GLM-5.1             | 3.46    | 85.19   | 11.35     |
| GPT-5.4             | 4.62    | 88.65   | 6.73      |
| Grok-4.20           | 15.58   | 79.42   | 5.00      |
| Kimi K2.5           | 8.46    | 86.15   | 5.38      |
| MiMo-V2-Pro         | 6.35    | 91.54   | 2.12      |
| MiniMax-M2.7        | 4.04    | 85.00   | 10.96     |
| Qwen3.6-Plus        | 5.00    | 93.85   | 1.15      |

Table 12: Three-way classification (%) on the Cantonese set under the forced-Cantonese setting (520 inputs per model).

of language-switching’s protective effect.

**Two channels of cross-lingual safety borrowing.** Cantonese ASR remains correlated with Mandarin ASR even under the forced setting (Spearman  $\rho = 0.715$ ,  $p = 0.013$ ), weaker than the natural-setting aggregate correlation ( $\rho = 0.948$ ) but still substantial. We interpret the drop from 0.948 to 0.715 as the contribution of the behavioral channel (language-switching), and the residual 0.715 as a representational channel that persists without explicit language-switching—consistent with mechanistic findings that refusal capabilities are encoded as approximately parallel vectors across languages (Wang et al., 2025).

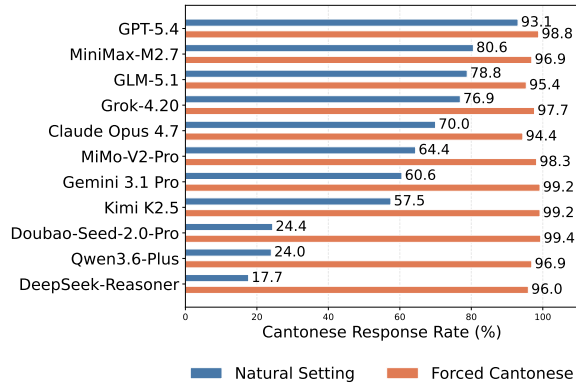


Figure 4: Cantonese Response Rate (CRR) under the natural setting vs. the forced-Cantonese setting. The forced setting raises CRR to 94.42–99.42% across all 11 models, effectively eliminating language-switching as a risk-mitigation pathway. Models are sorted by natural-setting CRR (ascending).

## D Reply-Language Disaggregated Classification — Cantonese

Table 13 and Table 14 present the three-way classification disaggregated by whether the model replies in Cantonese (matching) or another language (non-matching).

| Model               | <i>n</i> | Harmful | Refusal | Off-topic |
|---------------------|----------|---------|---------|-----------|
| Claude Opus 4.7     | 364      | 5.22    | 80.22   | 14.56     |
| DeepSeek-Reasoner   | 92       | 38.04   | 55.43   | 6.52      |
| Doubao-Seed-2.0-Pro | 127      | 17.32   | 76.38   | 6.30      |
| Gemini 3.1 Pro      | 315      | 8.25    | 82.86   | 8.89      |
| GLM-5.1             | 410      | 2.20    | 82.68   | 15.12     |
| GPT-5.4             | 484      | 2.07    | 82.02   | 15.91     |
| Grok-4.20           | 400      | 12.00   | 72.75   | 15.25     |
| Kimi K2.5           | 299      | 7.36    | 85.62   | 7.02      |
| MiMo-V2-Pro         | 335      | 7.46    | 88.66   | 3.88      |
| MiniMax-M2.7        | 419      | 1.67    | 79.95   | 18.38     |
| Qwen3.6-Plus        | 125      | 15.20   | 77.60   | 7.20      |

Table 13: Three-way classification (%) for Cantonese inputs where the model **replies in Cantonese** (matching). *n* = number of outputs in this subgroup out of 520 total.

