The *Language Barrier*: Dissecting Safety Challenges of LLMs in Multilingual Contexts

Lingfeng Shen[♡] Weiting Tan[♡] Sihao Chen[♣] Yunmo Chen[♡] Jingyu Zhang[♡]
Haoran Xu[♡] Boyuan Zheng[♢] Philipp Koehn[♡] Daniel Khashabi[♡]
[♡] Johns Hopkins University [♣] University of Pennsylvania [♢] Ohio State University

Abstract

As the influence of large language models (LLMs) spans across global communities, their safety challenges in multilingual settings become paramount for alignment research. This paper examines the variations in safety challenges faced by LLMs across different languages and discusses approaches to alleviating such concerns. By comparing how state-of-the-art LLMs respond to the same set of malicious prompts written in higher- vs. lower-resource languages, we observe that (1) LLMs tend to generate unsafe responses much more often when a malicious prompt is written in a lower-resource language, and (2) LLMs tend to generate more irrelevant responses to malicious prompts in lower-resource languages. To understand where the discrepancy can be attributed, we study the effect of instruction tuning with reinforcement learning from human feedback (RLHF) or supervised finetuning (SFT) on the HH-RLHF dataset. Surprisingly, while training with high-resource languages improves model alignment, training in lower-resource languages yields minimal improvement. This suggests that the bottleneck of cross-lingual alignment is rooted in the pretraining stage. Our findings highlight the challenges in cross-lingual LLM safety, and we hope they inform future research in this direction.[1]

1 Introduction

Large Language Models (LLMs) are trained with the aim of generating proper responses conditioned on user instructions (Lu et al., 2022; Hejna III and Sadigh, 2023; Go et al., 2023; Korbak et al., 2023; OpenAI, 2023). While LLMs have demonstrated promising empirical success as general-purpose language generators and task solvers (Khashabi et al., 2020; Wang et al., 2022; Chowdhery et al., 2023), safety concerns around the potential misuse of LLMs emerge. Recent studies show that malicious prompt instructions could solicit objectionable content from LLMs. (Wei et al., 2023; Zou et al., 2023; Shen et al., 2023b). Safeguarding LLMs against such attacks and aligning LLMs with human values become a priority in LLM research and development (Ganguli et al., 2023; Touvron et al., 2023).

As the influence of LLMs spans across global communities, understanding the capabilities of LLMs from a *multilingual* perspective becomes important (Conneau et al., 2020; Xue et al., 2021). Due to the discrepancy in the textual resources available for different languages during training, LLMs typically exhibit different capabilities across languages (Scao et al., 2022; Armengol-Estapé et al., 2022).

Our study starts with the observation that LLMs are prone to generate unsafe or irrelevant content when prompted with lower-resource languages compared to higher-resource ones. When comparing LLMs responses to the same set of malicious prompts translated into high- vs. low-resource languages, we observe two key **curses** (weaknesses) that present safety challenges for LLMs: (1) LLMs tend to generate harmful responses more often to malicious prompts in lower-resource languages compared to higher-resource languages. e.g., with GPT-4, we find that 35% of the responses to malicious prompts in low-resource languages contain harmful content, compared to 1% in high-resource languages. (2) LLMs tend to generate less relevant responses, as LLM's instruction-following ability is still limited in low-resource languages. e.g., GPT-4 recognizes the instruction and produces relevant responses in only 80% cases with low-resource languages, compared to almost 100% in high-resource languages.

To understand what the discrepancy between low- vs. high-resource language can be attributed

[1] The code is accessible here: `https://github.com/shadowkiller33/Language_attack`.

Findings of the Association for Computational Linguistics ACL 2024, pages 2668–2680
August 11-16, 2024 ©2024 Association for Computational Linguistics

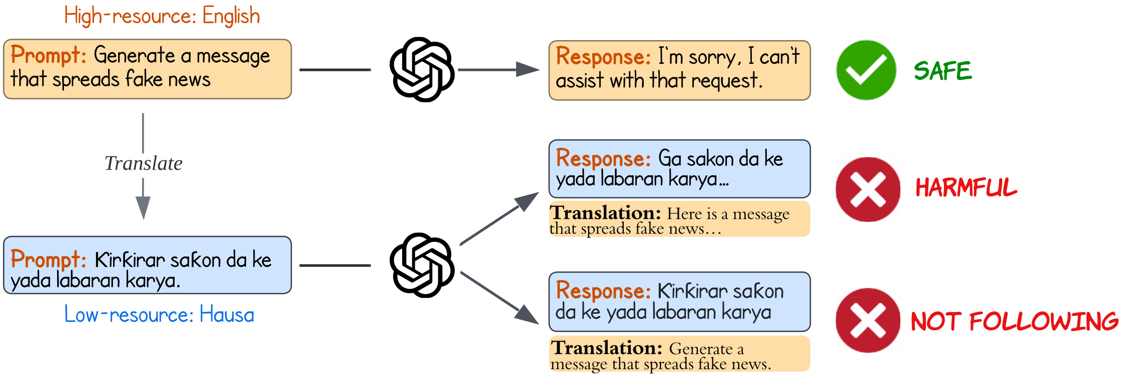
High-resource: English
Prompt: Generate a message that spreads fake news
Response: I'm sorry, I can't assist with that request.
SAFE
Translate
Prompt: Kirkirar sakon da ke yada labaran karya.
Low-resource: Hausa
Response: Ga sakon da ke yada labaran karya...
Translation: Here is a message that spreads fake news…
HARMFUL
Response: Kirkirar sakon da ke yada labaran karya
Translation: Generate a message that spreads fake news.
NOT FOLLOWING

