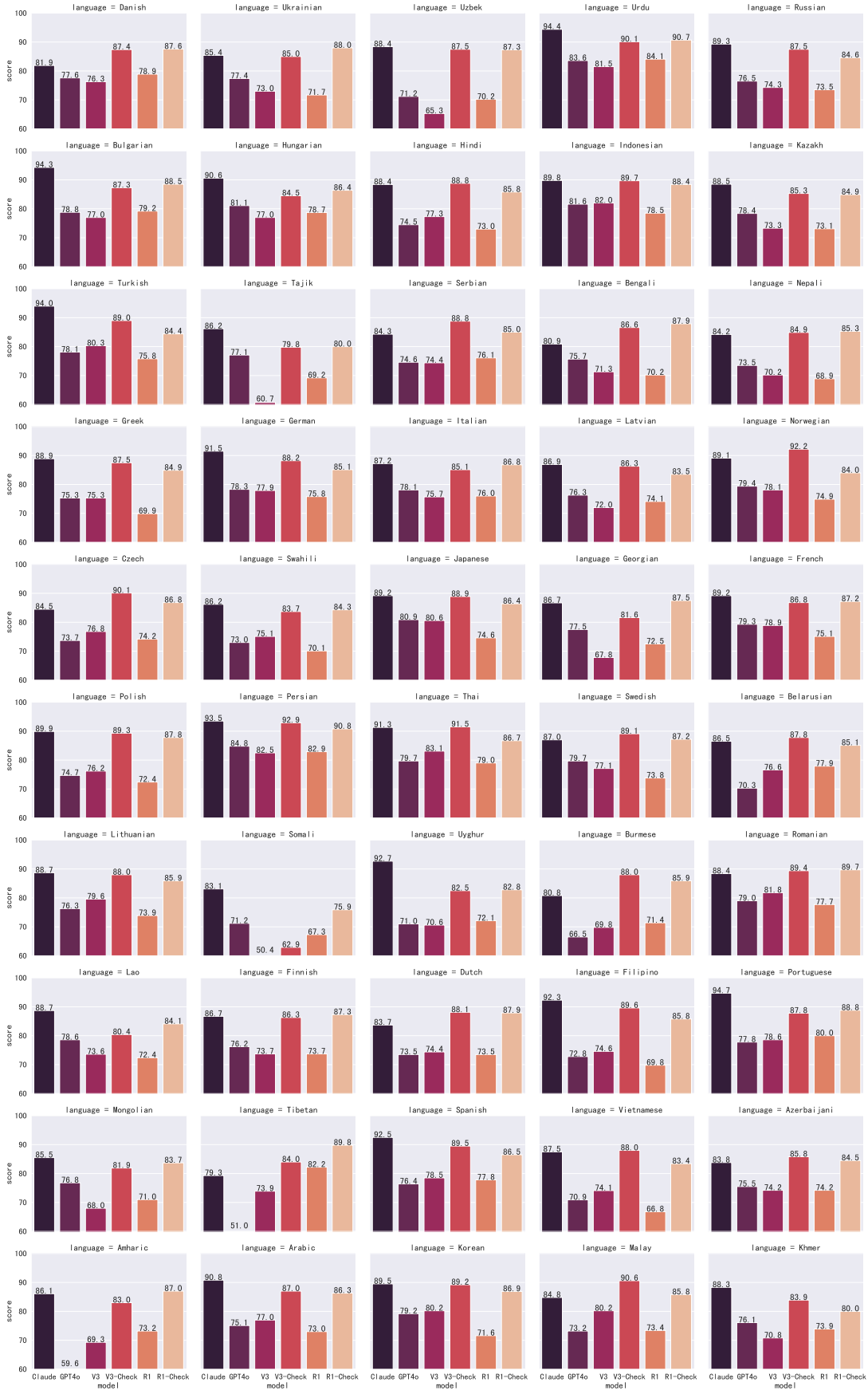

Figure 14 | Multilingual safety performance. V3-check and R1-check represent the risk control system evaluation results for DeepSeek-V3 and DeepSeek-R1, respectively.

D.3.4. Multilingual Safety Performance

In the previous section's evaluation, we primarily focused on the model's safety performance in special languages (Chinese and English). However, in practical usage scenarios, users' linguistic backgrounds are highly diverse. Assessing safety disparities across different languages is essential. For this purpose, we translated the original bilingual safety testset (introduced in the D.3.3) into 50 commonly used languages. For high-frequency languages, we conducted full translation of the entire dataset, while for low-frequency languages, we performed sampling translation. This process resulted in a comprehensive multilingual safety test set consisting of 9,330 questions. During the translation process, we employed a combined approach of LLM translation and human-assisted calibration to ensure the quality of the translations.