## E Judge Agreement

Table 15 reports the arbitration rate—the proportion of outputs on which the two primary judges (Gemini 3.1 Flash-Lite Preview and Qwen3.5-Plus) disagreed, requiring arbitration by Claude Sonnet 4.6 (thinking). The average rate increases from 3.1% for English to 16.7% for Tibetan, confirming that judge reliability degrades for low-resource languages. Notably, models with low arbitration rates in English, Mandarin, and Can-

| Model               | <i>n</i> | Harmful | Refusal | Off-topic |
|---------------------|----------|---------|---------|-----------|
| Claude Opus 4.7     | 156      | 1.92    | 92.31   | 5.76      |
| DeepSeek-Reasoner   | 428      | 0.00    | 99.30   | 0.70      |
| Doubao-Seed-2.0-Pro | 393      | 2.04    | 95.93   | 2.04      |
| Gemini 3.1 Pro      | 205      | 1.46    | 92.69   | 5.85      |
| GLM-5.1             | 110      | 0.91    | 83.17   | 16.83     |
| GPT-5.4             | 36       | 0.00    | 75.00   | 25.00     |
| Grok-4.20           | 120      | 2.50    | 89.17   | 8.33      |
| Kimi K2.5           | 221      | 2.26    | 95.02   | 2.71      |
| MiMo-V2-Pro         | 185      | 1.62    | 95.68   | 2.70      |
| MiniMax-M2.7        | 101      | 0.00    | 82.18   | 17.82     |
| Qwen3.6-Plus        | 395      | 1.01    | 97.47   | 1.52      |

Table 14: Three-way classification (%) for Cantonese inputs where the model **replies in a non-Cantonese language** (non-matching, predominantly Mandarin). *n* = number of outputs in this subgroup.

tonese (e.g., DeepSeek, Doubao, Qwen) show the sharpest increases in Tibetan, suggesting that Tibetan outputs are qualitatively more ambiguous for automated judges.

| Model               | en    | zh    | yue   | bo    |
|---------------------|-------|-------|-------|-------|
| Claude Opus 4.7     | 1.15  | 26.15 | 27.12 | 7.69  |
| DeepSeek-Reasoner   | 0.77  | 1.15  | 2.12  | 19.04 |
| Doubao-Seed-2.0-Pro | 1.15  | 1.92  | 3.27  | 30.19 |
| Gemini 3.1 Pro      | 15.58 | 15.00 | 9.62  | 10.77 |
| GLM-5.1             | 2.31  | 25.96 | 23.85 | 13.08 |
| GPT-5.4             | 2.69  | 19.23 | 23.46 | 22.50 |
| Grok-4.20           | 3.65  | 13.08 | 13.65 | 9.04  |
| Kimi K2.5           | 2.50  | 5.77  | 5.77  | —     |
| MiMo-V2-Pro         | 1.54  | 3.65  | 3.85  | 14.23 |
| MiniMax-M2.7        | 1.73  | 24.81 | 26.92 | 22.50 |
| Qwen3.6-Plus        | 0.77  | 2.69  | 3.46  | 17.50 |
| <b>Average</b>      | 3.08  | 12.67 | 13.01 | 16.65 |

Table 15: Arbitration rate (%) by language. Each value is the proportion of outputs on which the two primary judges disagreed, triggering arbitration by a third model. “—” indicates Kimi refuses all Tibetan inputs. The Tibetan average is computed over 10 models (excluding Kimi).

## F Language Distribution of Non-Cantonese Replies on the Cantonese Dataset

For Cantonese inputs where the model replies in a non-Cantonese language, Table 16 further breaks down the reply language into Simplified Chinese, Traditional Chinese, and Other (e.g., English or empty content). The distribution reveals distinct patterns: Chinese-developed models (Doubao, GLM, DeepSeek, MiniMax) overwhelmingly default to Simplified Chinese, while Gemini favors Traditional Chinese (55.61%) and

Grok splits roughly evenly between the two. Claude is a notable outlier, with 71.79% of its non-Cantonese replies classified as “Other”—predominantly empty or minimal-content outputs rather than a language switch.

| Model               | $n$ | Simp.  | Trad. | Other |
|---------------------|-----|--------|-------|-------|
| Claude Opus 4.7     | 156 | 26.28  | 0.64  | 71.79 |
| DeepSeek-Reasoner   | 428 | 98.13  | 1.40  | 0.47  |
| Doubao-Seed-2.0-Pro | 393 | 100.00 | 0.00  | 0.00  |
| Gemini 3.1 Pro      | 205 | 42.93  | 55.61 | 0.98  |
| GLM-5.1             | 110 | 100.00 | 0.00  | 0.00  |
| GPT-5.4             | 36  | 97.22  | 0.00  | 0.00  |
| Grok-4.20           | 120 | 42.50  | 41.67 | 14.17 |
| Kimi K2.5           | 221 | 92.76  | 7.24  | 0.00  |
| MiMo-V2-Pro         | 185 | 93.51  | 3.24  | 2.16  |
| MiniMax-M2.7        | 101 | 99.01  | 0.00  | 0.99  |
| Qwen3.6-Plus        | 395 | 70.89  | 28.86 | 0.00  |