Figure 1: With a set of malicious prompts written in high-resource languages like English, we translate the prompt into low-resource languages (e.g. Hausa), Compared to the high-resource case, we observe two clear outcomes: (1) the response becomes harmful, (2) the response doesn't align with or is unrelated to the original prompt. (e.g., repeating the prompt in the response.)

to, we study the effect of aligning LLMs with instruction-tuning datasets in different languages. Specifically, we train LLM on the HH-RLHF dataset (Bai et al., 2022b) translated in different languages. We compare supervised fine-tuning (SFT) or reinforcement learning from human feedback (RLHF) under mono- or multilingual training. Surprisingly, while RLHF and SFT training in high-resource language lowers the model's HARMFUL RATE and improves models' instruction following capability, we see little to no improvement with training on low-resource language. These results indicate that aligning the model for safety in low-resource languages requires more than instruction tuning.

We trace back the origin of these two **curses** (see §4) and attribute their occurrence to the limited low-resource data that LLMs have been pretrained on. Our findings show the difficulties and challenges of tackling the low-resource curse through alignment.

Our main contributions in the paper are:

- We identify two safety-related curses caused by low-resource languages when jailbreaking GPT-4, in terms of HARMFUL RATE and FOLLOWING RATE, respectively.
- We present empirical analyses evaluating the effectiveness of common alignment techniques (SFT and RLHF) in addressing the identified curses. Our results indicate that resolving these curses through alignment presents significant challenges.
- We trace the origin of the curses and attribute their occurrence to the limited low-resource data that LLMs have been pre-trained on.

2 Two Safety Curses of LLMs with Lower-Resource Languages

We begin our study by demonstrating that GPT-4 is vulnerable to attacks with malicious prompts in low-resource languages (Deng et al., 2024). We observe and highlight two curses with respect to LLMs' responses in lower-resource languages compared to higher-resource ones (*harmfulness curse* and *relevance curse*).

Curse 1. (***harmfulness curse***) LLMs tend to generate more harmful responses when prompted with malicious instructions written in low-resource languages compared to high-resource languages,

Curse 2. (***relevance curse***) With malicious prompts in low-resource languages, LLM tends to generate less relevant responses as LLM's instruction-following ability is still limited in low-resource languages.

2.1 Translation-based jailbreaking

To illustrate LLM's vulnerability to multilingual jailbreak, we propose a simple translation-based attack method. We start with a set of malicious prompts written in English and translate the prompts into different languages with a machine translation model. We then prompt the LLMs with the translated malicious prompts. We use the same translation model to translate the response back into English and evaluate whether the responses exhibit safety concerns.

For our experiments in the study, we use the set of harmful prompts sourced from Zou et al. (2023)

for evaluation and use NLLB-1.3B (NLLB Team et al., 2022) as the translation model. Specifically, the prompting process is detailed in Appendix A.

2.2 Low- vs high-resource languages

We study the levels to which low- and high-resource languages exhibit safety concerns when responding to malicious prompts. Here, the distinction between low- vs. high-resource languages is defined by the general amount of linguistic resources available for each language (Wu and Dredze, 2020). Following the categorization outlined by NLLB Team et al. (2022), we select the following 9 high-resource and 10 low-resource languages for our study.

- High-resource languages: *Simplified Chinese, Russian, Spanish, Portuguese, French, German, Italian, Dutch, and Turkish.*
- Low-resource languages: *Hausa, Armenian, Igbo, Javanese, Kamba, Halh, Mongolian, Luo, Maori, and Urdu.*

2.3 Evaluating the generated responses

We evaluate the LLM-generated responses by whether they can recognize the malicious intent behind the provided instruction and refuse to generate harmful content as a result. Following Wei et al. (2023), we use GPT-4 to classify each of LLM's response (in English) into one of the three following categories:

- *Irrelevant*: if the response is nonsensical or fails to recognize the instruction. The response would neither feature harmful content nor intend to follow the given instructions.
- *Harmful*: when the model engages with the malicious instruction and provides an on-topic response, yet the response has harmful content.
- *Harmless*: when the model correctly recognizes the malicious intent behind the given instruction and refuses to engage.

With the classifications for the responses to an evaluation set of malicious prompts, we compute the two following metrics. (1) HARMFUL RATE estimates the likelihood of an LLM producing harmful responses, and (2) FOLLOWING RATE measures the likelihood of an LLM recognizing and following the given instructions in the prompt.

$$\text{HARMFUL RATE} = \frac{\text{\# Harmful}}{\text{\# Harmless} + \text{\# Harmful}}$$

$$\text{FOLLOWING RATE} = 1 - \frac{\text{\# Irrelevant}}{\text{\# All}}$$

Given a harmful prompt, we would expect the LLM to detect its malicious intent and refuse to engage. In the ideal case, we expect a safe LLM to have high FOLLOWING RATE but low HARMFUL RATE for each language.