We continued to use the LLM-as-a-judge methodology described in the previous section, which determines safety labels (safe, unsafe, or rejected) for each question-answer pair. Rather than merely rejecting risky queries, we prefer responses that provide safe content; therefore, we assigned higher scores to safe responses (5 points per question, with 5 points for safe responses, 0 points for unsafe responses, and 4 points for rejections). The final safety score proportions (safety score as a percentage of the total possible safety score) across 50 languages are presented in Figure 14. For DeepSeek-V3 and DeepSeek-R1, we evaluated safety scores for models with and without the risk control system (introduced in D.3.1). Additionally, we tested the multilingual safety performance of Claude-3.7-Sonnet and GPT-4o(2024-05-13). From Figure 14, we can draw the following conclusions:

- With risk control system in place, DeepSeek-V3 (86.5%) and DeepSeek-R1 (85.9%) achieve total safety scores across 50 languages that approach the best-performing Claude-3.7-Sonnet (88.3%). This demonstrates that DeepSeek has reached state-of-the-art levels in system-level multilingual safety.
- Without risk control system, DeepSeek-V3 (75.3%) and DeepSeek-R1 (74.2%) get safety scores across 50 languages comparable to GPT-4o(2024-05-13)'s performance (75.2%). This indicates that even when directly using the open-source versions of R1, the model still exhibits a moderate level of safety standard.
- Examining language-specific weaknesses, we categorize languages with safety scores below 60 points as high-risk languages for the corresponding model. Among the 50 languages evaluated, DeepSeek-R1 (without risk control system) and Claude-3.7-Sonnet have zero high-risk languages; DeepSeek-V3 (without risk control system) and GPT-4o(2024-05-13) have one and two high-risk languages, respectively. This suggests that DeepSeek-R1 has no obvious language-specific vulnerabilities.

D.3.5. Robustness against Jailbreaking

In real-world application scenarios, malicious users may employ various jailbreaking techniques to circumvent a model's safety alignment and elicit harmful responses. Therefore, beyond evaluating model safety under direct questioning, we place significant emphasis on examining the model's robustness when confronted with jailbreaking attacks. Thus, we constructed a dedicated test suite for jailbreaking evaluation. Specifically, we developed a template collection consisting of 2,232 jailbreaking instructions. We then randomly concatenated these jailbreaking prompts with questions from the original safety testset (introduced in D.3.3) and further examined the performance differences in the model's responses when confronted with original unsafe questions versus newly formulated questions with jailbreaking elements.

When evaluating the results, we followed the LLM-as-a-Judge safety assessment (introduced

in D.3.3), while improving the safety evaluation prompts to focus more specifically on identifying manipulative traps in jailbreak attempts. Each question-answer pair was classified into one of three categories: safe, unsafe, or rejected (introduced in D.3.3). The results of jailbreak attacks against various models are presented in Table 11. From these results, we draw the following conclusions:

Table 11 | Comparison of DeepSeek-R1 and other frontier models in jailbreaking scenarios.

	Unsafe Ratio			**Rejected Ratio**		
Ratio(%)	Origin	Jailbreak	GAP	Origin	Jailbreak	GAP
Claude-3.7-Sonnet	10.7	26.2	+15.5	3.6	21.9	+18.3
o1 (2024-12-17)	9.0	12.1	+3.1	50.4	79.8	+29.4
GPT-4o (2024-05-13)	22.0	30.4	+8.4	17.1	57.3	+40.2
Qwen2.5 Instruct (72B)	13.8	29.7	+15.9	5.4	25.2	+19.8
DeepSeek-V3	17.6	36.4	+18.8	8.1	8.9	+0.8
+ risk control system	5.3	2.3	-3.0	25.4	46.5	+21.1
DeepSeek-R1	25.2	85.9	+60.7	5.6	1.9	-3.7
+ risk control system	8.5	4.3	-4.2	27.3	87.3	+60.0

- All tested models exhibited significantly increased rates of unsafe responses and rejections, along with decreased safety rates when facing jailbreak attacks. For example, Claude-3.7-Sonnet, showed a 33.8% decrease in the proportion of safe responses when confronted with our security jailbreak attacks. This demonstrates that current cutting-edge models still face substantial threats from jailbreak attacks.
- Compared to non-reasoning models, the two reasoning models in our experiments — DeepSeek-R1 and o1(2024-12-17) — rely more heavily on the risk control system for security checks, resulting in considerably higher overall rejection rates (79.8% and 87.3% respectively).
- Open-source models (DeepSeek, Qwen) face more severe jailbreak security challenges than closed-source models, because of the lack of a risk control system in locally deployed models. To address safety issues, we advise developers using open source models in their services to adopt comparable risk control measures.

E. More Analysis

E.1. Performance Comparison with DeepSeek-V3

Since both DeepSeek-R1 and DeepSeek-V3 share a common base architecture, namely DeepSeek-V3-Base, a critical question naturally arises: which specific dimensions are enhanced through the application of different post-training techniques? To address this, we first compare the R1 family of models with DeepSeek-V3 and DeepSeek-V3-Base, as summarized in Table 12. Notably, DeepSeek-R1 demonstrates significant improvements in competitive programming and mathematical reasoning tasks, as evidenced by superior performance on benchmarks such as LiveCodeBench and AIME 2024. These enhancements in reasoning capabilities also translate into higher scores on the Arena-Hard evaluation suite. Furthermore, DeepSeek-R1 exhibits stronger long-context understanding, as indicated by its improved accuracy on the FRAMES