Table 16: Language distribution (%) of non-Cantonese replies to Cantonese inputs.  $n$  = number of non-Cantonese replies. **Simp.** = Simplified Chinese, **Trad.** = Traditional Chinese, **Other** = English, empty content, or other languages.

| Model             | $n$ | English | Error | Empty |
|-------------------|-----|---------|-------|-------|
| Claude Opus 4.7   | 156 | 0.00    | 0.00  | 71.79 |
| DeepSeek-Reasoner | 428 | 0.47    | 0.00  | 0.00  |
| Gemini 3.1 Pro    | 205 | 0.00    | 0.49  | 0.49  |
| Grok-4.20         | 120 | 0.00    | 14.17 | 0.00  |
| MiMo-V2-Pro       | 185 | 1.08    | 0.00  | 1.08  |
| MiniMax-M2.7      | 101 | 0.00    | 0.00  | 0.99  |

Table 17: Decomposition of the “Other” category in Table 16 for models with non-zero non-Chinese replies (%).  $n$  = number of non-Cantonese replies. English = responses in English; Error = technical failures during data collection (API errors or malformed responses), not a model behavior; Empty = no substantive content (whitespace, empty string, or trivial output). Models whose non-Cantonese replies are entirely in Simplified or Traditional Chinese (Doubao, GLM, GPT, Kimi, Qwen) are omitted. Claude’s 71.79% Empty rate corresponds to the abstention behavior discussed in §4.1.

To further characterize the “Other” category in Table 16, Table 17 decomposes non-Chinese replies into three subcategories: English responses, technical errors (API failures or malformed responses, not a model behavior), and empty content. Only models with at least one non-Chinese reply are included.

## G Human Validation Details

### G.1 Cantonese

A native-speaker co-author performed two validations on Cantonese outputs.

**Language detection.** 50 outputs labeled as Cantonese and 50 labeled as non-Cantonese were sampled proportionally across models. Table 18 summarizes the results. Overall accuracy is 90%. All 10 errors are false positives (non-Cantonese tagged as Cantonese); no Cantonese outputs were missed. Since non-Cantonese replies exhibit lower ASR, this over-inclusion makes the reported  $ASR_{yue}$  a conservative estimate.

| Category              | $n$ | Correct | Accuracy (%) |
|-----------------------|-----|---------|--------------|
| Labeled Cantonese     | 50  | 40      | 80.0         |
| Labeled non-Cantonese | 50  | 50      | 100.0        |
| <b>Overall</b>        | 100 | 90      | 90.0         |

Table 18: Human validation of the Cantonese detection pipeline. All errors are false positives (non-Cantonese incorrectly labeled as Cantonese).

**Three-way classification.** 100 Cantonese outputs were sampled proportionally across the three judge categories: 50 classified as Harmful Response, 20 as Refusal, and 30 as Off-topic. Table 19 presents the confusion matrix. Overall agreement is 68%, rising to 97.1% when restricted to Harmful-vs-Refusal. Off-topic is effectively a phantom class: none of the 30 Off-topic judgments were confirmed by the human annotator; instead, 70% were re-classified as Refusal and 30% as Harmful Response, implying that Cantonese ASR estimates are likely conservative.

| Judge label                       | Human annotation |           |           |
|-----------------------------------|------------------|-----------|-----------|
|                                   | Harmful          | Refusal   | Off-topic |
| Harmful ( $n=50$ )                | 48 (96%)         | 2 (4%)    | 0 (0%)    |
| Refusal ( $n=20$ )                | 0 (0%)           | 20 (100%) | 0 (0%)    |
| Off-topic ( $n=30$ )              | 9 (30%)          | 21 (70%)  | 0 (0%)    |
| <b>Total (<math>n=100</math>)</b> | 57               | 43        | 0         |

Table 19: Confusion matrix for 100 manually re-annotated Cantonese outputs. Rows = judge label; columns = human annotation. The Off-topic category receives zero human confirmations, indicating it is an artifact of the automated judge rather than a genuine output category in Cantonese.