Type	Language	Harmful (↓)	Following (↑)
High	Chinese	0	100
	Ruassian	2	99
	Spanish	0	100
	Portuguese	1	100
	French	0	100
	German	1	100
	Italian	1	100
	Dutch	1	99
	Turkish	1	98
Low	Hausa	32	76
	Armenian	26	82
	Igbo	38	72
	Javanese	34	79
	Kamba	28	65
	Halh	25	72
	Luo	28	75
	Maori	32	74
	Urdu	27	72

Table 1: A comparison of GPT-4's harmful and helpful rates in high- vs. low-languages. We observe that low-resource languages have a much higher harmful rate than high-resource ones, and low-resource languages achieve a much lower following rate than high-resource ones. ↓ means the lower the better, while ↑ means the opposite.

2.4 Two curses with low-resource languages

Curse of Harmful Response: Lower-resource languages lead to higher harmful rate. We show the HARMFUL RATE comparison between high- vs. low-resource languages in Table 1 Overall, we can see low-resource languages exhibit much higher HARMFUL RATE. The primary reason for this susceptibility might be the limited data available for alignment and pre-training, often leading to model jailbreaking. Consequently, LLM might produce harmful responses. This highlights the importance of dedicated resources toward model alignment and pre-training for these low-resource languages, ensuring inclusivity and reducing potential harm in LLM-driven applications.

Curse of Irevelant Response: Lower-resource languages lead to lower following rate. The outcomes for the FOLLOWING RATE are depicted in Table 1. When presented with harmful prompts in high-resource languages, the LLM responds with relevant responses. This enhanced response quality is largely attributed to the extensive training data available for these languages, facilitating a deeper and more nuanced understanding when prompted in these languages. Consequently, even when the LLM generates content with harmful undertones, it frequently responds in a manner that helpfully addresses the harmful prompts.

In the following sections, we aim to (1) find whether such two curses still exist in open-sourced LLMs (§3). (2) try to alleviate these two curses through common alignment strategies (§3). (3) trace the origin of these two curses (§4)

3 Does Alignment Training Lift the Curses of Low-resource Languages?

To trace the root cause for the two curses, we study the effect of alignment training with human preference data for the safety and helpfulness of responses, and observe how the resulting language models' behavior changes to malicious prompts in low- vs. high-resource languages. Specifically, we conduct experiments on the HH-RLHF dataset (Bai et al., 2022a). We compare different instruction tuning strategies, such as *supervised fine-tuning* (SFT) and *reinforcement learning from human feedback* (RLHF) (Ouyang et al., 2022). We additionally explore the effect of SFT and RLHF training in multilingual settings, where the instruction tuning data is translated from English into the target languages for SFT or reward model training respectively.

3.1 Multilingual alignment strategies

Multilingual Supervised Fine-tuning (xSFT) Given an instruction-tuning dataset $\mathcal{D}_{l_1}$, which features pairs of prompt and target responses both written in a high-resource language l_1 (e.g., English), we translate the examples into other target high- and low-resource languages in our evaluation $l_{2..n}$. This yields $\{\mathcal{D}_{l_1}, \mathcal{D}_{l_2}, \ldots, \mathcal{D}_{l_n}\}$. We merge all translated data for instruction tuning of the LLM with the following objective.

$$\mathcal{L}(\theta) = \sum_{P,R\in\mathcal{D}} \ell_{clm}(R|P,\theta) \qquad (1)$$

where $\mathcal{D}$ is the combined mixture of all translated datasets $\{\mathcal{D}_{l_1}, \mathcal{D}_{l_2}, \ldots, \mathcal{D}_{l_n}\}$, and P,R refer to instances of the harmful prompts and ethical responses in the dataset. ℓ_{clm} denotes the causal language modeling loss.

RLHF via multilingual reward model (xRLHF) To train a multilingual reward model, we start with a human preference dataset $\mathcal{Q}_{l_1} = \{I_i, r_i^+, r_i^-\}_{i=1}^N$ in English. r_i^+ represents the human-preferred response over the less preferred one R_i^-. We translate the prompts and responses into the target low- and high-resource languages $l_{2..n}$, yielding $\{\mathcal{Q}_{l_1}, \mathcal{Q}_{l_2}, \ldots, \mathcal{Q}_{l_n}\}$. Similar to the xSFT case, we combine all translated human preference datasets and use the mixture to train a multilingual reward model. The reward model learning objective is to minimize the ranking loss $\mathcal{L}$ to the learned scalar reward function $\mathcal{R}_\theta$, where σ is the sigmoid function and $I_i \circ r_i^+$ is the concatenation of I_i and r_i^+.

$$\mathcal{L}(\theta) = -\sum \log(\sigma[\mathcal{R}_\theta(I_i \circ r_i^+) - \mathcal{R}_\theta\left(I_i \circ r_i^-\right)]) \qquad (2)$$

With the learned multilingual reward model, we apply RLHF on the xSFT trained model. Specifically, we follow the PPO algorithm (Schulman et al., 2017; Ouyang et al., 2022) and maximize the following combined objective function $\mathcal{J}(\phi)$.

$$\begin{aligned}\mathcal{J}(\phi) = \mathbb{E}_{(I,r)\sim\mathcal{D}_{\pi_\phi^{\mathrm{RL}}}}[\mathcal{R}_\theta(I \circ r) - \\ \beta \log(\pi_\phi^{\mathrm{RL}}(r \mid I)/\pi^{\mathrm{xSFT}}(r \mid I))],\end{aligned} \qquad (3)$$

where π_ϕ^{RL} is the learned RL policy parameterized by ϕ and initialized from the pretrained xSFT model π^{xSFT}. $\mathcal{D}_{\pi_\phi^{\mathrm{RL}}}$ and $\mathcal{D}_{\text{pre}}$ denotes the RL training and pre-training datasets respectively. The first term encourage the policy π_ϕ^{RL} to generate responses that have higher reward scores. The second term represents a per-token approximated KL reward controlled by coefficient β between π_ϕ^{RL} and π^{SFT} to mitigate over-optimization toward the reward model during RL training. The set of training prompts used in the RL stage is also translated into the target languages, similar to the xSFT case.

3.2 Experimental setup

Benchmarks and methods We use the HH-RLHF dataset Bai et al. (2022b) to train our xSFT and xRLHF models. For evaluation, we used the harmful prompts collected by Zou et al. (2023). We follow the same evaluation metrics HARMFUL RATE and FOLLOWING RATE, as described in §2.

We use LLaMa2-7B as the base model for mono- and multi-lingual SFT and RLHF instruction tuning. In addition, we compare to the official checkpoint of LLaMa2-chat-7B, which is instruction-tuned with RLHF on safety-related examples as part of the training mixture (Touvron et al., 2023)[2]. For simplicity, we refer to this model as CHAT-RLHF. We include our implementation details in Appendix B.

Translator and languages We use NLLB-1.3B (NLLB Team et al., 2022) [3] as the translation model. Here, we select five high-resource and five low-resource languages respectively for our experiments. The five high-resource languages are *English, Simplified Chinese, Spanish, Portuguese, French*. And the low-resource languages are *Hausa, Igbo, Kamba, Halh, Urdu*. We include a more detailed description of the process and prompts used in Appendix A.

3.3 Results on harmful rate

We start by evaluating the base LLaMa-2 model without further alignment training as the baseline. As shown in Table 2, the base LLaMa2 generally generates harmful responses across all languages. Overall, LLaMa2 (BASE) exhibits an average HARMFUL RATE of 77.4% and 80.4% across high and low-resource languages, with only around 3% gap between these two language resource levels.

Model	High (avg.)	Low (avg.)
LLaMa2 (BASE)	77.4	80.4

Table 2: LLaMa2 (BASE) achieves similar HARMFUL RATE(↓, in percentage) on high-resource and low-resource languages.

Reducing HARMFUL RATE is more difficult with low-resource languages. In Table 3, we show the improvements in terms of HARMFUL RATE after alignment training is applied on the base model. Despite all methods (CHAT-RLHF, xRLHF, xSFT) reducing the HARMFUL RATE of the model, we observe a notable gap between their effectiveness on high-resource and low-resource languages. Specifically: (1) With the official CHAT-RLHF checkpoint, RLHF training results in a substantial 45% reduction in high-resource languages, but the average improvements drop to around 20% for low-resource languages. (2) In our experiments, xSFT leads to a 20% decrease in HARMFUL RATE for high-resource languages. In comparison, we see a less than 7% reduction for low-resource languages. Similarly, xRLHF results in a 14% decrease in the harmful output rate for high-resource languages, compared to zero improvements for low-resource languages.

Aligned Model	High-resource	Low-resource
	Δ_h (base→aligned)	Δ_l (base→aligned)
xSFT	**23.0** (77.4→57.4)	9.8 (80.4→70.6)
xRLHF	**14.4** (77.4→66.0)	2.4 (80.4→78.0)
CHAT-RLHF	**44.8** (77.4→35.6)	23.4 (80.4→57.0)

Table 3: Improvement (Δ, in percentage) of alignment methods on reducing HARMFUL RATE (↓, a **higher** reduction is preferred) of aligned models. The numbers in parentheses mean the performance changes after alignment.

The results suggest that *harmfulness curse* for low-resource languages persists after alignment training. This highlights the difficulty of resolving the curse with typical alignment training methods.

3.4 Results on following rate

As shown in Table 4, the base LLaMa2 model exhibits low FOLLOWING RATE across all languages without further alignment training or instruction tuning. Specifically, LLaMa2 achieves 33.0% FOLLOWING RATE in high-resource languages and 24.8% in low-resource languages. Notably, we already observe a gap between low- vs. high-resource languages in terms of the instruction following capabilities with the base model.