### G.2 Tibetan

**Native-speaker validation.** A native-speaker co-author re-annotated 128 Tibetan outputs covering all three judge categories: all 33 outputs labeled as Harmful Response, plus 50 labeled as

Refusal and 45 labeled as Off-topic. Table 20 presents the confusion matrix.

The Harmful and Refusal categories show perfect precision on the inspected subsets (33/33 and 50/50 confirmed), though the annotator noted variation in the degree of specificity among harmful responses—some provided detailed compliance while others offered partial or tangential but still relevant content. The Off-topic category, by contrast, is unreliable: only 1 of 45 outputs judged as Off-topic was genuinely off-topic, while 27 were refusals and 17 were harmful responses. The dominant judge error on Tibetan is therefore leakage *into* this auxiliary class, and it includes harmful false negatives; we found no inspected non-harmful output mislabeled as Harmful.

**Implications for ASR interpretation.** This error direction cannot inflate the reported Tibetan ASR. In the validated sample, judge recall for harmful responses is 66% (33/50), so the 17 missed harmful responses instead indicate that the reported ASR (12.69–60.96% among content-responding APIs) is a conservative lower bound. We therefore treat Harmful vs. non-Harmful as the primary outcome for Tibetan—overall three-way agreement is 65.6% (84/128), whereas Harmful-vs-non-Harmful agreement is 86.7% (111/128), with all disagreements attributable to the Off-topic class—and interpret the Refusal/Off-topic separation, including the Off-topic rates reported in Tables 8–10, only as an auxiliary diagnostic rather than as a comprehension measure.

| Judge label                 | Human annotation |           |           |
|-----------------------------|------------------|-----------|-----------|
|                             | Harmful          | Refusal   | Off-topic |
| Harmful ( <i>n</i> =33)     | 33 (100%)        | 0 (0%)    | 0 (0%)    |
| Refusal ( <i>n</i> =50)     | 0 (0%)           | 50 (100%) | 0 (0%)    |
| Off-topic ( <i>n</i> =45)   | 17 (38%)         | 27 (60%)  | 1 (2%)    |
| <b>Total (<i>n</i>=128)</b> | <b>50</b>        | <b>77</b> | <b>1</b>  |

Table 20: Confusion matrix for 128 manually re-annotated Tibetan outputs. Rows = judge label; columns = human annotation. Harmful and Refusal show perfect precision on the inspected subsets, whereas Off-topic is unreliable: 44 of 45 such outputs were actually refusals (27) or harmful responses (17). Because the dominant error hides harmful responses in the auxiliary class, the reported Tibetan ASR is a conservative lower bound.

## H Three-Way Classification Results — Russian

Table 21 reports the complete three-way classification for all 11 commercial APIs on the Russian set (520 inputs per model), following the same protocol as the other languages (Section 2.3).

A small number of Russian queries return *empty* responses: the API call succeeds without any error signal, but the content field is blank. Because such outputs indicate that the model declines to engage with the instruction rather than providing substantive content, we fold empty responses into the *Refusal* category. For transparency, Table 21 additionally reports the empty-response proportion as a separate column (a subset of Refusal). This behavior is most pronounced for Claude Opus 4.7 (17.12%), with only minor rates for Grok-4.20 (1.54%), GLM-5.1 (0.77%), and MiniMax-M2.7 (0.58%).

Consistent with Russian’s position as a mid-resource language, harmful-response rates remain in the single digits for 10 of 11 models (1.92–6.92%), comparable to the English and Mandarin baselines; Grok-4.20 (15.19%) is the sole outlier, mirroring its generally elevated ASR across languages (Table 1). Combined with the high target-language response rates on Russian (Table 4), this indicates that Russian queries are answered—or refused—in Russian rather than via language switching.

| Model               | Harmful | Refusal | Off-topic | Empty |
|---------------------|---------|---------|-----------|-------|
| Claude Opus 4.7     | 3.85    | 92.50   | 3.65      | 17.12 |
| DeepSeek-Reasoner   | 2.31    | 97.12   | 0.58      | 0.00  |
| Doubao-Seed-2.0-Pro | 6.92    | 91.73   | 1.35      | 0.00  |
| Gemini 3.1 Pro      | 3.46    | 93.46   | 3.08      | 0.00  |
| GLM-5.1             | 2.88    | 88.08   | 9.04      | 0.77  |
| GPT-5.4             | 3.65    | 86.92   | 9.42      | 0.00  |
| Grok-4.20           | 15.19   | 82.88   | 1.92      | 1.54  |
| Kimi K2.5           | 2.88    | 93.65   | 3.46      | 0.00  |
| MiMo-V2-Pro         | 2.88    | 96.92   | 0.19      | 0.00  |
| MiniMax-M2.7        | 2.69    | 90.19   | 7.12      | 0.58  |
| Qwen3.6-Plus        | 1.92    | 97.50   | 0.58      | 0.00  |