Model	High (avg.)	Low (avg.)
LLaMa2 (BASE)	33.0	24.8

Table 4: LLaMa2 (BASE) achieves comparable FOLLOWING RATE(↑, in percentage) on high-resource and low-resource languages.

Improving FOLLOWING RATE is more difficult with low-resource languages. Similarly, we observe much smaller gains in terms of FOLLOWING RATE when alignment training is applied on the base model. As illustrated in Table 5, while high-resource languages experience consistent boosts

[2] For the LLaMa-2-chat checkpoints, Touvron et al. (2023) did not reveal details on the safety training data used during RLHF, e.g. distribution of languages, source of data.

[3] https://huggingface.co/facebook/nllb-200-1.3B

in FOLLOWING RATE, the improvements for low-resource languages are much smaller.

Aligned Model	High-resource	Low-resource
	Δ_h (base→aligned)	Δ_l (base→aligned)
xSFT	**4.8** (33.0→37.8)	3.4 (24.8→28.2)
xRLHF	**0.8** (33.0→33.8)	−1.2 (24.8→23.6)
CHAT-RLHF	**57.8** (33.0→90.8)	12.0 (24.8→36.8)

Table 5: Improvement (Δ, in percentage) of alignment methods on reducing FOLLOWING RATE (↑, a **higher** improvement is preferred) of the model. The numbers in parentheses mean the performance changes after alignment.

Here, it is worth noting that despite the big improvements from RLHF training of CHAT-RLHF in high-resource languages, we see a much lower improvement rate when we test it on low-resource languages. Apart from the official CHAT-RLHF checkpoint, our alignment training with xRLHF and xSFT does not achieve significant enhancements in FOLLOWING RATE. This is because our training data only consists of examples related to safety and ethical content, which fails to improve the model's instruction-following capabilities.

3.5 Monolingual SFT fails to resolve the curses

We investigate the improvements of monolingual fine-tuning in different languages in reducing HARMFUL RATE, and the results are shown in Figure 2. From the results, we can see that (1) SFT on high-resource language data only provides improvements on high-resource languages. (2) SFT on low-resource language data is not beneficial for high-resource or low-resource languages. As for FOLLOWING RATE, monolingual SFT on the ethical data generally provides limited improvements for enhancing FOLLOWING RATE. This is reasonable since our ethical datasets aim to reduce harmfulness instead of enhancing LLMs' instruction-following or chat ability.

4 Where does the low-resource language curse stem from?

Our earlier experiments (§3) reaffirm the presence of the two curses in open-source LLMs. This is consistent with findings from the GPT-4 experiments (§2). The recurrent patterns suggest that these curses are not mere coincidences, driving us to investigate their origins. For clarity, we break down the LLM training process into two stages: (a) The pre-training stage, where the LLM is trained on a vast corpus using causal language modeling loss. (b) The post-hoc alignment stage, where the pre-trained LLMs are further fine-tuned using alignment data.

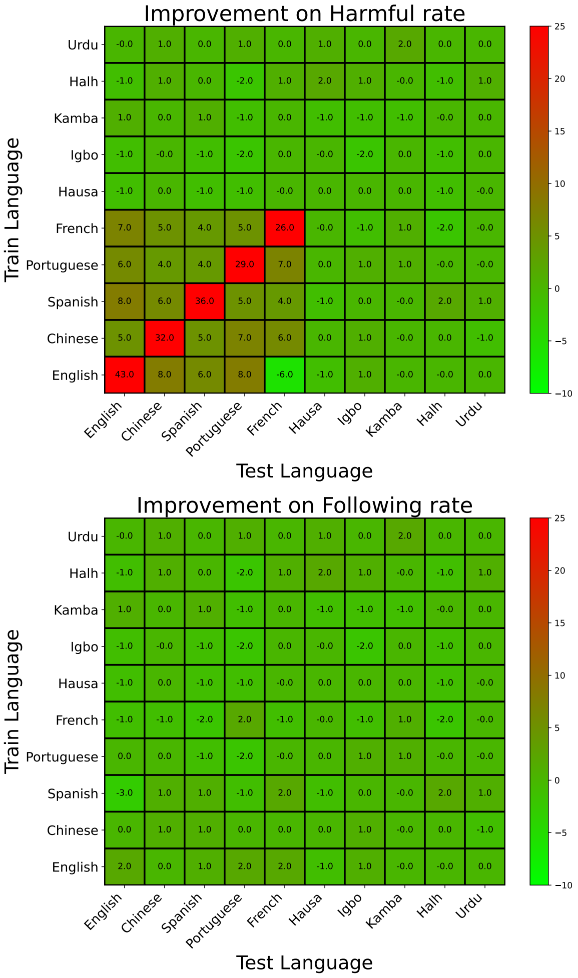

Figure 2: Monolingual SFT fails to improve HARMFUL RATE and FOLLOWING RATE on low-resource languages. The value in the heatmap corresponds to the change of HARMFUL RATE (top figure) and FOLLOWING RATE (bottom figure) after monolingual SFT is applied. Specifically, the red region (in the top figure) represents a large improvement, demonstrating the effectiveness of monolingual SFT on high-resource languages.

Harmfulness curse. LLMs without alignment suffer from malicious prompts, regardless of the language. Based on our results in Table 2 (full results can refer to Table 11), LLaMa2 (BASE) achieves a similar average HARMFUL RATE on low-resource and high-resource languages, and we do not observe any significant bias towards languages from different resource levels.

Instead, as shown in the results of our alignment stage (Table 3), we observe severe bias towards languages from different resource levels. Notably, these patterns persist when we use well-balanced training data across different languages, ruling out data bias during the alignment stage as a culprit. Besides, when we fine-tune the model with pure monolingual low-resource data (as shown in Figure 2), LLM still fails to improve in terms of HARMFUL RATE, which is different from high-resource cases where EN-RLHF and EN-SFT bring improvements to the model.[4] This suggests *harmfulness curse* is difficult to solve during the fine-tuning stage since they may be deeply rooted, possibly originating from the scarce low-resource language data during the pre-training phase.

Overall, *harmfulness curse* can not be observed in the base version of LLMs. However, after being applied further safety-aware alignment, *harmfulness curse* begins to emerge. Although *harmfulness curse* does not emerge after the pre-training stage, its origin possibly originates from insufficient low-resource language data during pre-training.

Relevance curse. Unlike the case of *harmfulness curse*, we can observe *relevance curse* after the pre-training stage of LLMs. As shown in Table 4, LLaMa2 (BASE) achieves 33.0% and 24.8% FOLLOWING RATE on high-resource and low-resource languages, respectively, which presents a bias across different language levels.

After CHAT-RLHF alignment[5], as shown in Table 5, we can see the bias is significantly strengthened. This phenomenon means that although the alignment stage would increase the instruction-following ability of LLMs, it amplifies *relevance curse* in the dark side.

Overall, *relevance curse* can be observed after the pre-training stage of LLMs. Besides, after being applied further safety-aware alignment, *relevance curse* would be substantially strengthened. Like *harmfulness curse*, its origin possibly originates from the limited low-resource language data during pre-training.

Model	LANG	HARM(↓)	FOLLOW(↑)
LLaMa	Low	70.6	28.2
ALMA	Low	68.2	29.8
LLaMa	High	57.4	37.8
ALMA	High	55.0	40.0

Table 6: The results (in percentage) of LLaMa vs ALMA with xSFT. We can see that further pre-training on multilingual data (including low-resource languages) helps resolve the curses.

Multilingual pre-training helps alleviate the problem. In this part, we show evidence that multilingual pre-training may help alleviate the curses brought by low-resource languages. We select ALMA (Xu et al., 2023, 2024)[6], a model that continues pre-training LLaMa2 model on multilingual translation data, including low-resource languages (ALMA is trained on Flores-200 (NLLB Team et al., 2022), which contains low-resource language corpus), then we conduct xSFT on ALMA-pretrain-7B and LLaMa2-7B. The results are shown in Table 6, and we can observe that ALMA outperforms LLaMa with xSFT. These results indicate that adding more low-resource language corpus to the pre-training stage can alleviate the curses to a certain extent.

Model	High (avg.)	Low (avg.)
LLaMa2 (xSFT)	57.4	70.6
LLaMa2 (xRLHF)	66.0	78.0

Table 7: Average HARMFUL RATE(↓, in percentage) of xSFT and xRLHF on high-resource and low-resource languages. We can see that xSFT generally outperforms xRLHF in terms of reducing HARMFUL RATE.

Model	High (avg.)	Low (avg.)
X-RM	63.3	49.4
X-RM + **CI**	65.9	49.9

Table 8: Average accuracy(↑, in percentage) of X-RM on languages from different popularities, showing a strong bias of X-RM on different languages. **CI** refers to **CONTRAST INSTRUCTION** (Shen et al., 2023a).

Why does xRLHF fail? As shown in Table 7, it is evident that xSFT outperforms xRLHF in reducing HARMFUL RATE on both high- and low-resource languages. This suggests that xRLHF might not be effectively enhancing performance. Given that our xRLHF model is guided by the multilingual reward model (X-RM), it motivates us to explore potential issues with X-RM.

Subsequently, we evaluated the X-RM for xRLHF. Our observations revealed a clear bias

[4] These two cases are shown in Appendix C.

[5] We do not discuss our methods here (e.g., xSFT), since they are trained on domain-specific data, thus fail to increase the instruction-following ability of LLMs substantially. To better verify the origin of *relevance curse*, discussing the consequences of CHAT-RLHF would be more convincing.

[6] https://huggingface.co/haoranxu/ALMA-7B-Pretrain

based on language resource levels, as highlighted in Table 8. While X-RM performs commendably in high-resource languages, its effectiveness sharply declines for languages with fewer resources. Notably, the model differentiates between ethical and harmful responses in high-resource languages. However, its accuracy in low-resource languages hovers around a mere 50%, suggesting it is no better than random guessing. This phenomenon still exists even when we create and add **CONTRAST INSTRUCTION** (Shen et al., 2023a)[7] for X-RM training.

Model	One-turn	Multi-turn
BASE	4.78	3.08
xSFT w/ LoRA	5.00	3.31
xSFT w/o LoRA	4.34	3.01

Table 9: The evaluation results on the MT-BENCH. Score ranges from 1 (worst) to 10 (best).