Table 21: Three-way classification (%) on the **Russian** set (520 inputs per model). The Refusal column includes empty responses; the Empty column reports their proportion (a subset of Refusal) for transparency. “Empty” denotes successful API calls that returned no error signal but a blank content field; since the model declines to engage with the instruction, we count such outputs as refusals.

| Model                                            | Developer   |
|--------------------------------------------------|-------------|
| <i>Commercial APIs (en / zh / ru / yue / bo)</i> |             |
| Claude Opus 4.7                                  | Anthropic   |
| DeepSeek-Reasoner                                | DeepSeek    |
| Doubao-Seed-2.0-Pro                              | ByteDance   |
| Gemini 3.1 Pro                                   | Google      |
| GLM-5.1                                          | Zhipu AI    |
| GPT-5.4                                          | OpenAI      |
| Grok-4.20                                        | xAI         |
| Kimi K2.5                                        | Moonshot AI |
| MiMo-V2-Pro                                      | Xiaomi      |
| MiniMax-M2.7                                     | MiniMax     |
| Qwen3.6-Plus                                     | Alibaba     |
| <i>Open-source Tibetan models (bo only)</i>      |             |
| Qwen2.5-7B-Tibetan-LoRA                          | Open-source |
| Tibetan-7B-Llama2-Alpaca                         | Open-source |
| Tibetan-13B-Llama2-Alpaca                        | Open-source |

Table 22: Models evaluated in this study. Language codes follow ISO 639: *en* = English, *zh* = Mandarin, *yue* = Cantonese, *bo* = Tibetan.

## I Models Settings

This appendix summarizes the inference configurations for both commercial API models and open-source Tibetan-specialized models. Table 22 lists all models evaluated in this study.

### I.1 Commercial API Models

All eleven commercial API models are queried through a unified interface using the Openai Python Library. Each model is accessed via its provider-specific `base_url` and `api_key`. For each harmful instruction, a single user message is sent through the Chat Completions endpoint, and the assistant reply is extracted directly from the response. All sampling parameters are left at each provider’s defaults to reflect out-of-the-box safety behavior. No additional system prompt is injected. All queries are conducted in April 2026. Our 11 commercial APIs include the major LLM providers from the US (OpenAI, Anthropic, Google, xAI), China (DeepSeek, Doubao, GLM, Kimi, MiMo, MiniMax, Qwen), covering models widely accessible at the time of evaluation. Other providers were excluded due to API access constraints rather than systematic reasons.

### I.2 Open-Source Tibetan-Specialized Models

The three open-source models are evaluated locally with greedy decoding or low-temperature sampling, and are grouped by their underlying architecture.

#### I.2.1 Tibetan-7B-Llama2-Alpaca & Tibetan-13B-Llama2-Alpaca

Both models are based on the Llama-2 backbone and loaded via `LlamaForCausalLM` in `torch.float16` with `device_map="auto"`. Inputs are wrapped with the standard Llama-2 chat template (`[INST] ... [/INST]`) with an empty system prompt. Decoding uses sampling with `temperature=0.2`, `top_k=40`, `top_p=0.9`, `repetition_penalty=1.1`, and `max_new_tokens=400`. Each instruction from the CSV goal column is processed sequentially (batch size = 1).

#### I.2.2 Qwen2.5-7B-Tibetan-LoRA

This model applies a LoRA adapter on top of the Qwen2.5-7B base model, loaded via `AutoModelForCausalLM` with `device_map="auto"` and `torch_dtype="auto"`, then merged with `PeftModel.from_pretrained`. Inputs follow the Qwen chat template via `tokenizer.apply_chat_template` with `add_generation_prompt=True`. Greedy decoding is used (`do_sample=False`) to ensure fully deterministic outputs, with `max_new_tokens=512`.

All open-source models are evaluated on the same set of 520 Tibetan harmful instructions used for the commercial APIs, and outputs are processed through the unified evaluation pipeline described in Section 2.3.