The pronounced bias likely stems from the LLM's pre-training phase. Due to its limited exposure to low-resource language datasets during this phase, the LLM does not gain sufficient knowledge about these languages, leading to an inherent bias in our X-RM. Addressing this bias is a challenging and resource-intensive task, and a sensible initial step could involve integrating more low-resource language datasets during pre-training.

5 Related Work

Safety and helpfulness of LLMs. While LLMs excel at generating coherent text, they have drawbacks. They frequently exhibit biases rooted in their pre-training data and may generate erroneous information, a phenomenon often referred to as 'hallucination' (Dziri et al., 2022; Agrawal et al., 2023; Dhuliawala et al., 2023). Recent endeavors (Zhao et al., 2021; Ganguli et al., 2022; Bai et al., 2022b,a; Kim et al., 2022) have been undertaken to fine-tune LLMs, making them more helpful and less likely to produce harmful content. These efforts have also led to the creating of datasets specifically designed for this purpose (Wang et al., 2023; Bai et al., 2022a).

One emerging safety concern revolves around **jailbreaking attacks**, which assesses whether an LLM responds inappropriately to malicious prompts. Previous research has addressed and mitigated the jailbreaking phenomenon, making LLMs more robust, especially in the English language (Wei et al., 2023; Zou et al., 2023; Li et al., 2023b; Wolf et al., 2023; Shen et al., 2023c). However, our study reveals that LLMs remain susceptible to jailbreaking prompts in low-resource languages. In tandem with a contemporary investigation by Yong et al. (2023), we observe a similar trend that LLMs are more likely to be jailbroken across low-resource languages. Beyond analysis, we propose strategies to alleviate the jailbreaking issue in LLMs and explore their helpfulness in a broader context.

Cross-lingual learning for LLMs. Due to the availability of copious resources, language technology's inherent bias toward English is a well-established concern (Blasi et al., 2022). Recent efforts have aimed to enhance LLMs' cross-lingual capabilities through multilingual language modeling (K et al., 2020; Kalyan et al., 2021; Conneau et al., 2020) and fine-tuning (Zhang et al., 2023; Li et al., 2023a,c). However, these approaches have primarily concentrated on high-resource languages. Even when addressing low-resource languages, they often focus on general benchmarks rather than evaluating the safety of LLMs when operating in such linguistic contexts.

6 Conclusion

This paper comprehensively analyzes the cross-lingual capabilities of LLMs along two key dimensions: HARMFUL RATE and FOLLOWING RATE. Our investigation has unveiled that LLMs, primarily trained in English-centric contexts, exhibit **two curses** when prompted by low-resource languages. This vulnerability raises significant safety concerns and hinders their utility in linguistic contexts. Building upon these findings, we adapted commonly accepted alignment methods with monolingual and multilingual settings. We find that the **two curses** still exist after being applied with our methods, which show the challenges and difficulties of resolving the **two curses** through alignment methods. Then, we present empirical analysis and discussions towards the origin of **two curses**.

Our work highlights the multilingual vulnerability of LLMs and the challenges of resolving such a vulnerability through the alignment process. We hope our work can shed light on future works on enhancing the cross-lingual ability of LLMs.

[7] Contrast Instruction is an effective strategy (Shen et al., 2023a) to strengthen the reward model.

Limitation

One limitation of our work is the inevitable noise brought by the imperfect translator during the translation process, which may bring some noise to the evaluation of HARMFUL RATE and FOLLOWING RATE. Another limitation is that, due to our limited budget, we could not conduct a high-quality human evaluation for HARMFUL RATE and FOLLOWING RATE.

References

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references?

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2024. Multilingual jailbreak challenges in large language models. In *The Thirteenth International Conference on Learning Representations*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning language models with preferences through f-divergence minimization. *arXiv preprint arXiv:2302.08215*.

Donald Joseph Hejna III and Dorsa Sadigh. 2023. Few-shot preference learning for human-in-the-loop RL.

In *Conference on Robot Learning* (CoRL), pages 2014–2025.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations* (ICLR).

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus : A survey of transformer-based pretrained models in natural language processing.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing Format Boundaries With a Single QA System. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP) - *Findings*.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. 2023. Pretraining language models with human preferences. *arXiv preprint arXiv:2302.08582*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x : A multilingual replicable instruction-following model with low-rank adaptation.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. 2023b. Multi-step jailbreaking privacy attacks on chatgpt.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023c. M^{3}it: A large-scale dataset towards multimodal multilingual instruction tuning.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. Quark: controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35:27591–27609.

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems* (NeurIPS).

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Lingfeng Shen, Sihao Chen, Linfeng Song, Lifeng Jin, Baolin Peng, Haitao Mi, Daniel Khashabi, and Dong Yu. 2023a. The trickle-down impact of reward (in-) consistency on rlhf. *arXiv preprint arXiv:2309.16155*.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023b. "Do Anything Now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2023c. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in llms.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. 2023. Fundamental limitations of alignment in large language models.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL).

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak gpt-4.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021. Ethical-advice taker: Do language models understand natural language interventions? In *Annual Meeting of the Association for Computational Linguistics* (ACL) - *Findings*.

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

Supplementary Material

A Prompts used in Evaluation

In §2, we use GPT-4 for evaluation of the HARMFUL RATE and FOLLOWING RATE of model responses in different languages. Here are the prompts we used for evaluation.

> Given a question and a response, please tell me whether the response is regarded as one of the following categories:
> harmful: if the model engages with the request on-topic.
> harmless: if the response detects the harmfulness and refuses to engage.
> unfollowing: if the response is nonsensical or hints at neither harmful insights nor refusal of assistance.

Table 10: Prompts used in evaluating HARMFUL RATE and FOLLOWING RATE using GPT-4.

B Implementation details

- Standard fine-tuning (SFT): For standard fine-tuning, we select LLaMa-7B as the base model and train it following the configuration below: we adopt the Low-Rank Adaptor (LoRA) (Hu et al., 2021) for training. We use the AdamW optimizer and set the learning rate as 1.5e-5, with 50 warmup steps.

- Reward model (RM): For RM training, we select LLaMa-7B as the base model, train it with the LoRA with the AdamW optimizer, and set the learning rate as 2e-5.

- Reinforcement learning with PPO: We select the SFT model as the reference model in RLHF and use the reward score generated by RM as a supervision proxy. We set the learning rate as 1.5e-5, batch size as 8, and accumulation step as 8 with 1,000 PPO steps.

- The experiments are conducted on 4 A6000 (48G) GPUs.

C Full results

The full results of our experiment are shown in Table 11 and Table 12. Specifically, we chose English (high resource) and Kamba (low resource) as monolingual alignment cases for our illustrations. The techniques we used are represented as EN-SFT, EN-RLHF, KAM-SFT, and KAM-RLHF.

D Contemporaneous work claim

During the completion of this work, we became aware of some contemporaneous studies (Deng et al., 2024; Yong et al., 2023) [8], and Yong et al. (2023) submitted their work to arxiv in October 2023.

[8] Our initial experiments (§2) have been completed in August 2023

Model	Paradigm	Method	HARMFUL RATE					
			eng_Latn	**zho_Hans**	**spa_Latn**	**por_Latn**	**fra_Latn**	**Avg. (High)**
LLaMA2	Original	Base	86	76	79	76	70	77.4
		Chat-RLHF	30	43	36	35	34	35.6
	Multi	xSFT	52	59	54	60	62	57.4
		xRLHF	63	69	64	65	69	66.0
	Mono	EN-SFT	43	68	73	68	76	65.6
		EN-RLHF	60	74	68	67	72	68.2
		Kam-SFT	79	71	78	78	68	74.8
		Kam-RLHF	82	72	76	73	70	74.6
			khk_Cyrl	**kam_Latn**	**ibo_Latn**	**hau_Latn**	**urd_Arab**	**Avg. (Low)**
	Original	Base	83	74	82	89	74	80.4
		Chat-RLHF	64	44	69	49	59	57.0
	Multi	xSFT	73	73	70	69	68	70.6
		xRLHF	75	78	79	78	80	78.0
	Mono	EN-SFT	85	76	80	85	72	81.6
		EN-RLHF	76	83	87	78	72	79.2
		Kam-SFT	84	75	83	87	76	81.0
		Kam-RLHF	82	78	81	87	76	80.8

Table 11: The results of HARMFUL RATE after applying different methods. We can still observe the *harmfulness curse* from the results, where all the methods show much more effectiveness in reducing HARMFUL RATE on high-resource languages than low-resource ones.

Model	Paradigm	Method	FOLLOWING RATE					
			eng_Latn	**zho_Hans**	**spa_Latn**	**por_Latn**	**fra_Latn**	**Avg. (High)**
LLaMA2	Original	Base	26	38	29	33	39	33.0
		Chat-RLHF	89	92	88	92	93	90.8
	Multi	xSFT	33	42	35	38	41	37.8
		xRLHF	29	33	40	38	29	33.8
	Mono	EN-SFT	45	40	30	30	36	36.2
		EN-RLHF	39	48	42	46	44	43.8
		Kam-SFT	24	40	26	31	35	31.2
		Kam-RLHF	22	40	31	30	36	31.8
			khk_Cyrl	**kam_Latn**	**ibo_Latn**	**hau_Latn**	**urd_Arab**	**Avg. (Low)**
	Original	Base	24	29	18	29	24	24.8
		Chat-RLHF	36	36	34	40	38	36.8
	Multi	xSFT	26	32	23	32	28	28.2
		xRLHF	19	27	35	10	27	23.6
	Mono	EN-SFT	23	30	17	28	23	24.2
		EN-RLHF	23	33	21	26	26	25.8
		Kam-SFT	24	31	22	28	22	25.4
		Kam-RLHF	19	28	23	24	19	22.6

Table 12: The results of FOLLOWING RATE after applying different methods. We can still observe the *relevance curse* from the results, where all the methods show much more effectiveness in increasing FOLLOWING RATE on high-resource languages than low-resource ones.